SHORTFALL UTILITARIANISM

A THEORY FOR VARIABLE POPULATION DECISIONS

ROBYN KATH

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy
Faculty of Arts and Social Sciences
University of Sydney
2017
This thesis is all my own work. To the best of my knowledge, all sources have been duly acknowledged.
Acknowledgements

Many, many thanks to my supervisors—Michael McDermott, David Braddon-Mitchell, Josh Parsons, and Duncan Ivison—for encouraging, reading, signing, debating, teaching, cajoling, and more.

Thanks to Matthew Hanser for enabling my visit to UCSB and for many fruitful discussions during that visit, and to Jonathan Winterbottom for private tuition.

Thanks to Matthew Clark, Pierrick Bourat, Johann Hariman, and Teru Thomas for reading chapter drafts. Thanks to Martin Pickup and the TWiPers for reading a draft chapter and for particularly enjoyable philosophising.

Thanks to Mark Colyvan for being unnecessarily supportive and interested, and to Kristie Miller for coordinating masterfully.

Thanks to David Boonin, Gustaf Arrhenius, Derek Parfit, and Ralf Bader for sharing manuscripts.

Thanks to Michelle Boulous-Walker, Phil Dowe, Marguerite La Caze, Michael Ure (and Anita, Louie, and Jules), and Peter Cryle for inspiration, encouragement, and my earliest encounters with academia and philosophy.

Thanks to Josh Combes, Elizabeth Kath, and Sue Cusbert for support of many kinds. Thanks to David Maze, Genevieve Wilks, Bronwen Nicholls, Calyn van Wyk, Johann Hariman, and Samantha Groenestyn for good times.

Thanks to my family, for love and support and lots of putting-up-with.

And thanks to John, for everything.
Abstract

In this dissertation I propose a novel utilitarian moral theory, Shortfall Utilitarianism (SU). The main advantage of SU over more familiar versions of utilitarianism (including totalist, averagist, and harm-minimising versions) is that it agrees with several common intuitions about variable population decisions. These intuitions concern ‘the Asymmetry’, ‘the non-identity problem’, and ‘the repugnant conclusion’.

According to SU one ought to minimise two kinds of worseness among available outcomes. Intersectional worseness is a matter of how well the people who exist in both of a pair of outcomes fare in those outcomes. Complementary worseness is a matter of how well the people who exist in either (but not both) of a pair of outcomes fare in those outcomes. Underpinning these two kinds of worseness are many respects of (standardly structured) betterness among outcomes. Very roughly, minimising intersectional and complementary worseness amounts to minimising the extent to which people’s lives are worse, in certain respects, than what might have happened instead.

Part of what makes SU a utilitarian theory is that it treats fixed population decisions in the standard utilitarian way. It also has three further characteristics required of a utilitarian theory: it is a consequentialist, axiological, welfarist theory. As I define them, these characteristics do not require an overall betterness relation among outcomes. According to SU there is no such relation, which is part of what enables SU to deal intuitively with variable population decisions. My secondary goal in the dissertation is to encourage further exploration of the utilitarian possibilities that rejecting the overall betterness requirement opens up.
## Contents

Abstract

1 Deciding who exists
   1.1 Outline .................................................. 1
   1.2 Moral decision-making ................................. 4
   1.3 Method .................................................. 13

2 General utilitarian theories ................................. 19
   2.1 Classical Fixed Population Utilitarianism ............ 20
   2.2 Further constraints .................................... 22
   2.3 Three contenders ...................................... 40

3 Variable population decisions ................................ 49
   3.1 Wretchedness .......................................... 50
   3.2 Neutrality .............................................. 61
   3.3 Non-identity ........................................... 72
   3.4 Repugnance ............................................. 77

4 Shortfall Utilitarianism ....................................... 91
   4.1 Introducing SU .......................................... 91
   4.2 SU is a general utilitarian theory ................... 110
   4.3 SU and variable population decisions ............... 119

5 Defending Shortfall Utilitarianism ........................ 133
   5.1 Menu-dependence ....................................... 134
   5.2 More on overall betterness ............................ 148
   5.3 Some conclusions ...................................... 163
   5.4 SU and risk ............................................. 178
CONTENTS

6 Conclusion 185

A List of Decisions 193

Bibliography 201
Chapter 1

Deciding who exists

Sometimes, our decisions affect who comes to exist. This dissertation is about how we ought to make decisions, and it is especially about how we ought to make decisions that affect who comes to exist. In it I have two objectives. First, to propose and defend a particular moral theory, Shortfall Utilitarianism. Second, to draw attention to some unexplored possibilities in a broad class of moral theory: what I call ‘general utilitarian theories’. Shortfall Utilitarianism is a general utilitarian theory. It is also a theory that accords with some moral beliefs and intuitions that are both widely shared and notoriously difficult to reconcile. This, I argue, gives us reason both to prefer Shortfall Utilitarianism to other known theories, and to further explore the class of general utilitarian theories.

1.1 Outline

Let me begin with an outline of the dissertation, chapter by chapter. This first chapter has three sections (of which this outline constitutes the first). In Section 1.2 I set the scene. I introduce some useful terminology, and adopt what I hope is a fairly uncontentious framework in which to represent decisions and moral theories. This enables me to more precisely frame the question of how we ought to make decisions that affect who comes to exist. I also explain why I think that this question is an important one. In Section 1.3 I make some comments on my methods; on the roles of intuitions, consistency, and hypothetical cases in this
dissertation.

In Chapter 2 I narrow my focus to the question of how we ought, as utilitarians, to make decisions that affect who comes to exist. My goal in this chapter is to carve out a space of possible utilitarian moral theories; to characterise the class of general utilitarian theories. I begin, in Section 2.1, by assuming Classical Fixed Population Utilitarianism (CFPU), the theory that one ought to maximise total wellbeing in decisions that don’t affect who comes to exist. The first constraint on a general utilitarian theory is that it agrees with this restricted theory, but also tells us what to do in decisions that do affect who comes to exist. It is in this sense an ‘extension’ of CFPU. In Section 2.2 I further constrain the class of general utilitarian theories by appealing to some characteristics of CFPU that seem central to the utilitarian spirit. I stipulate that a general utilitarian theory is consequentialist, axiological, and welfarist (and provide definitions for each of these characteristics). In this section I also consider and reject another possible constraint, which I call the ‘overall betterness requirement’. The overall betterness requirement features frequently in the population ethics literature, and its rejection will prove important. Having thus characterised the class of general utilitarian theories, in Section 2.3 I introduce three of its well known members: Total Utilitarianism (TU), Average Utilitarianism (AU), and Harm Utilitarianism (HU).

In Chapter 3 I endorse four widely shared moral intuitions about decisions that affect who comes to exist. Those intuitions, roughly described, are as follows. First, in Section 3.1: that a life can be so bad for the person who lives it, that one ought not create such a person. Second, in Section 3.2: that a life cannot be so good for the person who lives it, that one ought to create such a person. Third, in Section 3.3: that one ought not make it that one happy person exists, instead of a different happier person existing. And fourth, in Section 3.4: that one ought not make it that many happy people exist, instead of fewer happier people existing. Anyone who shares these intuitions should prefer a general utilitarian theory that agrees with them all to one that doesn’t, other things being equal. As I introduce and motivate each intuition, I use it to ‘test’ the general utilitarian theories introduced in the previous chapter (TU, AU, and HU), as well as some variations thereof. None of the theories that I consider agrees with all four intuitions.

In Chapter 4 I propose Shortfall Utilitarianism (SU), and show that it is a
1.1. OUTLINE

general utilitarian theory that agrees with all four of the intuitions endorsed in the previous chapter. In Section 4.1 I set out SU as clearly and simply as I can. It is a theory according to which one outcome can be worse (or better) than another in two ways. It might be intersectionally worse, which is a matter of how well the people who exist in both outcomes fare in each; and it might be complementarily worse, which is a matter of how well the people who exist in one of the outcomes but not the other fare, in the outcome in which they exist. According to SU one ought to make decisions on the basis of these two kinds of worseness. (Exactly how will have to wait.) In Section 4.2 I show that SU is a general utilitarian theory: that it is a consequentialist, axiological, welfarist extension of CFPU. I also show that SU does not meet the overall betterness requirement. In Section 4.3 I show that SU agrees with the four test intuitions, and argue that this provides some reason to prefer SU to any of the other general utilitarian theories considered.

In Chapter 5 I defend Shortfall Utilitarianism from some possible objections. One feature of SU is a kind of ‘menu-dependence’: what one ought to do depends in interesting ways on what the available options are. In Section 5.1 I consider whether this kind of menu-dependence might be problematic; for example, whether it might lead to cyclical decision-making. I argue that there is no space for these apparent problems in the framework I have adopted. In Section 5.2 I address possible concerns about the absence of an overall betterness relation among outcomes. I argue that this should not concern us, but also endorse the option of reclaiming overall betterness by ‘fine-graining’ outcomes. In Section 5.3 I consider some conclusions of SU in particular decisions, which may strike some as unintuitive. I defend the conclusions of SU, and also suggest some possible avenues of alteration of the theory. In Section 5.4 I suggest a way of extending SU to deal with risky decisions.

Finally, in Chapter 6 I draw my conclusions. I conclude that, based on my investigations so far, SU is the most compelling general utilitarian theory of which I am aware. However, I recommend further investigation of the class of general utilitarian theories, and in particular those that do not meet the overall betterness requirement. This latter kind of theory has, I suggest, been significantly overlooked until now, and may include further compelling options.

The Appendix (A) contains a list of all of the named decisions discussed in
1.2 Moral decision-making

What is a decision? When do I face one, and how do I make one? What is the relationship between a moral theory and a decision? Which decisions affect who comes to exist, and why are they important? These are the questions I shall address in this section. First I’ll define a decision, and a moral theory. Then I’ll focus in on some particularly relevant aspects of decisions: outcomes, populations, and wellbeings. Then, finally, I’ll be able to pick out more precisely the decisions that affect who comes to exist, and explain why they are important.

I hope that most of what I say here is uncontentious, at least in the context of my project of developing a utilitarian moral theory. That is, where I do take a stance on some difficult issue, I intend nothing that I go on to say in the rest of the dissertation to depend on this stance (except where I note otherwise). My intention here is just to establish a convenient framework in which to represent the issues I want to address. I follow existing population ethics literature fairly closely in doing so.¹

1.2.1 Decisions and moral theories

A decision is a set of alternatives, of which one and only one must be chosen. I face a particular decision when the choice among a particular set of alternatives is open for me to make. For example, this morning I faced the decision between wearing red suspenders, wearing blue suspenders, and wearing no suspenders (let’s suppose I couldn’t have worn both). I make a decision when I choose one of the alternatives and close the matter. So, this morning I made the decision to wear blue suspenders, ruling out wearing red suspenders or none.

What kind of thing are the alternatives in a decision? It might at first seem that they can be various kinds of things: sometimes I seem to face a decision between puddings, sometimes between careers, and sometimes between physical movements. I assume, however, that the alternatives in a decision are always acts. An apparent decision between puddings, for example, is always really a

¹I am particularly influenced by Parfit 1984, McDermott 1982b, and Broome 2004.
decision between various acts of eating (or ordering, or mentally committing to order) one pudding or another. Sitting still in an indecisive frame of mind is just as much an act as is jumping up and down (in such-and-such particular way).

A person faces a decision at a particular time between all and only the acts that are available to that person at that time—available in roughly the sense that it is possible for the person to perform that act. In practice, when people deliberate they often focus on some subset of the acts that they could possibly perform. I often wonder whether to keep writing or make a cup of tea. I much less often consider the equally possible option of pouring apple juice on my head. We also seem to individuate acts more or less finely in different contexts. I often wonder whether to keep writing or make a cup of tea; at other times, and although what is possible doesn’t seem to have changed, I wonder whether to make an Earl Grey or an English Breakfast. For these and other reasons, the issue of which decision a person faces at a particular time (i.e. which acts are available to a person at a time) is not straightforward. Throughout most of this dissertation, though, I think that we can rely on a fairly intuitive sense of what the available acts are. In Section 5.1 I return to this issue.

So, a decision is a set of available acts. It will be useful to have a shorthand way of labelling acts and decisions. To label acts I will use lower-cased letters: \(a\), \(b\), \(c\), etc. To label decisions I’ll use a ‘D’ with numerical subscripts: \(D_1\), \(D_2\), \(D_3\), etc. If I write, for example, ‘\(D_1 = \{a, b, c\}\)’, this should be read as saying that \(D_1\) is the decision consisting of (or between) the three available acts, \(a\), \(b\) and \(c\).

We all face decisions all the time. And, inevitably, we make those decisions. We choose one of the available acts, making it the act that we actually perform. This dissertation is about moral decision-making; about how one ought, morally, to choose from a given set of available acts. A moral theory takes a decision as an input, and returns some kind of choice-guiding output. A moral theory might, for example, do one or more of the following: pick out some subset of the available acts in the input decision as impermissible, permissible,
or obligatory in that decision; identify an available act in the decision as the one that ought to be chosen in that decision; order the available acts in the input decision in terms of choice-worthiness in that decision.

Each of the moral theories that I consider does all three of these things. Mostly I focus on the ordering of the available acts in a decision in terms of choice-worthiness in that decision. What I mean by the choice-worthiness of an act in a decision is how morally good it would be to choose that act when faced with that decision. To say that act \( a \) is more choice-worthy in decision \( D_1 \) than act \( b \) is to say that it would be morally better to choose \( a \) than to choose \( b \) when faced with \( D_1 \). I use the terminology of ‘choice-worthiness’, rather than that of ‘moral goodness’, to clearly distinguish this notion from another (likely related, but importantly different) kind of goodness. I want to guard against the collapse of these two notions: how good some act is in itself, or intrinsically (perhaps in a morally relevant respect, or perhaps overall); and how morally good it is to choose some act from a particular set of available acts. It is at least conceptually possible that some act is very good in itself (in some morally relevant respect, or even overall), and yet that it is very morally bad to choose that act when faced with some particular decision. I shall have a lot more to say about the relationship between these kinds of goodness, and I think that I am able to address that relationship more clearly by assigning them distinctive terminologies.

Each of the moral theories that I consider orders all of the available acts in an input decision for choice-worthiness in that decision. This ordering is complete. It says for each pair of available acts in the decision, either that one is more choice-worthy in that decision than the other, or that they are equally choice-worthy in that decision. This means that it also identifies some available acts in the decision as the ones that are most choice-worthy in that decision. I make the further general assumption that in any given decision, only the available acts most choice-worthy in that decision are permissible, and all other available acts are impermissible. If only one available act is permissible in some decision, that act is obligatory—i.e. ought to be chosen—in that decision.\(^4\)

Choice-worthiness can be represented in terms of betterness relations. There

\(^4\)That is, some \( a \in D_1 \) is permissible in \( D_1 \) iff no \( b \in D_1 \) is more choice-worthy in \( D_1 \) than \( a \); any \( a \in D_1 \) is impermissible in \( D_1 \) iff \( a \) is not permissible in \( D_1 \); if some \( a \in D_1 \) is permissible in \( D_1 \) and no \( b \neq a \in D_1 \) is permissible in \( D_1 \), then \( a \) is obligatory in \( D_1 \). I realise that these maximising assumptions are contentious, but I don’t think they play a significant role in what follows.
is a betterness relation corresponding to the choice-worthiness ordering for each decision: the relation of being more choice-worthy in that decision (that is, morally better to choose when faced with that decision). It will be useful to have a shorthand way of representing these relations. For the relation of being more choice-worthy in $D_1$ than, I will write $\succ_{D_1}$; for the relation of being more choice-worthy in $D_2$ than, $\succ_{D_2}$; and so on. If I write $a \succ_{D_1} b$, for example, this should be read as saying that $a$ is more choice-worthy in $D_1$ than $b$ (or that it would be morally better to choose $a$ than to choose $b$ when faced with $D_1$).

Each of the moral theories that I consider provides even more detailed choice-guiding output than this. They say not only when one available act is more choice-worthy in a particular decision than another, but also how much more choice-worthy in that decision the one is than the other. This means that we can also represent each theory as a function from decisions to choice-worthiness rankings; as taking a set of acts and returning an integer for each act, representing the choice-worthiness of that act in that decision. I don’t make any use of this further information, so I won’t go into details. Note, though, that the orderings have interval but not ratio structure; and that the different theories likely use different choice-worthiness ‘scales’, so that comparing a difference in choice-worthiness according to one theory with a difference in choice-worthiness according to a second theory is not necessarily straightforward.

---

A note on two alternatives here. First: instead of representing choice-worthiness in terms of many binary betterness relations (one for each decision), one might represent it in terms of a single three-place relation, taking as its relata two acts and a decision. One might then for example write ‘$Rab\, D_1$’ to represent that claim that it would be at least as morally good to choose act $a$ as to choose act $b$, when faced with decision $D_1$. I use the many binary betterness relations here just because they are familiar. Second: the choice-worthiness output of some theories can be represented in terms of a single (standard, binary) betterness relation. Some of the theories that I consider are like this, but others are not. Why not will become evident in the next chapter. In the mean time I ask the reader just to go along with me in using the more accommodating representation.

I’ll label the related relations, for example of being at least as choice-worthy in a decision, similarly. For example, I’ll write $a \succeq_{D_1} b$ for the claim that $a$ is at least as choice-worthy in $D_1$ as $b$. I say more about the structure of betterness relations, and at-least-as-good-as relations, in Section 2.2.2.

A common illustration of an interval (and not ratio) scale is temperature. If two things have temperatures, it is always the case either that one temperature is higher than the other or that they are the same, and we can say by how much one is higher than the other (having settled on a particular scale, Celsius or Fahrenheit for example). Yet, it is not possible to say that one temperature is twice as high as another. The point about the difficulty in inter-theoretic comparisons is potentially important in the context of moral uncertainty. When we are uncertain which moral theory is correct, we nonetheless need to have some way of making decisions. If theories shared a common scale of choice-worthiness it might be easier (though perhaps still not easy) to know how to proceed. On moral uncertainty see for example Sepielli 2006 and Riedener 2015.
A moral theory ‘makes’ a decision by ordering the available outcomes for choice-worthiness in that decision. The correct moral theory is the theory that orders the available outcomes in every decision in accordance with their actual (correct) choice-worthiness (setting aside concerns about whether there is a correct moral theory). In this dissertation I will evaluate a number of moral theories, each identified by the choice-worthiness orderings they provide.

1.2.2 Outcomes, populations, and wellbeings

One thing that may affect choice-worthiness in a decision is what the consequences of each available act in the decision would be, if it were chosen (i.e. performed). What would happen if this act were chosen, and what would happen if I chose the other? Throughout most of this dissertation I assume that each available act in a decision would, if chosen, make it the case that some particular possible world is the actual world.\(^8\) I assume that this information is part of the content of a decision. I shall label the possible worlds using the upper case letters: \(A, B, C, \ldots,\) etc, and call them outcomes.

I thus assume that all decisions are decisions under certainty: each available act would, if performed, bring about some particular outcome, and we know which outcome that would be for each act. This is a significant simplification. In reality, two kinds of uncertainty might infect our decisions: subjective and objective uncertainty. As a rule, I don’t know which outcome will result from my actions. In some cases, there may also be no fact of the matter, even beyond my epistemic access. A more complete picture of a decision would include for each available act a prospect, consisting of the various outcomes the act might (if chosen) lead to, together with the probabilities with which the act would lead to each of those outcomes. Nonetheless, for the time being I make this assumption of certainty. The issues I have to deal with are quite complex enough even with it. In Section 5.4 I broach the issue of how to deal with the reintroduction of uncertainty.

One feature of outcomes with which I am particularly concerned is which people exist in them. Call a person who exists in at least one possible world a possible person. I label the possible people: \(p_1, p_2, p_3, \ldots\) (except when I

\(^8\)This claim is sometimes called counterfactual determinism. It is often endorsed in consequentialist literature, and in population ethics: for example in Bykvist 2003, 30, Bykvist 2002, 46, and Brown 2011b, 752.
1.2. MORAL DECISION-MAKING

sometimes give them more familiar names, like Anne and Bob). I call the set of people who exist—that is, ever exist—in an outcome the population of that outcome. To be clear, I mean the set of people and not the number of people in that set (so there are many different populations of the same size). Each outcome has a population (possibly the empty population). For the population of outcome \( A \) I’ll write \( P_A \), for the population of \( B \), \( P_B \), and so on. If I write \( P_A = \{p_1, p_3, p_5\} \), this should be read as saying that the only people who ever exist in outcome \( A \) are \( p_1 \), \( p_3 \) and \( p_5 \).

I’m interested not only in whether people exist, but also in how well they fare. I assume that each person has a lifetime wellbeing in each outcome in which she exists, which is a measure of how good her life in that outcome is for her. I also assume that a person does not have a wellbeing in an outcome in which she doesn’t exist (though I’ll revisit this assumption in Section 3.1). I assume that these lifetime wellbeings are both intra- and inter-personally comparable and can be represented by integers. Think of one particular, fully detailed life lived by some person; and of a second particular, fully detailed life lived by some person (whether the same person or not). There is a fact about how well the first life goes for the person who lives it, and a fact about how well the second life goes for the person who lives it, and these are measures on the same scale. The first life is either higher, lower, or equal to the second, in terms of lifetime wellbeing. The scale of lifetime wellbeing is an interval scale. That scale has no privileged zero-point (for the time being—again, I’ll come back to this in Section 3.1).

I make no assumptions about what lifetime wellbeing consists in, or is determined by.\(^9\) It seems plausible to me that lifetime wellbeing is a function of momentary wellbeings at each time throughout a life. However, I’m not sure what the relationship between momentary (or ‘temporal’) wellbeings and lifetime wellbeing would be. Is lifetime wellbeing the sum of all of the momentary wellbeings in a life? Or perhaps their average? Or does the ‘shape’ of a life matter? Is it better, for example, to have a life that starts poorly but ends well rather than the other way around? I am also not sure which features of lives contribute to (momentary or lifetime) wellbeing. One suggestion I find plausible is that wellbeing is determined by happiness, or pleasure, broadly.

\(^9\)For more on wellbeing, what determines or constitutes it, see Griffin 1986, Feldman 2010, and Wolf 2010.
construed. Although this suggestion is somewhat appealing, it is also fairly uninformative in the absence of a detailed account of happiness or pleasure (notoriously also difficult to provide). Other plausible proposals include a preference satisfaction account and an objective list account of wellbeing. I leave all of these questions about wellbeing open. From now on I will mostly refer to lifetime wellbeing, the kind of wellbeing with which I am primarily concerned, simply as ‘wellbeing’.

1.2.3 Variable population decisions

Now I can more clearly distinguish between decisions that do affect who comes to exist, and those that don’t. First, let me introduce a way of partially representing a decision in a table. Here is such a table.

\[
\begin{array}{l|cc}
D_1: & \text{Anne} & \text{Bob} \\
\hline
a \Rightarrow A & 10 & 8 \\
b \Rightarrow B & 9 & 9 \\
\end{array}
\]

This table represents \( D_1 = \{a, b\} \), where if act \( a \) is chosen outcome \( A \) will result and if act \( b \) is chosen outcome \( B \) will result. Some features of those outcomes are represented in the two right-most columns: in \( A \) two people exist, Anne with wellbeing 10 and Bob with wellbeing 8; in \( B \) the same two people exist, each with wellbeing 9. No-one else exists in either outcome. Here is another table.

\[
\begin{array}{l|cc}
D_2: & \text{Anne} & \text{Bob} \\
\hline
a \Rightarrow A & 10 & 8 \\
c \Rightarrow C & 9 & * \\
\end{array}
\]

This table represents \( D_2 = \{a, c\} \). As above, if \( a \) is chosen \( A \) will result. If \( c \) is chosen \( C \) will result, and in \( C \) only one person exists: Anne with wellbeing 9. I represent Bob’s non-existence in \( C \) with a ‘*’.

The first decision, \( D_1 \), is a decision that doesn’t affect who comes to exist. Exactly the same people (ever) exist no matter which of the available acts in \( D_1 \) is chosen. The second decision, \( D_2 \), is a decision that does affect who comes to exist. In \( D_2 \), if \( a \) is chosen Anne and Bob exist, while if \( b \) is chosen only Anne exists. I call a decision that doesn’t affect who comes to exist a ‘fixed
population decision’, and a decision that does affect who comes to exist a ‘variable population decision’. Here are more precise definitions of each.

*Fixed population decisions*: Any decision in which the populations of the outcomes of each available act are identical is a fixed population decision. No other decision is a fixed population decision.

*Variable population decisions*: Any decision in which the populations of the outcomes of at least two available acts are not identical is a variable population decision. No other decision is a variable population decision.

Every decision is either a fixed population decision or a variable population decision, and none is both.

Note that in some variable population decisions the size of the population is fixed, but the identities of the members of the populations is not. Here is a table representing a decision like that.

<table>
<thead>
<tr>
<th></th>
<th>Anne</th>
<th>Bob</th>
<th>Carl</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a \implies A$</td>
<td>10</td>
<td>8</td>
<td>*</td>
</tr>
<tr>
<td>$d \implies D$</td>
<td>9</td>
<td>*</td>
<td>9</td>
</tr>
</tbody>
</table>

In $D_3$ Anne and Bob will exist if $a$ is chosen, whereas Anne and Carl will exist if $d$ is chosen. This decision affects who comes to exist even though it doesn’t affect how many people come to exist.

### 1.2.4 Motivation

Now that I’ve defined a variable population decision I can explain why they are important. In brief, there are three reasons: first, they are very common; second, they are morally perplexing and divisive; and third, how we treat them will have significant consequences for our assessment of many pressing practical decisions.

---

10 This distinction and terminology come from McDermott 1982b. Common alternatives come from Parfit 1984 (see page 356). Parfit calls fixed population decisions ‘same person decisions’, and identifies two other categories: ‘same number decisions’ and ‘different number decisions’. The category of same number decisions includes all fixed population decisions as well as some variable population decisions (such as $D_3$ below). The category of different number decisions covers the rest of the variable population decisions (such as $D_2$ above).

11 Any decision, $D_1$, is a fixed population decision iff for any pair of outcomes of acts in $D_1$, $A$ and $B$, $P_A = P_B$.

12 Any decision, $D_1$, is a variable population decision iff for some pair of outcomes of acts in $D_1$, $A$ and $B$, $P_A \neq P_B$. 
I’ll begin with the first point: that variable population decisions are very common. This comes as a surprise to some people, when they first begin to think about variable population decisions. These decisions are so common because the existence of any particular person is highly contingent. Think of any person existing now: her existence depended on countless decisions made by her parents and ancestors, as well as many other people.

Consider me. I exist because my parents decided to have a fourth child, just when they did so. If they had not decided to have a fourth child, I would never have existed. If they had decided to have a fourth child earlier or later than they in fact did, I would never have existed (and some other person(s) would have existed instead). If they had not both decided to go to the university where they met, I would never have existed (and nor would any of my siblings, nieces or nephews). My paternal grandparents migrated to Australia; my great-grandparents married on a Tuesday; an English judge sent my great-great-grandmother to Australia; Alexander Bell became a scientist and invented the telephone. I think that it is safe to say that if any of these people had acted differently, I would never have existed. And of course, those people faced decisions about whether to act as they did; decisions which, though they may not have realised it, affected who would come to exist.

My existence, then, was a very fragile matter indeed. Similarly, the existence of very many possible future people depends on the decisions that I make. Already, through my decisions about whom to befriend, where to study and work, and in all likelihood what to have for lunch one Wednesday, I have made it the case that some possible people will come to exist, and that many more will not. It is possible that most decisions are in fact variable population decisions.

Exactly which decisions are variable population decisions depends on how persons are individuated. I don’t assume any particular theory of personal identity.\(^\text{13}\) I think that throughout this dissertation we can rely on an intuitive understanding of the boundaries of personhood. I have so far relied on the idea that the same person generally does not grow from two different pairs of gametes. This seems intuitively compelling to me, but is not important for anything that follows. Whenever I discuss a case in which this supposition plays a role, an analogous case in which it does not is readily available.

The sheer prevalence of variable population decisions provides some mo-

\(^{13}\)For more on personal identity and variable population decisions see Parfit 1984.
1.3. METHOD

Before I begin in earnest I’d like to make some notes on method, and on how I see this dissertation relating to existing work. I take myself to share broad goals and methodology with much of the large literature on population ethics. Having realised the significance of the fact that many of our decisions affect who comes to exist, we seek an answer to the question of how we ought to make such decisions. We seek an answer to this question largely by the
dual guidance of intuitions and consistency, and we explore our intuitions by considering particular decisions and cases.

1.3.1 Intuitions and consistency

The intuitions at play can be divided roughly into two categories: the general, and the specific. (Or they may be better thought of as on a scale from one to the other.) More general intuitions tend be expressed in the form of principles or ‘views’. One example of a fairly general intuition is the consequentialist idea that all that matters morally is how good the consequences of our action are. More specific intuitions often relate to particular decisions, or kinds of decision. An example of a more specific intuition is that one ought to create a happier rather than a less happy person, if possible.

Intuitions, particularly the more general ones, can be difficult to spell out or capture. However, we do our best. We try carefully to represent the principle, view, or claim that we find compelling. For example, we might rephrase the two intuitions presented casually above in this more precise way: ‘choice-worthiness in any given decision is fully determined by the outcomes of the available acts in that decision’; and ‘in a decision between creating some person with a given wellbeing, and creating a different person with a higher wellbeing, other things being equal, one ought to choose the latter’. Having captured our intuitions as best we can, we can then test them for consistency. However intuitively compelling a set of claims may be, if they are inconsistent they cannot all be true.

Consistency has turned out to be quite a powerful tool in the field of population ethics. At least since Parfit wrote Reasons and Persons (1984), inconsistencies between compelling intuitions have been at the heart of discussion. The precise locations of, and possible responses to, some of these inconsistencies have been examined by, among others, Arrhenius (2000, 2009b, and 2016), Broome (2004), Roberts (2003 and 2011a), Temkin (2011), and Thomas (2015). The intuitions I appeal to in this dissertation feature often in that literature, and have been thought to be inconsistent. I aim to show that this is not the case. I diagnose the mistake made by those who claim to have proven them inconsistent, in their spelling out of the intuitions. The way that I spell them out, they are not inconsistent after all.

At least, that is my preferred way of seeing things. Here is another way.
1.3. METHOD

The claims of inconsistency among intuitions made in the population ethics literature are in fact correct. The claims of consistency among the intuitions that I defend are also correct. The apparent disagreement is illusory, and what explains the illusion is that the intuitions involved in the two kinds of claim are different. Their intuitions are inconsistent, while mine are consistent.

Deciding which way of seeing things is more accurate is difficult, and I think not particularly significant (if there even is a fact of the matter). I appeal to textual evidence throughout the dissertation where I think this indicates that someone shares my intuitions, rather than having a similar but importantly different intuition. However, there is a reading of my conclusions for the reader who shares my intuitions from the outset, and one for the reader who begins with some slightly different intuitions. The former I reassure: our shared intuitions are consistent after all! The latter I console: although your intuitions are inconsistent, my very similar and I think equally compelling intuitions are not.

I do more than show that a set of intuitions is consistent. I defend a particular utilitarian theory consistent with those intuitions. I do not, however, claim that this is the only, or even the best, utilitarian theory consistent with those intuitions. There are likely to be many such theories, and not until we have explored all (or at least more) of them will we be able to say with any confidence which is best. I hope that the reader will not find my conclusions disappointingly tentative. I admit that it would be in some ways more satisfying to see a particular theory emerge as the only possible theory.\(^{16}\) However, this would require me to commit to more or stronger intuitions, which I am wary of doing. For one thing, in sticking to familiar intuitions I can be sure that they will be widely shared. Also, as I’ve already noted, spelling out intuitions is difficult. The more I commit to, the greater the chance that I will make a mistake. At this stage it seems to me preferable to accept some uncertainty about which utilitarian theory to adopt, rather than risk ruling out good candidate theories by inadvertently representing my intuitions too restrictively.

\(^{16}\)As we do, for example, in Broome 2004.
1.3.2 Decision scenarios

Finally, a note on the use and representation of more specific intuitions—intuitions about particular decisions or kinds of decision. I have found it somewhat difficult to choose, or describe, decisions. Say I want to use a decision with some particular structure to illustrate a conclusion or principle. Should I use a realistic decision, so that the reader can easily picture herself in the decision scenario? Or should I use an unrealistic but simpler decision, with potentially distracting elements removed? A more realistic case might better prompt ‘genuine’, pre-theoretic intuitions; on the other hand, it might prompt intuitions influenced by confounding details. People are perhaps able to imagine realistic cases well and in detail; but it is not clear that they are able to do this rich imagining without including some extraneous, distracting information. A less realistic case might better focus on the target issue, but might not prompt intuitions in a very direct way.

An example. Say I want to address the question of whether one ought to create a very happy person, other things being equal. Among the ways in which I could (and shall) address this question are the following. First, I could try to access direct, general intuitions on the matter by asking: ‘Ought one to create a very happy person, other things being equal?’ Second, I could introduce a realistic case in which I try to isolate the act of creating a person from other plausibly morally relevant considerations: ‘You and your partner must decide whether to have a child. The child would, if she existed, be very happy. Your decision will not affect anyone’s wellbeing. Ought you to have a child?’ The decision whether to have a child is a very realistic and indeed common one, so an advantage of this line of enquiry might be that people can readily access pre-theoretic intuitions about the case. A disadvantage is that, although I have stipulated that things are otherwise equal and the decision is made under certainty, intuitions about this case might not reflect these stipulations, because in real decisions about having children they do not hold. Your intuition that you ought to have a child might be influenced by your belief that, in fact, parents are always made better off by having children. Or your intuition that you ought not have a child might be influenced by your belief that, in fact, having a child always makes other people worse off. The realism that makes the case so easy to imagine and care about also makes it difficult to exclude other realistically covariant factors.
Finally, third, I could appeal to a less realistic case: ‘You could press a button that will bring into existence, on a distant and well-endowed planet, a person. The person would be very happy and pressing the button will not affect anyone’s wellbeing. Ought you to press the button?’ This is a decision that I don’t believe anyone will ever actually face. It is in obvious ways quite far removed from any of the decisions that anyone has actually contemplated, and so people might be fairly theoretically led in their response to it. However, it has some of the advantage of appealing to the imagination and intuitive ethical concern, and seems unlikely to be prone to much corruption by realistically covariant factors. Because it is so unrealistic there aren’t many realistically covariant factors, and so people’s intuitions will be plausibly better focussed on the issue I am investigating: the moral status of creating a very happy person, other things being equal.

A similar concern arises regarding the use of particular integers to represent wellbeings in describing cases. The concern is that people may have one intuition about, for example, whether one ought to create a person with wellbeing 2, and a different intuition about whether one ought to create a person at wellbeing 1 million, even in the absence of any explicit information about what kind of life those wellbeing levels correspond to. It seems somewhat natural to think of 2 as a fairly low number, and to think of 1 million as a fairly high number. Similarly it seems natural to think of the difference between 2 and 3 as small, and the difference between 500 and 1 million as large. These kinds of numerical assumptions may infect moral intuitions.

My general approach in this dissertation is to describe cases with minimal content. For example, regarding the question of whether one ought to create a very happy person, other things being equal, I focus on a case thus described: ‘You could create a very happy person. Doing so will not affect anyone else. Ought you to do so?’ Using particular integers is too useful to avoid, but I try to vary my use of ‘small’ and ‘large’ integers. I encourage the reader to experiment with different, more detailed versions of my ‘bare cases’, and to experiment with using different integers (that maintain the structure of the decision) in those cases. Considering a variety of person-creation cases seems likely to enable one to develop a fuller picture of one’s intuitions. Sometimes, for example when I think that a bare case is quite difficult to fill out realistically, I will draw attention to a particularly realistic case. Likewise, when I think
that realistic covariant factors are likely to be misleading, I will introduce a particularly unrealistic case.
Chapter 2

General utilitarian theories

In this chapter I characterise in some detail the class of general utilitarian theories. This is meant to enable me to focus on the question of how we ought, *as utilitarians*, to make variable population decisions; or what the *utilitarian component* of morality says about variable population decisions. This does not mean that only card-carrying utilitarians, or those who believe that there *is* a utilitarian component of morality, should keep reading. Those people should perhaps be especially interested. But others might accept this assumption of utilitarianism temporarily, as a tool for narrowing our focus to some particular problems that arise in variable population decisions. For it may be that these problems arise not only in a relatively simple utilitarian framework (or close analogues), but also more widely. The lessons we learn within this framework, about how best to deal with these problems, might also be useful outside it.¹

So I set aside, at least for the time being, potentially morally relevant issues like equality, rights violation, and deontic constraints. I take it for granted that we know how we ought, as utilitarians, to make fixed population decisions. The difficulty is in deciding how to extend this common core of utilitarianism to variable population decisions too.

Let me remind the reader of what is to come in this chapter. In Section 2.1 I introduce Classical Fixed Population Utilitarianism (CFPU), a restricted version of utilitarianism that only makes fixed population decisions. A general

¹Of course, one might also be interested in exploring the difficulties and possibilities associated with utilitarianism without thinking that this is likely to lead one closer to the moral truth, just out of curiosity or in order to know one’s opponent in debate.
utilitarian theory must agree with CFPU in fixed population decisions, and must also make variable population decisions. In Section 2.2 I place three further constraints on a general utilitarian theory: it must be consequentialist, axiological, and welfarist. I also consider and reject a fourth constraint, the overall betterness requirement. In Section 2.3 I introduce three competing general utilitarian theories: Total Utilitarianism (TU), Average Utilitarianism (AU), and Harm Utilitarianism (HU).

2.1 Classical Fixed Population Utilitarianism

Here is a restricted version of utilitarianism, which only makes fixed population decisions.

Classical Fixed Population Utilitarianism (CFPU): For any fixed population decision, $D_1$, one available act, $a$, is at least as choice-worthy in $D_1$ as a second available act, $b$, just in case the total wellbeing of the population of the outcome of $a$ in the outcome of $a$ is at least as high as the total wellbeing of the population of the outcome of $b$ in the outcome of $b$.

The total wellbeing of the population of the outcome of an act in that outcome is, straightforwardly, the sum of the wellbeings of all of the people in that population in that outcome. CFPU orders all of the available acts in a given decision for choice-worthiness in that decision, according to the total wellbeings of the populations of their outcomes: the higher the total wellbeing, the more choice-worthy the act. This applies, of course, only to fixed population decisions. Regarding the choice-worthiness of acts in variable population decisions this theory is silent.

I take CFPU, and indeed any other moral theory, to be properly identified by the choice-worthiness orderings of acts in decisions that it delivers. So, although I have presented CFPU in terms of total wellbeings, there are a number of other

---

2What if the relevant population is the empty population? This question will arise at various points throughout the dissertation. I don’t think it is a very important question, outcomes with the empty population being unavailable in any decision faced by an already-existing person. (And what other kind of decision could there be?) However, for completeness, I think it is best to stipulate that the total wellbeing of the empty population is zero. The implication of this for CFPU is that the available acts are all equally choice-worthy in any fixed population decision where the fixed population is the empty one.
2.1. CLASSICAL FIXED POPULATION UTILITARIANISM

equivalent ways of presenting the theory. For example, ordering acts by average wellbeing (of the population of their outcome in that outcome) amounts to the same thing as ordering them by total wellbeing, in fixed population decisions. Imagine you face a fixed population decision. Exactly the same people, and thus the same number of people, exist in the outcome of each of your available acts. This means that putting those outcomes in order according to the total wellbeings of their populations amounts to the same thing as putting them in order according to the average wellbeings of their populations. Dividing each of the totals by some particular positive number doesn’t change their order. There are many other ways of presenting CFPU, as we shall see throughout this dissertation. I have chosen to present it in terms of total wellbeings merely for the sake of simplicity and familiarity.

CFPU represents, I think, the historical and conceptual core of utilitarianism. Though early proposals of utilitarianism, including Bentham’s (1879) and Mill’s (1871), are not explicitly or deliberately restricted to fixed population decisions, neither do they pay much attention to variable population decisions. Sidgwick recognises this oversight and realises, in his brief but insightful discussion of the issue in The Methods of Ethics, that utilitarians who agree about fixed population decisions face a decision about how to address variable population decisions (1907, 414–416). He recognises two possible general versions of utilitarianism: the one that maximises total wellbeing, and the one that maximises average wellbeing. (Sidgwick favours the former.)

The first defining feature of a general utilitarian theory is that it gives the same choice-worthiness orderings as CFPU in fixed population decisions. It must also order all of the available acts in each variable population decision for choice-worthiness in that decision. As every decision is either a fixed or a variable population decision, such a theory is general in the sense of giving a choice-worthiness ordering of available acts for every possible decision. A theory that meets these two conditions—that agrees with CFPU in fixed population decisions but also makes variable population decisions—I call an extension of CFPU.

Just to be clear: choice-worthiness is always choice-worthiness in a particular

---

3 Again, the empty population presents a difficulty. Should we stipulate that the ‘average’ well-being of the empty population is zero, or (more in line with mathematical convention) undefined? I don’t think it matters much here. In either case we reach the conclusion that all of the available acts in a fixed population decision in which the fixed population is the empty one are equally choice-worthy in that decision.
decision, and an extension of CFPU need only replicate the choice-worthiness orderings of CFPU in fixed population decisions. The choice-worthiness orderings in variable population decisions of extensions of CFPU are not constrained. Say that $D_1 = \{a, b\}$ is a fixed population decision, and that according to CFPU $a$ is more choice-worthy in $D_1$ than $b$. It may nonetheless be that, according to some extension of CFPU, $b$ is more choice-worthy in some variable population decision (say, $D_2 = \{a, b, c\}$) than $a$.

This way of framing my project—in terms of extending CFPU—closely follows McDermott in ‘Utility and Population’ (1982b). The idea is that we know how we ought, as utilitarians, to make fixed population decisions. Our motivating goal is to make people as well off as possible, and we spell this out most straightforwardly in fixed population decisions in terms of maximising total wellbeing. However, when we come to consider variable population decisions, it is less clear how to spell out the motivating idea. We realise that maximising total wellbeing is equivalent to following a number of other procedures (such as maximising average wellbeing) in fixed population decisions, and that these procedures come apart in variable population decisions. Which procedure represents the best general version of utilitarianism?

### 2.2 Further constraints

Here is an extension of CFPU.

**Cheesy Extension** (CE): For any fixed population decision, $D_1$, one available act, $a$, is at least as choice-worthy in $D_1$ as a second available act, $b$, just in case the total wellbeing of the population of the outcome of $a$ in the outcome of $a$ is at least as high as the total wellbeing of the population of the outcome of $b$ in the outcome of $b$.

For any variable population decision, $D_2$, one available act, $c$, is at least as choice-worthy in $D_2$ as a second available act, $d$, just in case the volume of cheese in the outcome of $c$ is at least as great as the volume of cheese in the outcome of $d$.

This theory is simply CFPU with an additional clause, to the effect that in variable population decisions choice-worthiness is determined by the cheese content of outcomes. This theory is absurd. It is also, I think, clearly not a
version of utilitarianism. No theory according to which choice-worthiness ever ultimately depends on cheesiness is a utilitarian theory. In this section I complete my definition of a general utilitarian theory, adding more constraints to rule out absurd theories like CE and restrict my focus to theories that might plausibly be thought to maintain the utilitarian spirit motivating CFPU. I stipulate that a general utilitarian theory must be consequentialist, axiological, and welfarist. I take each of these characteristics in turn, propose a definition of that characteristic, and incorporate it into the definition of a general utilitarian theory.\footnote{CE is a consequentialist theory, and an axiological theory (according to my definitions). It is the welfarist constraint that will rule out CE.}

The definitions of consequentialism and axiology that I propose differ from some recent definitions of these terms (whereas my definition of welfarism is fairly standard). Both have been associated with what I call the ‘overall betterness requirement’, whereas I propose broader, more inclusive definitions. I try to motivate these different definitions to some extent, but I don’t think that the terminological dispute is very important for present purposes. I focus on the issue of which characteristics to require in a general utilitarian theory, rather than which names to give them. The overall betterness requirement has been widely adopted in population ethics (usually in association with the term ‘axiology’). I do not make the overall betterness requirement a defining feature of a general utilitarian theory. In Section 2.2.3 I explain why not.

\subsection{Consequentialism}

A general utilitarian theory should be a consequentialist theory. Utilitarianism is often cited as the quintessential consequentialist theory, and the idea that consequences are all that matter morally is central to utilitarian thought. To make this a useful constraint, however, I need to spell out more precisely what it means to be a consequentialist theory.

A moral theory, I said, makes decisions. It takes a set of available acts and returns that set ordered for choice-worthiness. Each available act in a decision will, if performed, lead to a particular outcome. These outcomes are precisely the consequences with which a consequentialist theory is concerned: a consequentialist moral theory orders the available acts in a decision for choice-worthiness in that decision, based solely on the outcomes of those
acts. (Remember that I have limited my attention for the time being to decisions under certainty; otherwise, the consequentialist focus might better be thought of as the prospects of acts.) A moral theory that takes anything other than the outcomes of available acts into account is not a consequentialist theory. Here is my definition of consequentialism.

**Consequentialism**: The choice-worthiness relation in any decision, $D_1$, is fully determined by the available outcomes in $D_1$.

A consequentialist theory is a moral theory according to which consequentialism is true. How can we distinguish theories according to which consequentialism is true, from theories according to which it is not true? I propose that a consequentialist theory can be represented as making decisions between available outcomes, rather than between available acts each of which will lead to an outcome; and that a non-consequentialist theory cannot be represented in this way. Take any set of outcomes, say $\{A, B, C\}$. A consequentialist theory will order any set of three acts, $\{\text{act-leading-to-}A, \text{act-leading-to-}B, \text{act-leading-to-}C\}$, for choice-worthiness in the same way, regardless what other features those acts may have. Such a theory can be represented as ordering the set of outcomes $\{A, B, C\}$ directly, rather than as ordering each of the corresponding sets of acts. A non-consequentialist theory cannot be represented this way, for according to a non-consequentialist theory something other than the outcomes of the available acts plays a role in determining the choice-worthiness of the acts in a decision. A non-consequentialist theory will not order every set of three acts, $\{\text{act-leading-to-}A, \text{act-leading-to-}B, \text{act-leading-to-}C\}$, in the same way (at least, it will not do so for some sets of outcomes).\(^5\)

I want to constrain my attention to consequentialist theories. Consequentialism is a defining feature of a general utilitarian theory. Talk of acts is superfluous to the presentation of consequentialist theories. So, I now propose

\(^5\)What other features might acts have, outside the outcomes they will lead to (and their prospects, which I have set aside for the time being)? It is not entirely clear that there are any. It is possible that features of acts that have traditionally been associated with non-consequentialist moral theories can in fact be thought of as features of outcomes. One such feature is who faces a decision, or performs the chosen act. Call a moral theory according to which choice-worthiness can depend on who faces the decision (according to which the identity of the decision-maker matters morally) an agent-relative theory. It is natural to think that two different agents can face decisions between acts leading to the very same outcomes. However, if facts about which decisions have been faced by which agents are part of outcomes (which it seems they might be), then it is not obvious that agent-relativity is incompatible with consequentialism. Perhaps all moral theories can be ‘consequentialised’. For more on this debate see Vallentyne 1988, Dreier 1993, Louise 2004, and Brown 2011b. I’ll address a related issue in Section 5.2.2.
to simplify my discussion by dropping acts out of the picture. Strictly speaking, a decision is a set of available acts; each available act leads to a particular outcome; and a consequentialist theory orders the available acts in a decision for choice-worthiness in that decision on the basis of the outcomes of those acts. However, I shall now proceed as though a decision were a set of available outcomes, and a consequentialist theory ordered the available outcomes in a decision for choice-worthiness in that decision. This allows me to leave out a lot of instances of phrases like ‘the act that would lead to’. For example, ‘the decision between an act that would lead to one outcome and another act that would lead to another outcome’ can become ‘the decision between one outcome and another’. I shall talk of deciding which outcome to bring about, rather than which act to perform. Sometimes, for elocutionary simplicity, I’ll still talk in terms of deciding what to ‘do’. This should be read as synonymous with talk of deciding which outcome to bring about.

Here is CFPU represented as making decisions between outcomes. From now on I refer to this statement of the theory unless I indicate otherwise.

\[ \text{CFPU, restated: For any fixed population decision, } D_1, \text{ one available outcome, } A, \text{ is at least as choice-worthy in } D_1 \text{ as a second available outcome, } B, \text{ just in case the total wellbeing of the population of } A \text{ in } A \text{ is at least as high as the total wellbeing of the population of } B \text{ in } B. \]

An extension of CFPU only counts as a general utilitarian theory if it is a consequentialist theory. A consequentialist theory can be represented as ordering available outcomes for choice-worthiness in decisions. CFPU itself is a consequentialist theory, and consequentialism is a central part of utilitarianism. Now I’ll further constrain the class of general utilitarian theories by reference to the role of axiology.

2.2.2 Axiology

A general utilitarian theory should be an axiological consequentialist theory. An axiological theory is a moral theory according to which what one ought to do is determined by the goodness of the available options; according to which choice-worthiness is determined by goodness. In the consequentialist framework I’ve just adopted this amounts to the claim that the choice-worthiness in
any decision of the available outcomes in that decision is determined by the
goodness of those outcomes. It is central to the utilitarian idea, I think, that
the choice-worthiness ordering in any decision must be explicable in terms of
the goodness of the available outcomes. If, say, \( A \) is more choice-worthy in a
decision between \( A \) and \( B \) than \( B \), it must be that \( A \) is better than \( B \) in some
morally relevant way. Why, according to CFPU, should I increase someone’s
wellbeing, other things being equal? Because the outcome in which that person
has higher wellbeing is better in a morally relevant way, than the outcome in
which she has lower wellbeing. (Whether morally relevant goodness is a mat-
ter of total wellbeing, average wellbeing, or something else is in part what this
dissertation is about.)

Before I try to spell out what it means to be an axiological consequentialist
theory more precisely, let me pause to make some assumptions about goodness.
First: I conduct my discussion of goodness primarily in terms of relations of
betterness or worseness. I shall be concerned with when, and by how much,
one outcome is better or worse than another. I’m not sure whether there is
anything to goodness beyond these comparative claims. Can we say how good
something is, as well as how much better and worse it is than other things? I’m
not sure; but I assume that we don’t need to. I assume that I can capture all of
the necessary facts about goodness in terms of betterness relations.\(^6\)

I leave open that outcomes might be better and worse than one another
in various respects. My focus will be on any respects of betterness that are
morally relevant, but there might also be respects of betterness that are not
morally relevant. There might be facts about betterness with respect to beauty,
for example, which I think probably aren’t morally relevant (except in so far as
they affect people’s wellbeings). Each respect of betterness can be represented
as a two-place betterness (or worseness) relation. I assume that each betterness
relation is structured as many people intuitively think they are.\(^7\) For each
relation (or respect) of betterness, I assume that the following claims are true.
Betterness is transitive. If one thing is better in some respect than a second

\(^{6}\)See Broome 2004 on this issue. Some people think that an outcome can be simply good or
bad, without being (necessarily) better or worse than other outcomes. I’ll discuss this possibility
in Section 3.2, but until then I assume that betterness is all we need.

\(^{7}\)Perhaps best known for defending these ‘standard’ structural constraints on betterness relations
is Broome (1991 and 2004). See also Arrhenius 2016 and Thomas 2015. People who have proposed,
or at least considered, the rejection of these constraints include Temkin (2011, 1987), Chan
(2003), and Rachels (2004).
and the second is better in that same respect than a third, then the first thing is better in that respect than the third. Betterness is irreflexive. No thing is better in any respect than itself. Betterness is anti-symmetric. If one thing is better in some respect than a second, then the second is not better in that same respect than the first. I do not assume that any betterness relation is complete. Sometimes, when one thing is not better in a particular respect than a second and the second is not better in that respect than the first, the two things are equally good in that respect. However, it may be that two things are neither better than one another in some respect, nor equally good in that respect. In that case I will say that the two are incomparable in that respect.

Worseness and betterness are inter-definable, as are the notions of being equally as good as, at least as good as, and at least as bad as (and, as we have seen, incomparable). It is easiest to show how these notions are related by starting with a relation of being at least as good, as primitive. (Alternatively, one could start with the relations of being better than and equally as good as.) Label all of the respects of betterness, \( r_1, r_2, r_3, \ldots \) Then we can represent betterness in any respect, \( r \), in terms of the relation of being at least as good in that respect, which I’ll represent like this: \( \succeq_r \). We can use this relation to define the others: better than in respect \( r \) (\( \succ_r \)), worse than in respect \( r \) (\( \prec_r \)), equally as good as in respect \( r \) (\( \equiv_r \)), and at least as bad as in respect \( r \) (\( \preceq_r \)).

Here is how those definitions proceed. \( A \) is better than \( B \) if \( A \) is at least as good as \( B \) and \( B \) is not at least as good as \( A \); that is, \( A \succ_r B \) if \( A \succeq_r B \) and \( B \not\succeq_r A \). \( A \) is worse than \( B \) if \( B \) is at least as good as \( A \) and \( A \) is not at least as good as \( B \); that is, \( A \prec_r B \) if \( B \succeq_r A \) and \( A \not\succeq_r B \). These are sometimes called ‘strict’ betterness and worseness. \( A \) is equally as good as \( B \) if \( A \) is at least as good as \( B \) and \( B \) is at least as good as \( A \); that is, \( A \equiv_r B \) if \( A \succeq_r B \) and \( B \succeq_r A \). \( A \) is at least as bad as \( B \) if \( B \) is at least as good as \( A \) and \( A \) is not at least as good as \( B \); that is, \( A \preceq_r B \) if \( B \succeq_r A \) and \( A \not\succeq_r B \).

We have already encountered some betterness relations. The wellbeing scale, for example, can be represented as a betterness relation among lives. A life higher on the scale is better to live (better for the person who lives it) than a life lower on the scale. Choice-worthiness orderings can also be represented as betterness relations. A moral theory provides a choice-worthiness ordering for each decision, each of which represents the relation of being better to choose when faced with that decision. Each of these relations is standardly structured
(and, for all of the theories that I consider, complete). This means, for example, that if \( A \) is more choice-worthy in \( D_1 \) than \( B \) (if it would be better, faced with \( D_1 \), to bring about \( A \) than to bring about \( B \)), and \( B \) is more choice-worthy in \( D_1 \) than \( C \), then \( A \) must be more choice-worthy in \( D_1 \) than \( C \). These facts do not constrain the betterness relation, choice-worthiness in \( D_2 \). It might well be that \( C \) is more choice-worthy in \( D_2 \) than \( A \). This would be an instance of an interesting kind of menu-dependence of choice-worthiness: \( A \) is more choice-worthy in one decision than \( C \), and \( C \) is more choice-worthy in another decision (with a different ‘menu’ of available outcomes) than \( A \).\(^8\)

So much for goodness and betterness relations. Now I can return to the task of defining axiological consequentialism, and identifying axiological consequentialist theories. I characterised axiological consequentialism as the view that the choice-worthiness in any decision of the available outcomes in that decision is determined by the goodness of those outcomes.\(^9\) Here is my final definition.

**Axiological consequentialism:** The choice-worthiness relation in any decision, \( D_1 \), is fully determined by the morally relevant betterness relations among the available outcomes in \( D_1 \).

An axiological consequentialist theory is a moral theory according to which axiological consequentialism is true.\(^10\) How can we distinguish theories according to which axiological consequentialism is true, from theories according to which it is not? I propose that an axiological consequentialist theory should have two parts. It should identify: a) the morally relevant betterness relations among all possible outcomes; and b) the method for using those relations to establish the choice-worthiness of the available outcomes in each possible decision. I’ll call an account of all of the morally relevant betterness relations among outcomes an axiology. An axiological consequentialist theory consists of an

\(^8\)The theory that I endorse has this feature, as does one of the theories that I introduce in the final section of this chapter. I address some possible objections to menu-dependent choice-worthiness in Section 5.1.

\(^9\)This is supposed to be fairly uncontroversial, agreeing for example with Oddie and Milne: ‘According to the axiologist the value concepts (good, evil, better than and so on) are basic, and the deontic concepts (permission, obligation and so on) are in some sense derivative. Briefly, the moral or deontic properties of an act are determined by the value of the act’ (1991, 42). See also Stocker’s ‘axiological credo’, ‘Rightness depends solely on goodness’ (1973, 87).

\(^10\)There may also be axiological non-consequentialist moral theories; theories according to which choice-worthiness is fully determined by betterness relations among something other than outcomes.
axiology and a normative component: a procedure or function that determines choice-worthiness in decisions from the axiology.

My claim is that any utilitarian theory should be axiological and consequentialist, and so able to be represented by an axiology and a normative component. So far, I haven’t represented CFPU in this way. I haven’t identified the morally relevant betterness relations among outcomes according CFPU. One way of doing this is, I think, fairly obvious and simple. It would not misrepresent the theory to say that according to CFPU there is only one morally relevant betterness relation, which we might call betterness with respect to total wellbeing and identify by the following condition: A is at least as good with respect to total wellbeing as B iff the total wellbeing of the population of A in A is at least as high as the total wellbeing of the population of B in B. If this is the only relevant betterness relation, the normative component of CFPU says that axiology determines choice-worthiness in fixed population decisions straightforwardly: in any fixed population decision, available outcome A is at least as choice-worthy as available outcome B iff A is at least as good as B in the morally relevant respect. However, this is just one of the combinations of axiology and normative component that produce (or are equivalent to) CFPU. In looking for an axiological consequentialist extension of CFPU we will discover many other combinations. I have identified CFPU by its choice-worthiness orderings. I want to leave open, for now, which axiology and normative component best explain the theory.

2.2.3 The overall betterness requirement

A general utilitarian theory must be an axiological consequentialist theory. Both axiology and consequentialism have been associated with what I will call the ‘overall betterness requirement’. However, I do not wish to require a general

\[\text{On consequentialism see in particular Brown 2011b. Brown talks of ‘betterness’ rather than ‘choice-worthiness’, but I think that in my terminology the best representation of his claim is that a consequentialist theory provides a single choice-worthiness ordering of all possible outcomes (rather than a choice-worthiness ordering for each decision, as I propose). An alternative reading of his claim is that there is a single overall betterness relation among outcomes, which straightforwardly determines choice-worthiness in each decision. On axiology see Carlson 1998, Broome 2004, Arrhenius 2016, and Thomas 2015, who all define an axiology as (an account of) a single betterness relation. I do not mean to indicate that the overall betterness requirement is universally made. McDermott does not make it. Roberts also rejects it, I think, as here: ‘According to the traditional account of the basic maximizing idea, what agents must do is create the most aggregate good that they can. According to this less conventional account, what agents must do is create the most good that they can for each person’ (Roberts 2002, 316).} \]
utilitarian theory to meet this requirement. Before I introduce the final defining feature of a general utilitarian theory (welfarism), let me explain why I don’t want to endorse the overall betterness requirement.

First, let me explain what the overall betterness requirement is and why it might seem appealing. The idea is roughly that according to an axiological consequentialist theory (whether as a defining feature of axiology or consequentialism) there must be a single overall betterness ordering of outcomes, which either is, or straightforwardly determines, a choice-worthiness ordering of all possible outcomes in any decision. This is an attractive, simple way of representing the idea that goodness determines choice-worthiness. We might arrive at it by something like the following thoughts. Some outcomes are better than others. It might be that there are various respects in which one outcome can be better than another, but these various respects must be able to be combined somehow—balanced against each other, traded off, perhaps weighted—to establish a relation of overall betterness. We must be able to arrive at a conclusion about how any given pair of outcomes compare overall, or all things considered. And this final, overall betterness is the kind of betterness that determines how we ought to make decisions. It either is choice-worthiness, or (if we want to maintain the conceptual distinction between the axiology and the normative component of a theory) it determines choice-worthiness straightforwardly: the better an outcome is overall, the more choice-worthy it is (in any decision).\footnote{The way that overall betterness determines choice-worthiness must be somewhat more subtle if the overall betterness relation is (allowed to be) incomplete. For example, a common approach is to say that in any decision, two available outcomes that are overall incomparable and neither overall worse than any other available outcome are equally choice-worthy. For an illustration of this kind of connection between incomparability and choice-worthiness in population ethics see Rabinowicz 2009.}

The overall betterness requirement is stronger than the requirement of axiological consequentialism that I make. It rules out candidate general utilitarian theories that I allow. I allow that there might be multiple morally relevant betterness relations, which cannot be combined to produce an overall betterness relation, but determine choice-worthiness in decisions in some other way. Although I see the appeal of the overall betterness requirement, and agree that it represents the structure of some axiological consequentialist theories well (including CFPU), I think that it is a mistake to accept it in the context of my project. I am trying to extend CFPU to variable population decisions consistently with the utilitarian spirit. Part of that spirit, I think, is the idea...
that what determines choice-worthiness is betterness among the available outcomes. However, I don’t think that it is essential to that spirit that what determines choice-worthiness is some single relation of (overall) betterness among outcomes. This is the first reason I have defined axiological consequentialism fairly weakly.

A second, related reason for resisting the overall betterness requirement is that it has been shown convincingly to be incompatible with very compelling intuitions about variable population decisions. In the next chapter I identify four intuitively compelling conclusions about variable population decisions, and adopt the goal of developing a general utilitarian theory that agrees with these conclusions. I am not the first to find these conclusions compelling, and it has been shown that they cannot be accommodated within any moral theory according to which choice-worthiness is straightforwardly determined by a single betterness relation. In the face of these impossibility results, a lot of effort has gone into deciding which counterintuitive theory is less counterintuitive than the others and thus ought to be accepted. I think that these results provide a powerful reason rather to revisit our structural assumptions, and in particular to reject the overall betterness requirement. In Chapter 4 I propose a particular theory, Shortfall Utilitarianism (SU), which is an axiological consequentialist theory according to my definitions. In that sense, it is a theory according to which choice-worthiness is determined by betterness among outcomes. It also agrees with those four compelling variable population conclusions. It does not meet the overall betterness requirement. I think this strengthens the reason to reconsider that assumption.

At this point I would like to respond more directly to the reasons for which others do endorse the overall betterness requirement. Unfortunately, it’s not always clear what people take their reasons to be. Sometimes, the overall betterness requirement seems to be adopted because it is a fairly intuitive spelling out of the rough axiological consequentialist idea, and an effective means of curtailing the options. Those seem to me to have been acceptable, even good, reasons in the past. It is sensible to adopt a fairly narrow framework in the early stages of understanding the problem. However, as I noted earlier, sometimes the required betterness relation is not explicitly called an overall betterness relation. In some cases, that label might reasonably be disputed. However, the important point here is that these theorems show that there is no consequentialist axiological theory that employs only a single betterness relation and is compatible with certain compelling intuitions.
stages of exploring an area as complex as population ethics. And the strategy has been very useful. It is only because people adopted the overall betterness requirement that we now know that it is inconsistent with particular axiological and normative intuitions. However, having established these inconsistencies, I don’t think that these are any longer good reasons for maintaining the overall betterness requirement. At other times it seems to be assumed that the overall betterness requirement is not only a good first attempt, but the most permissive possible spelling out of the idea that betterness among outcomes determines choice-worthiness. I don’t think that this is the case. I think that the axiological consequentialist constraint I have adopted is another good spelling out of this idea.\textsuperscript{14}

One possible explanation for the widespread endorsement of the overall betterness requirement is that it can be difficult to imagine how betterness of various kinds could determine choice-worthiness, other than via an overall betterness relation. This would also help to explain why reasons for adopting the requirement are rarely provided. There is some textual evidence for this explanation, in the way that people make claims about ‘goodness’ or ‘betterness’ that fairly clearly relate to overall goodness or betterness, without making this clear. Broome, admirably clear in much of his discussion of goodness, is sometimes guilty of this I think. Broome does occasionally explicitly recognises his focus on overall goodness, as here in *Weighing Lives*:

> The problem I have set myself in this book is to compare distributions of wellbeing and look for principles that determine which is better than which. ... This can be thought of as a problem of aggregation: how is wellbeing aggregated across a grid? ... How does the wellbeing of different people together determine overall good? (2004, 16).

At other times, however, he seems to slide into the view that he has exhausted the possibilities concerning goodness without qualification, saying for example ‘This book is about goodness. If there are ethical considerations that are not to do with goodness—purely deontic rules, for example—they are beyond this book’s scope’ (2004, 145).\textsuperscript{15} Similarly, when Arrhenius says that ‘Population

\textsuperscript{14}I address some other possible objections to the lack of an overall betterness relation among outcomes in Section 5.2.

\textsuperscript{15}A terminological note: Broome associates the overall betterness requirement with the term
axiology concerns how to evaluate populations in regard to their goodness, that is, how to order populations by the relations “is better than” and “is as good as” (2009b, 23), it is clear that he means something like overall goodness, though he doesn’t make this explicit or acknowledge an alternative.

The route via overall betterness is, I don’t dispute, a very natural way to attempt to combine betterness in a number of respect to make decisions. I soon introduce some moral theories that use respects of betterness to determine choice-worthiness in other ways. Before I do that, let me give an analogy to illustrate how this might happen, and why I think we should resist the overall betterness assumption from the outset, however natural it may be.

An analogy

Imagine you’re trying to develop a procedure for deciding which of your local beaches to go to. You have five local beaches: Bondi, Tamarama, Bronte, Gordon’s Bay, and Coogee. On some occasions you can choose among all of them; on others you have to choose from a subset (say because you only have time to get to a close one, or because your friend has proposed a shortlist). You want to develop a method for ordering any given subset of the five beaches for what we might call go-worthiness—how good it would be for you to go to that beach rather than one of the other available ones. You want these orders to be based only on how good the beaches in the relevant subset are. You want an axiological beach decision theory.

One part of such a theory must be an axiology of the local beaches: an account of the relevant betterness relations among those beaches. Which features of beaches make them better or worse than one another in ways that are relevant to how you should decide which beach to go to? Imagine you think for a while about this question, and come up with the hypothesis that there are three relevant features, or respects: conditions in the ocean, conditions on the sand, and conditions in the pool. You also develop a betterness relation to represent each of these respects: call them the relations of ocean-betterness, sand-betterness,
and pool-betterness. Each relation has sufficient structure that it can also be represented on a scale of the integers 1-5, and in such a way that a difference of one unit in one of the three respects matters (for go-worthiness) just as much as a difference of one unit in another respect. That beach $A$ is ocean-better than beach $B$ by 2 has just as much impact on go-worthiness as the fact that beach $B$ is pool-better than beach $A$ by 2 (though these facts will push, so to speak, in opposite directions).

So far so good, I hope. Now, let’s focus on the details of the hypothesised relations. Each beach has ocean conditions and sand conditions, so the relations of ocean-betterness and sand-betterness are complete. For each pair of beaches, you’re able to say either that one is better than the other (and if so by how much) or that they are equally good, in each of these respects. You’re able to assign an ocean-rating to each beach, and a sand-rating to each beach (based on the relevant 1-5 scales). However, only some beaches have pools, and thus pool conditions. Bondi, Bronte, and Coogee have pools, but Tamarama and Gordon’s Bay don’t. To the question, ‘How are the conditions in the pool at Tamarama?’ there is no meaningful answer. Still, there is a pool-betterness relation among Bondi, Bronte, and Coogee. You are able to assign a pool-rating to each of these three beaches. Imagine these three respects of betterness are summarised in this table.

<table>
<thead>
<tr>
<th></th>
<th>Ocean</th>
<th>Sand</th>
<th>Pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bondi</td>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Tamarama</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Bronte</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Gordon’s Bay</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Coogee</td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

This seems on the face of it to be a plausible axiology. I see no reason to think that all of the respects of betterness among some set of alternatives must be complete. In this case, it seems most natural to me to think that Tamarama and Gordon’s Bay are neither better nor worse with respect to pool conditions than any other beach. They are incomparable, in that respect, with every beach. However, this kind of incompleteness makes trouble for the task of combining

---

17 These relations might be irreducible, or they might in turn be produced by some sub-factors, weighed against each other. For example, ocean-betterness may be determined by things like swell size, underwater visibility, and the prevalence of rips.
these three respects of betterness into an overall betterness relation. Or, another way of seeing the problem, forcing these respects of betterness into an overall betterness relation constrains the way we might use them. First let me suggest a way of using the respects of betterness to make decisions directly, and then I’ll explain why this way of using respects of betterness can’t be captured by an overall betterness relation.

Here is my proposal. Whenever you face a decision between beaches, you should make a series of pairwise comparisons to find out, for each available beach, how much (strictly) worse that beach is, in any of the relevant ways, than a single other available beach. Imagine your motivation in choosing between beaches is to avoid going to a beach that is worse in any of the three relevant ways, than another beach you might have gone to. (You might think of this as a kind of regret-avoidance strategy.) So, when you have a subset of beaches to choose from, you should consider each beach in turn and ask, ‘Is this beach worse than another of the beaches in the set in any of the three ways; and if so, by how much in total? What is the greatest total amount by which this beach is strictly worse than one other available beach?’ This will tell you how much reason you have, based on the goodness of the available beaches in the set, not to go to that beach. Having done this for each beach in the set, you know how much reason you have not to go to each beach in the set, which gives you the order of go-worthiness for that decision—the less reason you have not to go to a particular beach, the more go-worthy it is in that decision.

Consider the decision between going to Bondi or Bronte. Here is a representation of that decision.

<table>
<thead>
<tr>
<th></th>
<th>Ocean</th>
<th>Sand</th>
<th>Pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bondi</td>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Bronte</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Bondi is sand-worse than Bronte by 2, and pool-worse than Bronte by 2. Bronte is ocean-worse than Bondi by 2 and not worse in any other way. These being the only available beaches, in this decision, Bondi is worse than a single other available beach by 4 in total and Bronte is worse than a single other available beach by 2 in total. In this decision things tell against Bondi more than they tell against Bronte, so you should go to Bronte.\(^\text{18}\)

\(^{18}\)It is significant that this decision procedure is equivalent to maximising the sum of the scores (or ‘going via overall betterness’) where every available option gets a score in each respect, as in
CHAPTER 2. GENERAL UTILITARIAN THEORIES

How does this approach apply to decisions involving incomparability? Imagine you face the decision between going to Bondi or Tamarama. Here is that decision represented in table form.

<table>
<thead>
<tr>
<th></th>
<th>Ocean</th>
<th>Sand</th>
<th>Pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bondi</td>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Tamarama</td>
<td>4</td>
<td>4</td>
<td>-</td>
</tr>
</tbody>
</table>

Tamarama is ocean-worse than Bondi by 1, and Bondi is sand-worse than Tamarama by 1. Neither beach is pool-worse than the other. So Tamarama is worse than another available beach by 1 in total; likewise for Bondi. In this decision things tell against each beach equally, so the two are equally go-worthy.

This is one way to make decisions by appeal to respects of betterness directly (even when those respects are incomplete). The idea behind this approach is that Bondi and Tamarama, for example, are incomparable pool-wise. The fact that Bondi has a pool (with some particular conditions) and Tamarama has none does not tell in favour of going to Bondi rather than Tamarama, nor vice versa. Pool-ratings are only relevant to decisions between beaches that both have pools. You don’t mind whether you go to a beach with a pool or without; but if you go to one with a pool you’d rather one with a nicer pool.

This is the beginning of one possible explanation or justification of using betterness relations in this way. It would need elaborating. It might be based on your psychology: if you don’t encounter a pool you don’t think about pools, but if you do encounter a pool you are reminded of other pools you might have been at. So if you go to a beach without a pool you don’t miss it; but if you go to a beach with a worse pool rather than a better pool, you have a negative emotional reaction (perhaps regret, envy, or dissatisfaction).

Some may find this decision procedure implausible in the beach case. (Thanks to Matt Clark for very useful discussion of this point.) They may think that pool conditions of any kind must be better than no pool (so that any beach with no pool is pool-worse than any beach with a pool), or that ‘no pool’ must slot in somewhere else on the pool-rating scale. They may find the decision procedure more plausible in the case of judging a ‘battle-of-the-bands’. Say ‘Bondi’, ‘Tamarama’, ‘Bronte’, ‘Gordon’s Bay’, and ‘Coogee’ are the names of five bands; and instead of ocean-, sand-, and pool-betterness the relevant respects of comparison among the bands are guitar-, vocal-, and drum-betterness. Then the table above might represent how good each of the five bands are with respect to guitar, vocals, and drums; ‘Tamarama’ and ‘Gordon’s Bay’ being bands with no drummer. Plausibly, I think, it should not count against a band that they have no drummer, though it should count against them if their drummer is worse than another in the competition.

The reader may suspect that this explanation reveals a fault in my representation of the decision procedure. If this explanation is correct, it sounds as though what really matters is regret (or what have you), and pool, sand, and ocean conditions are only instrumentally valuable insofar as they...
lot like what Broome calls ‘conditional goodness’ 2004, 154–157. He tries to make sense of this notion but concludes that it is incoherent. I think that this is because he tries to make sense of it within the framework of overall goodness. I’ll explain why this is problematic shortly.

I don’t claim that this is the only way to make use of respects of betterness, nor even that it is always the best way. I certainly don’t claim that this is the best axiological beach decision procedure; such a claim would require a lot of support. My point is only that this is one possible, plausible axiological beach decision procedure. Bear with me while I apply it to one more decision. Then I’ll be able to explain why this plausible axiological beach decision theory makes trouble for overall betterness. Consider the decision between Tamarama and Bronte.

\[
\begin{array}{l|c|c|c}
D_6: & \text{Ocean} & \text{Sand} & \text{Pool} \\
Tamarama & 4 & 4 & \\
Bronte & 3 & 5 & 5
\end{array}
\]

This is quite like the decision between Tamarama and Bondi we just looked at. Considerations of ocean- and sand-betterness push in opposite directions, with equal strength; and the beaches are incomparable with respect to pool-betterness. Tamarama is worse than Bronte by 1 in total, and Bronte is worse than Tamarama by 1 in total. In this decision, thing tell against each beach equally and the two are equally go-worthy.

Now recall the three pairwise decisions we’ve just looked at. Tamarama and Bronte are equally go-worthy, in a decision between the two. Likewise, Tamarama and Bondi are equally go-worthy in a decision between the two. In a decision between Bronte and Bondi you should go to Bronte; Bronte is more go-worthy in that decision than Bondi. This means that go-worthiness cannot be represented as a single overall betterness relation. (Nor can it be directly determined by such a relation.) If Bronte is overall better than Bondi, it produce that thing. So it might seem that the more accurate way of representing this axiological decision procedure is in terms of an overall betterness relation between things like ‘Bondi-rather-than-Coogee-or-Tamarama’. I don’t think this particularly matters for my current goal, which is just to illustrate the possibility of using respects of betterness directly rather than forcing them into overall betterness. Also, I’ll address an analogous representational shift in the population ethics context in Section 5.2.2.

I’ve already indicated that one other option is to assign to beaches without pools the lowest pool-rating (1) and calculate overall betterness straightforwardly. This might be the right approach if you prefer (or it is better for you in whichever is the relevant way) to go to a beach with a pool than one without. Another option is to say that a beach with a pool and a beach without a pool are always overall incomparable and thus equally choice-worthy.
cannot be that Tamarama is overall equally as good as both Bronte and Bondi. Tamarama cannot have the same value as two things with different values.  

I have tried to illustrate what seems to me to be a reasonable approach to developing an axiological beach decision procedure. You might try to figure out what the relevant respects of betterness among beaches are, and how those respects of betterness determine which beach you should decide to go to. I’ve also tried to illustrate how this approach might lead you to an axiology and associated beach decision procedure that does not provide an overall betterness relation. If you assume that go-worthiness either is, or is straightforwardly determined by, overall betterness, you rule out decision procedures of this kind. You will be forced to revise your decision procedure—perhaps by disregarding conditions in the beach pools altogether, perhaps by accepting that it would be better to go to a beach with a pool than to one without, perhaps in some other way. You will be forced to admit that your intuitive idea of what it is that makes one outcome better than another in a relevant way is incorrect, because it cannot be fitted into the formal requirements of overall betterness.

I seek an axiological consequentialist moral theory. Some people have spelled out one or another of these characteristics in terms of the overall betterness requirement. I have spelled them out differently. I seek a moral theory according to which the choice-worthiness of the available acts in any decision is determined by relations of betterness among the outcomes of those acts. In this section I have tried to explain why I don’t endorse the overall betterness requirement. To summarise: if I assume that choice-worthiness either is, or is straightforwardly determined by, an overall betterness relation among outcomes, I might rule out some plausible axiological consequentialist moral theories. I see no reason to make that assumption, and numerous counterin-

---

22 This is analogous to Broome’s argument against his ‘neutrality intuition’ (2004, 156–157). Broome assumes an overall betterness relation, which leads him to conclude that betterness cannot have this structure. My claim here is that (incomplete) respects of betterness, and a decision procedure that appeals to them directly, can have this structure. More on this analogous problem for neutrality in Section 3.2. See also Rabinowicz 2009, who defends widespread overall incomparability; and Cusbert and Kath 2016, in which we argue that neutrality can be captured in a multi-dimensional consequentialist moral theory.

23 Here is this methodology endorsed in Broome 2004: ‘Goodness has a structure; for instance, the relation of betterness is transitive. A good way of clarifying our thinking is by working to bring it within this structure. For instance, if we think one thing \( A \) is better than another \( B \), and \( B \) is better than \( C \), but we do not think \( A \) is better than \( C \), we have more thinking to do’ (2004, 73). I also endorse these structural constraints, and I agree that we have more thinking to do. But I think one of the possibilities we should consider is that there are respects of betterness at play, which cannot be squeezed into an overall betterness relation.
2.2. FURTHER CONSTRAINTS

Intuitive conclusions are known to follow from doing so. Now let me introduce my final constraint on a general utilitarian theory: welfarism.

2.2.4 Welfarism

Recall the Cheesy Extension (CE) from Section 2.2. CE is a consequentialist axiological moral theory; a theory according to which the choice-worthiness relation in each decision is fully determined by the morally relevant betterness relations among the available outcomes in that decision. The reason CE doesn’t seem like a candidate general utilitarian theory is that it says that morally relevant goodness depends, in part, on the wrong features of outcomes. The axiology it posits is implausible. Cheese content of the world is intuitively intrinsically morally irrelevant, and in particular is intuitively irrelevant to the utilitarian goal. The utilitarian goal is roughly to make people well off—not to make the world cheesy. I have assumed that each person has a wellbeing in each outcome in which she exists, which represents how well off that person is in that world, or how good the life she leads in that world is for her. Any consequentialist axiological theory that captures the motivating utilitarian idea must be a theory according to which the axiology of outcomes (and thus choice-worthiness in each decision) is determined by people’s wellbeings.

Welfarist axiological consequentialism: The choice-worthiness relation in any decision, $D_1$, is fully determined by the morally relevant betterness relations among the available outcomes in $D_1$, which in turn are fully determined by people’s wellbeings in those outcomes.

A welfarist axiological consequentialist theory is a moral theory according to which welfarist axiological consequentialism is true.

I can now make another simplification. An outcome is a possible world, and possible worlds are (often) massive and complex. However, I am only interested in welfarist theories, and the only features of possible worlds that are relevant to welfarist theories are the wellbeings people have in those worlds. So, from now on I consider a possible world to be picked out by the people who exist in that world and the wellbeings they have in it; as wellbeing-assigned populations.\(^{24}\) An outcome, then, can be identified by a set of pairs: people and their wellbeings. For example the outcome in which only Michael,

\(^{24}\)Broome and others call these ‘distributions’ of wellbeing.
Vincent, and Duncan exist, each with wellbeing 20, might be represented like this: \{Michael, 20; Vincent, 20; Duncan, 20\}. Strictly speaking, there might be a large number of possible worlds captured by this description—worlds in which only Michael, Vincent, and Duncan exist, each with wellbeing 20. However, a welfarist axiological consequentialist theory will treat each such world the same way. We might say that according to such a theory outcomes with the same wellbeing-assigned populations are morally identical. I hope that this is common ground for utilitarians (though it is, of course, controversial in a broader context).

This also means that the kind of partial representation of decisions that I introduced in Section 1.2.3 includes all of the information a welfarist axiological consequentialist theory might need. In fact those tables also included acts, which I no longer (in my consequentialist framework) need. Here is an example of the kind of representation of a decision that I will use from now on.

\[
\begin{array}{c|cccc}
D_7: & p_1 & p_2 & p_3 & p_4 \\
\hline
A & 10 & 10 & 10 & * \\
B & 8 & * & * & 12
\end{array}
\]

This table represents the decision between two possible outcomes, A and B. In A three people exist: \(p_1, p_2\) and \(p_3\), each with wellbeing 10. In B only two people exist: \(p_1\) with wellbeing 8, and \(p_4\) with wellbeing 12. I use a ‘∗’ to represent a person’s non-existence in a particular outcome. Also, all of the possible people not featured in the table do not exist in either outcome.

Welfarism is my final constraint on a general utilitarian theory. I now define a general utilitarian theory as a welfarist, axiological, consequentialist extension of CFPU. A general utilitarian theory is an extension of CFPU—it agrees with CFPU about fixed population decisions, but also makes variable population decisions. A general utilitarian theory is also consequentialist, axiological, and welfarist. This means that it gives a single order of choice-worthiness for each subset of welfare-assigned populations, on the basis of betterness relations among those populations. I do not endorse the overall betterness requirement.

### 2.3 Three contenders

I’ve now narrowed my focus somewhat. Still, there are many general utilitarian theories. In this section I introduce three that have been discussed extensively
in the population ethics literature. In the next chapter I will explore some of the well-known counterintuitive implications of these theories for variable population decisions. In Chapter 4 I will propose a fourth general utilitarian theory, Shortfall Utilitarianism (SU), which avoids these troubling implications.

Perhaps the most obvious, and certainly the most discussed general utilitarian theories, are Total Utilitarianism and Average Utilitarianism.\(^{25}\)

**Total Utilitarianism** (TU): For any decision, \(D_1\), one available outcome, \(A\), is at least as choice-worthy in \(D_1\) as a second available outcome, \(B\), just in case the total wellbeing of the population of \(A\) in \(A\) is at least as high as the total wellbeing of the population of \(B\) in \(B\).

**Average Utilitarianism** (AU): For any decision, \(D_1\), one available outcome, \(A\), is at least as choice-worthy in \(D_1\) as a second available outcome, \(B\), just in case the average wellbeing of the population of \(A\) in \(A\) is at least as high as the average wellbeing of the population of \(B\) in \(B\).

TU and AU are both extensions of CFPU: they agree with CFPU in fixed population decisions.\(^{26}\) That TU does so is obvious: TU is simply CFPU as I presented it above, without the restriction to fixed population decisions. That AU also agrees with CFPU in fixed population decisions should also be evident. For I mentioned earlier that CFPU can also be presented in terms of average, rather than total, wellbeings. AU is simply CFPU presented in that way, without the restriction to fixed population decisions. TU, AU, and CFPU all order the available outcomes in fixed population decisions for choice-worthiness in that decision according to the total (equivalently, according to the average) wellbeing of the population of the outcome: the higher the total (average) wellbeing, the more choice-worthy the outcome.

---

\(^{25}\)These are the two extensions that Sidgwick considers. He favours Total Utilitarianism, though curiously he presents it not in terms of total happiness (his discussion being in terms of happiness rather than wellbeing), but in terms of ‘the product formed by multiplying the number of persons living into the amount of average happiness’ (1907, 415–416).

\(^{26}\)The issue of the empty population arises here again (see footnote 2 on page 20). I said that the total wellbeing of the empty population is zero. I now stipulate that the average wellbeing of the empty population is zero too. Fans of AU, if there were any, might disagree with this representation of the theory, but this makes for the simplest version of the theory and I won’t use it against the theory in what follows.
CHAPTER 2. GENERAL UTILITARIAN THEORIES

TU and AU disagree in some variable population decisions. Consider, for example, the decision whether to bring into existence one ‘extra’ person who will have a wellbeing below the prior average but still positive (without changing anyone else’s wellbeing). Here is such a decision.

<table>
<thead>
<tr>
<th>D₈:</th>
<th>Anne</th>
<th>Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>*</td>
</tr>
</tbody>
</table>

According to TU one ought to bring this person into existence, because doing so would result in a higher total wellbeing. According to AU one ought not bring this person into existence, because doing so would result in a lower average wellbeing.  

The axiologies most naturally associated with TU and AU are simple, each consisting of a single betterness relation. According to TU, outcome A is at least as good as outcome B just in case the total wellbeing of the population of A in A is at least as high as the total wellbeing of the population of B in B. According to AU, outcome A is at least as good as outcome B just in case the average wellbeing of the population of A in A is at least as high as the average wellbeing of the population of B in B. TU and AU thus satisfy the overall goodness requirement: they can be represented in terms of a single, complete ‘overall’ betterness relation among outcomes, which either is or straightforwardly determines choice-worthiness in every decision.

TU and AU are not the only totalist and averagist general utilitarian theories. Another general utilitarian theory, which can be presented either in totalist or in averagist terms (just like CFPU itself), differs from TU and AU in ‘scope’—which people it takes to be morally significant. According to this other theory—call it Inevitable Utilitarianism (IU)—in any decision one ought to maximise the total (equivalently average) wellbeing of the set of people who exist in every available outcome in that decision. I mention this theory just to illustrate the range of alternatives. IU is an unappealing theory and I won’t spend any time discussing it here.

---

27 Clearly, it is significant for TU (as it is for some other general utilitarian theories) where the zero point in the wellbeing scale is set. I have so far refrained from addressing this issue but will do so in Section 3.1. At this point it is enough to note that wherever it is set, this difference between TU and AU will exist.

28 These theories can also be represented using more complex axiologies, and correspondingly more complex normative components.

29 For more on IU see Arrenius 2016, 282.
The third promising general utilitarian theory I want to introduce is Harm Utilitarianism (HU). HU is what is known as a ‘person-affecting’ theory.\(^{30}\) A person-affecting theory takes effects on individuals, rather than features of populations or aggregation within groups of people, to be the primary moral concern. HU cashes out ‘effects on individuals’ by way of a particular notion of harm. Let me now define a harm, as it is employed in this theory.\(^{31}\)

\[\text{Harm: For any decision, } D_1, \text{ person, } p_1, \text{ and available outcome in } D_1, A; \text{ } p_1 \text{ incurs a harm of amount } x \text{ in } A \text{ just in case the maximum wellbeing of } p_1 \text{ in any available outcome in } D_1 \text{ is greater than } p_1's \text{ wellbeing in } A \text{ and the difference between these two wellbeing is } x.\]

So, harms are relative to a decision. A person incurs a harm in a particular outcome in a particular decision iff they have lower wellbeing in that outcome, than they have in some other outcome in the decision. We can also say that I harm someone if, facing some decision, I choose an available outcome in which that person has lower wellbeing rather than an available outcome in which she has higher wellbeing.

According to HU one ought to minimise harms. Here is HU spelled out more precisely.

\[\text{Harm Utilitarianism (HU): For any decision, } D_1, \text{ one available outcome, } A, \text{ is at least as choice-worthy in } D_1 \text{ as a second available outcome, } B, \text{ just in case the total harm incurred by the population of } A \text { in } A \text{ is at least as low as the total harm incurred by the population of } B \text{ in } B.\]


\(^{31}\)There is a significant literature debating the virtues of various definitions of ‘harm’ (and ‘benefit’)—see Bradley 2012a, Bradley 2012b, Hanser 1990, Hanser 2008, Hanser 2009, Hanser 2011, Harman 2004, Roberts and Wasserman 2009, Shiffrin 1999, and Thomson 2011. Some people take as basic the claim that harms—whatever they are—are morally significant. From this starting point, the question of how to define a harm is also morally significant. I don’t think that this is a promising approach to population ethics. I grant that, intuitively, harms are morally significant. However, (and whatever the virtues of conceptual analysis in general) I don’t think that we should assume that the concept of harm produced by people who have focussed almost exclusively on fixed population decisions will track morality in variable population decisions. My definition of a harm here is stipulative. The question is then whether the thing thus defined is the (only) morally significant thing—a question we would have to ask whatever the definition given.

\(^{32}\)HU is very like the theory that McDermott and Roberts endorse. However, they do not assume that people have wellbeings only in outcomes in which they exist. I’ll return to this assumption,
Like TU and AU, HU is an extension of CFPU. Here is one way to see that this is the case. First, imagine a fixed population decision such that in one of the available outcomes everyone has their maximum wellbeing—that is, the wellbeing of every person in the population of that outcome is at least as high in that outcome as it is in any other available outcome in the decision. That will be the outcome with the highest total wellbeing, and the outcome with the least (zero) harm. According to both HU and CFPU it is the most choice-worthy outcome in that decision. Each other (non-identical) outcome in the decision will have a lower total wellbeing, and some amount of harm. The amount of harm on each outcome is the amount by which the total wellbeing of that outcome is lower than the total wellbeing of the most choice-worthy outcome. So the order of outcomes from highest total wellbeing to lowest is the same as the order of outcomes from lowest total harm to highest (and can be represented on the same scale).

What about fixed population decisions that are not like this? Decisions in which there is no available outcome in which everyone has their maximum wellbeing? In these decisions, we can use a similar line of reasoning. Here is the more general argument. Imagine any fixed population decision. Take the maximum wellbeing, in that decision, of each person who exists in the outcomes in that decision (we’ve allowed, now, that these maximum wellbeings may occur in different outcomes). Sum these wellbeings and call the total $M$. The total wellbeing of each available outcome in the decision must be less than or equal to $M$. Any available outcome with a total wellbeing equal to $M$ must be an outcome in which there is no harm (everyone has their total wellbeing). If there is such an outcome it is the (perhaps jointly) most choice-worthy outcome in the decision according to CFPU and HU. Any available outcome with a total wellbeing lower than $M$ must involve some harm, and in fact the difference between the total wellbeing of an outcome and $M$ must be the

and show how it matters for HU, in Sections 3.1 and 3.2. Here is how McDermott describes his theory (using ‘best’ where I would use ‘most choice-worthy’).

... the best alternative is the one which does least harm. More generally: for any two of the many alternatives, the better is the one which does less harm. ... The amount of harm done to [a person] is the difference between his happiness on the best alternative for him (the one on which he would have been happiest), and his happiness on the alternative I chose. If my choice affects many people, the total harm done is simply the sum of the harms done to all those harmed (1982b, 170–171).

McDermott provides another in 1982b, 171–173.
sum of the harms incurred in that outcome (the only way the total wellbeing of an outcome can be lower than $M$ is if someone on that outcome has a wellbeing lower than their maximum). So again, the order of available outcomes from highest total wellbeing to lowest must be the same as the order of available outcomes from lowest total harm to highest.

HU disagrees with both TU and AU in some variable population decisions. Consider $D_8$ again.

<table>
<thead>
<tr>
<th></th>
<th>Anne</th>
<th>Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_8$:</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>*</td>
</tr>
</tbody>
</table>

TU says that $A$ is more choice-worthy in $D_8$ than $B$. AU says that $B$ is more choice-worthy in $D_8$ than $A$. HU says that $A$ and $B$ are equally choice-worthy in $D_8$. In neither outcome is anyone harmed; the total harm incurred in each outcome is 0.34

HU is a general utilitarian theory but, unlike TU and AU, does not meet the overall betterness requirement. TU and AU can each be represented in terms of a single overall betterness relation among outcomes. We could say that, according to TU, $A$ is overall better than $B$ iff the total wellbeing of the population of $A$ in $A$ is higher than the total wellbeing of the population of $B$ in $B$. Analogously for AU. If we present things this way (and I think it is a fairly natural way to present them), then according to both TU and AU, choice-worthiness either is, or is straightforwardly determined by, overall goodness.

We cannot represent HU this way. HU does not produce a single overall betterness relation. I think that the axiology most naturally associated with HU consists of a number of morally relevant betterness relations, one for each possible person and each the relation of betterness for that person.35 According to HU each of these betterness relations matters morally, but they can’t be combined into an overall betterness relation. This is partly because each relation is incomplete. Any outcome on which a particular person, $p_1$, does not exist, is incomparable with every other outcome with respect to goodness-for-$p_1$. Whether anything counts against choosing some available outcome, $A$, in some

---

34 Again, note that HU is potentially quite a different theory if people have wellbeings in outcomes in which they don’t exist. For example, if Bob has wellbeing 0 in $B$, in which he doesn’t exist, is he harmed in $B$? I’ll return to the possibility of people having wellbeings when they don’t exist in Section 3.2.

35 We set out a theory like HU in this way in Cusbert and Kath 2016.
decision, $D_1$, depends on which other outcomes are available in that decision. Think about $D_8$ above. HU says that $A$ and $B$ are equally choice-worthy in $D_8$. It is tempting to think that this means that according to HU $A$ and $B$ are overall equally good. But then consider a different decision.

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>$B$</td>
<td>10</td>
<td>*</td>
</tr>
<tr>
<td>$C$</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

$A$ and $B$ are not equally choice-worthy in $D_9$ according to HU. In this decision, there is some harm in $A$, because in this decision $p_2$ does not have her maximum wellbeing in $A$. According to HU, $B$ and $C$ are equally choice-worthy in $D_9$, and $A$ is less choice-worthy in $D_9$ than both (and thus impermissible).\(^{36}\)

This is an example of an interesting kind of menu-dependence of choice-worthiness. According to HU the relative choice-worthiness ordering of a pair of outcomes in a decision depends on which other outcomes are available in that decision. I discuss the menu-dependence of choice-worthiness at length in Section 5.1. This feature in HU can be traced back to the definition of harm at play in the theory. Harms are defined relative to a decision. Whether a person incurs a harm in a particular outcome is a matter of how that outcome compares for that person, with the other outcomes in a decision. It should come as no surprise that choice-worthiness in a decision, which is determined by harms, is menu-dependent in this way.

Although it produces this kind of menu-dependence, HU seems to me to be a welfarist, axiological, consequentialist theory. It simply is, of course, according to the definitions of these terms that I have stipulated. What I mean, however, is that HU seems intuitively to have these characteristics; or that my intuitive understanding of these terms seems to include theories like HU. This

---

\(^{36}\)For a similar reason Broome claims that person-affecting theories cannot be accounted for axiologically (2004). Again, I think what he shows is that they cannot be accounted for in terms of overall betterness. I think Roberts argues something similar here:

The inconsistency argument interprets the person-affecting intuition as a vehicle for the ranking of alternatives in terms of their betterness, with the ultimate goal of assessing the permissibility of alternatives. I argued, however, that the intuition could instead be understood as ranking alternatives, for each existing and future person, in terms of their betterness-for-that-person—again, with the ultimate goal of assessing permissibility. Under this second interpretation, the inconsistency was avoided (2003, 35).
2.3. THREE CONTENDERS

is part of my motivation for defining consequentialism and axiology as I did in the previous section, and rejecting the overall betterness requirement. We have begun to see, through HU, that the class of general utilitarian theories is broader than the class of theories one is restricted to by adoption of the overall betterness requirement. This is not entirely novel. McDermott uses roughly this framework, and clearly thinks that HU is a ‘generalized utilitarianism’. 37 Roberts also champions the claim that a harm minimising theory should be thought of as a consequentialist theory (2002 and 2004). But others have excluded person-affecting theories like HU from axiological, consequentialist, or utilitarian categories.

Conclusion

I have now defined a general utilitarian theory. A general utilitarian theory is an extension of CFPU that is consequentialist, axiological, and welfarist. Such a theory need not produce an overall betterness ordering of outcomes. I’ve introduced three general utilitarian theories: TU, AU, and HU. As we will see, each of these theories conflicts with compelling intuitions in some variable population decisions. In the next chapter I endorse four intuitions about variable population decisions in particular, and show that none of TU, AU, and HU agrees with all four of these intuitions. (I also explore some possible adjustments of these theories, with the same result.) I then frame the question of how we as utilitarians ought to make variable population decisions, in terms of the quest for a general utilitarian theory that does agree with those intuitions (without implying anything more counterintuitive).

37 Of his version of HU he says, ‘it seems correct to call it consequentialist’ (1982b, 176). He also clearly thinks it is a version of utilitarianism.
Chapter 3

Variable population decisions

The four intuitions I endorse in this chapter are well-known, and widely (I don’t claim unanimously) shared.\(^1\) The intuitions relate to three issues known in the population ethics literature as ‘the Asymmetry’, ‘the non-identity problem’, and ‘the repugnant conclusion’. I’ll take the intuitions one at a time, in each section first introducing and motivating an intuition, and then testing the familiar general utilitarian theories for consistency with that intuition. We’ll see that none of those theories agrees with all four of the intuitions. In the next chapter I’ll propose Shortfall Utilitarianism (SU), a general utilitarian theory that does agree with all four. This tells in favour of SU itself, and also demonstrates the richness of the class of general utilitarian theories.

Of course, my list of four is not intended to be a complete catalogue of compelling intuitions about variable population decisions. I focus on four that I find particularly powerful, which seem to be widely shared, and have proven difficult to reconcile. These intuitions are of the more specific kind (whereas I characterised the class of general utilitarian theories by appeal to more general intuitions). At least as I spell them out here, they concern mostly particular two-way decisions (decisions in which only two outcomes are available). Occasionally, I also endorse intuitions about more complex decisions.

Some people have argued that, despite their appeal, we must reject one or another of these intuitions, because it is not possible to find a moral theory that agrees with all four—at least within some utilitarian, consequentialist, or

---

\(^1\) The goal of finding a theory that agrees with all four has been adopted by many before me. This is a version of Parfit’s pursuit of ‘Theory X’ (1984, 361). See also Ng 1989 and Sider 1991.
axiological constraints.\(^2\) Others reject one of these intuitions simply because they don’t find it intuitive at all. I do my best to motivate and explain my intuitions in this chapter, but I accept that some people may not share them. However, even someone whose intuitions conflict with mine may find the moral theory that I present in the next chapter appealing, or find some aspects of that theory useful. Its structure is quite flexible, and I will make some notes in the next chapter about how the theory might be adapted to reflect different intuitions.

3.1 Wretchedness

Think of the worst possible life: the life that is as bad for the person who lives it as it is possible for a life to be; the life with the lowest possible wellbeing (or one such life—there need not be a unique worst possible life). This might be a life full of pain and frustration, with little pleasure or autonomy; perhaps the life lived by someone born into tortured slavery, or with an aweful disease. Now imagine you face the decision whether or not to bring into existence a person who will, if she exists, live that life. (This is of course a variable population decision.) Imagine that no-one else is affected by the decision. My intuition is that you ought not create that person. It seems to me that it is worse, in some morally relevant way, that a person lives such a life than that they don’t live at all, and that this tips the moral balance against creating such a person. Creating that person is less choice-worthy than not creating them, in the decision between the two. Call a life of which this is true a normatively-wretched life (or just a wretched life, for short).\(^3\)

\((\text{Normatively-})\text{Wretched}\): A wretched life is a life such that, in any decision between creating a person who will, if created, live that life, and not creating that person, other things being equal, one ought not create that person.

Moral theories disagree on whether any possible lives are wretched. I have the intuition that some, but not all, possible lives are wretched. Lives at the


\(^3\)I am using the term ‘wretched’ here is a non-standard way, which is why I have attached the ‘normatively’ qualifier to it. More on this below, but it is worth emphasising from the outset that I define (normative-)wretchedness in terms of how one ought to make particular decisions.
lowest possible wellbeing are wretched. Lives at wellbeings just barely higher than the lowest possible wellbeing are surely also wretched. But I have the intuition that a life at the highest possible wellbeing is not wretched; and that it is not the case that you ought not create a person who will, if created, live a non-wretched life (in a two-way decision, other things being equal). Call a life of this kind (a non-wretched life), happy.

**Happy**: A happy life is a non-wretched life such that, in any decision between creating a person who will, if created, live that life, and not creating that person, other things being equal, it is not the case that one ought not create that person.4

My intuition is that whether a life is wretched or happy is a matter of how good that life is for the person who lives it. It is a matter of that person’s wellbeing. Some lives, including the worst possible lives, are wretched; some lives, including the best possible lives, are happy; and there must be some threshold wellbeing that divides the wretched from the happy lives. Call this the wretchedness threshold.

**Wretchedness threshold**: There is a wretchedness threshold iff there is some wellbeing level, \(x\), such that all and only lives below that wellbeing are wretched, and all and only lives at or above that wellbeing are happy; in which case, \(x\) is the wellbeing threshold.

My first intuition is that there is a wretchedness threshold. If there were no wretchedness threshold, this would mean one of three things. First, it might mean that no lives are wretched; that a life cannot be so bad for the person who lives it that one ought not create such a person. Second, it might mean that all lives are wretched; that every life is so bad for the person who lives it that one ought not create such a person. Second, it might mean that all lives are wretched; that every life is so bad for the person who lives it that one ought not create such a person. Third, it might mean that whether

---

4Further to the terminological note above: I take the terms ‘wretched’ and ‘happy’ from Parfit 1984, as have others before me. However, I define these terms with reference to normative moral claims, whereas they are often defined with reference to some wellbeing identified non-normatively. I prefer my way of setting things up because I think it makes clearer the distinction between the question of whether there are wretched and non-wretched lives, and the question of which lives are which. (I discuss some possible answers to the latter question shortly, but don’t endorse one in particular and rather set the question aside.) I don’t think there is any greater significance to this terminological shift. This setup has similarities with Wolf’s criterion for a life ‘just barely good’: ‘if you knew that any child someone else conceived would enjoy at least that level of well-being, you would not regard that person as having a reason (deriving from consideration of the child’s welfare) not to conceive a child’ (2004, 77).
a life is wretched or happy depends on something other than whether or not it is below a particular wellbeing. For example, there might be intervals of wretchedness, separated by intervals of happiness, along the wellbeing scale. I find each of these possibilities highly counterintuitive. There must, I think, be a wretchedness threshold.\footnote{Of the four intuitions endorsed in this chapter, this is probably the one on which there is widest agreement. People who I believe share my intuitions about wretchedness include: Parfit (1984, 2015a), Broome (2004), McDermott (1982b), Roberts (2011b), and McMahan (2009). Benatar (2006) disagrees, I think, and believes either that every possible life is wretched (granting the claim that every possible life contains some pain), or that wretchedness is not determined by wellbeing level (but rather by pain-content of lives). As far as I know, no-one has defended the claim that there is no wretchedness threshold because no possible lives are wretched (though this is the case according to AU, as I will show shortly).}

3.1.1 Which lives are wretched?

The requirement that there is a wretchedness threshold leaves open a lot of questions about where that threshold is; about which lives are wretched and which happy. As much as possible, I intend to leave these difficult questions open, but let me acknowledge and briefly address three kinds of question about the nature of the wretchedness threshold.

First: what are wretched lives like? In Section 1.2 I embraced agnosticism about what wellbeing consists in, or what makes one life go better than another. This is one reason I won’t say very much about what a wretched life is like, or what features it has. I have assumed that wretchedness is a matter of wellbeing, that a wretched life is a life that falls below some threshold wellbeing. But what makes a life bad for the person who lives it? Is it a lack of pleasure? A surfeit of frustrated desires? A low level of autonomy? Does it matter how these features are distributed temporally through a life? As I said in Section 1.2, I will not defend a particular account of wellbeing (neither momentary nor lifetime). Nor do I think that it would be helpful to do so here. It is preferable, I think, to largely set aside the question of what wretched lives are like, for the purposes of developing a general utilitarian theory. Still, it is probably safe to say that lives of unrelenting pain and torture are among the worst possible lives, and so likely fall below the wretchedness threshold. The reader should feel free to plug in her favourite account of, or intuitions about, wellbeing, where that would be useful.

The second question is: where on the wellbeing scale is the wretchedness threshold?
threshold? I don’t mean which number should represent the wretchedness threshold. (I’ll get to that next.) I mean, looking down the list of lives ordered by wellbeing, when do they start to be wretched? Deciding on an account of wellbeing, and even assigning each possible life a particular numerically represented wellbeing, still wouldn’t settle the matter of which lives are wretched and which happy. Having done this (as I assumed we could in section 1.2), the question remains as to where on the wellbeing scale the wretchedness threshold lies. This is a substantive moral question. You and I might agree about the wellbeing ordering of all possible lives, and agree that there must be a wretchedness threshold, and yet disagree about where that threshold is. This would be a disagreement about when it is that one ought not create a person. Of some particular wellbeing, you might think that one ought never create a person with that wellbeing (in a two-way decision, other things being equal), and I might think that this is not the case.

I intend to leave this question open too. But I will consider in some detail two candidate proposals of where the wretchedness threshold is, to illustrate the debate—for it may not be obvious what such a proposal might look like, or what kind of arguments might be given for or against any particular proposal. First, consider the ‘non-existence comparative account’ of the wretchedness threshold. This account relies on being able to compare how good it would be, for a particular person, a) to live a given life, and b) not to exist at all. There is some wellbeing, it is claimed, at and above which life is at least as good as non-existence; below which life is worse than non-existence. This is the wretchedness threshold. Wretched lives are all and only those lives that are worse than non-existence, for the person who lives them.\(^6\)

\textit{The non-existence comparative account:} Wretched lives are all and only those lives such that it is worse for a person that she lives a life at that wellbeing than that she does not exist at all.

To some people the non-existence comparative account has the ring of truth,

\(^6\)See especially Adams 1979, Parfit 1984 (esp. Appendix G), Holtug 2001, Roberts 2003, Bykvist 2007, Cusbert and Greaves 2015, and Bader 2015 (most defending the view as at least plausible, Bader arguing strongly against it). Sometimes this account is couched in terms of lives ‘worth living’ (though that terminology is also sometimes employed in a different way). A life worth living is supposed to be a life such that it would be better for a person that she lives that life than that she doesn’t exist at all. A life of which this is not the case is called a life ‘not worth living’ (or, sometimes, a life ‘worth not living’). Sometimes, this issue is put in terms of the question of whether existence can be a harm or a benefit.
CHAPTER 3. VARIABLE POPULATION DECISIONS

or at least of plausibility. To others it has the ring of absurdity or confusion.\(^7\) I find myself in this latter group. There seems to me to be something wrong with the claim that a particular life can be either better or worse for a person than not existing at all. There seems to me to be a very useful and well-defined notion picked out by talk of things being better for a particular person, which is a matter of comparing how a person fares (her wellbeing) in one case, with how she fares in another case. Further, I don’t think that a person fares in any way—that things go either well or poorly for her (comparatively or non-comparatively)—in a world in which she doesn’t exist. So claims to the effect that some life would be either better or worse for a person than non-existence seem to me to make a kind of category error. An analogy: if you tell me that this marble is heavier than the number four, I’ll be inclined to correct you by explaining that a number is not the kind of thing that has a mass and not the kind of thing than which something can be heavier (or lighter). Similarly, if you tell me that it is worse for Jane to live a particular life than not to exist at all, I’ll be inclined to correct you by explaining that not existing is not the kind of thing that is any way at all for Jane, and not the kind of thing than which something can be worse (or better) for Jane.\(^8\)

Part of the appeal of the comparative account, I think, is that what is better or worse for people seems to be just the familiar thing with which utilitarianism is concerned. So, if it is true of some lives that it would be worse for a person to live that life than not to exist at all, that would seem to provide a nice, simple utilitarian justification for the wretchedness threshold. Just as one ought not kick a person if that makes the person worse off, one ought not create a person if that makes her worse off! However, I don’t think things are quite

---

\(^7\)For example, Broome writes:

[I]t cannot ever be true that it is better for a person that she lives than that she should never have lived at all. If it were better for a person that she lives than that she should never have lived at all, then if she had never lived at all, that would have been worse for her than if she had lived. But if she had never lived at all, there would have been no her for it to be worse for, so it could not have been worse for her (1999, 168).

\(^8\)Some people have argued that these comparative claims don’t make sense; that it is nonsensical to say that it would be better (or worse) for Jane to live at wellbeing \(x\) than not to exist at all. Cusbert and Greaves (2015) have argued convincingly that there is no semantic obstacle to such claims. I understand the claim that it would be better for Jane to live some life than not to exist at all, and think that probably all such claims are false. Likewise, I understand the claim that this marble is heavier than four and think that it is false. Bader provides (a smorgasbord of) better, largely metaphysical arguments against this kind of comparativist view. He argues that the betterness relation appealed to ‘requires (a) concrete lives as relata, and (b) concrete properties as good-makers’, which non-existence does not provide (2015, 13).
3.1. WRETCHEDNESS

this straightforward. As I’ve just indicated, whether comparisons of this kind between existence (with some particular life) and non-existence are possible is contested. Even if there is a sensible notion of when one thing is worse for a person than another, according to which they are possible, it is not obvious that this is the notion of worseness-for-a-person with which a utilitarian is concerned. If by ‘what is better or worse for people’ you mean this broader thing, it seems plausible that the utilitarian concern is narrower than this. The utilitarian should perhaps be concerned only with betterness of one lived life than another, and not with betterness of one lived life than non-existence.

Consider another account of the wretchedness threshold, which I find more appealing. I call it the ‘shortest truncation account’ of wretchedness. This account relies on being able to compare how good it would be, for a particular person, a) to live one life, and b) to live a truncation of that life. By a truncation of life \( L \) I mean a life that is identical to some initial part of \( L \); that is the same as \( L \) up to some point in the life, and then simply ends (by immediate, painless death). I have assumed that we can evaluate how good any given life is for the person living that life—this is the role of wellbeings. We can use wellbeings to compare how good it would be, for a person, to live one life or another, even if one of those lives is a truncation of the other. According to the shortest truncation account of wretchedness, a life is wretched iff it would be better to live the shortest truncation of that life, rather than the whole life.

*The shortest truncation account:* Wretched lives are all and only those lives during which there is some point, such that it would be better for a person to live a truncation of the life up to that or any earlier point, than to live out the full life.

It seems thoroughly plausible to me that sometimes, a truncation of a life is better (has higher wellbeing) than the full life. Think of someone facing the prospect of a painful, drawn-out death. It might be better for that person to die immediately and painlessly. Now, we would all hope that such a point occurs later in our life, rather than earlier. But it might occur very early indeed. Consider again the worst possible life, a long life of unrelenting torture and misery. Imagine that someone has just been born who will live this life, the

\[9\]This formulation avoids any reference to an earliest point or moment in a life, which would be problematic according to some views about the structure of time (e.g. if time is continuous). Another issue in this area is when a life begins, and whether it does so instantaneously or gradually (see Pummer 2016). The shortest truncation account is similar to the account in Bader 2015.
only alternative being that she dies immediately and painlessly. I think that it would be better for that person if she dies immediately and painlessly. And if the decision were mine, if I had to choose whether this person lives out her long life of misery or dies now, I think that I ought to choose that she dies. (At least, according to utilitarian concerns, and remembering that this is a decision under certainty. A realistic decision will be much more difficult.)

The shortest truncation account, like the non-existence comparative account, would require justification. Why should this kind of comparison between lives correspond to a moral threshold regarding *creating* lives? Some motivational force can be drawn from the apparent similarity between immediately and painlessly killing someone just after they come into existence, and preventing a person from coming into existence. Imagine an evil scientist has just set in motion, in some room, a process that will create a person and then slowly torture her to death. You have one opportunity to intervene, by destroying the room and its contents just after the person comes to exist. Say the life would be such that you ought to intervene—a brief flicker of that life would be better for the person than the long life of torture. Now imagine a slightly different case, in which your one opportunity to intervene occurs just before the person comes into existence. Could this slight difference in timing mean that it is not the case that you ought to intervene? This seems unlikely, to me.\(^\text{10}\)

This kind of similarity is persuasive, though it doesn’t really provide an *explanation* of the proposed link between truncated lives and wretchedness. One possible explanation again appeals to non-existence comparativism. It might be the fact that the tortured person’s life would be worse for her than non-existence that justifies your obligation in the second case. There are alternative explanations, however. For example, Bader argues that allowing someone to come into existence with a life such that you ought, if possible, end that life as soon as possible, fails a ‘consistency condition’ (2015).

A final note on the shortest truncation account. I have presented it as a necessary and *sufficient* condition for wretchedness, but in fact it seems most compelling to me as a sufficient condition. Its status as a necessary condition seems more questionable. Perhaps the wretchedness threshold should be higher than the immediate-death threshold. Perhaps we should expect more

---

\(^{10}\) Admittedly this kind of analogy is tricky, complicated by the fact that coming into existence is probably not an instantaneous matter.
from someone’s life than being just better than immediate death, before we rescind the obligation not to create a person.\textsuperscript{11}

The question of where the wretchedness threshold is, is a difficult one. The accounts I have discussed are two among many possible candidates. Though I have indicated my tentative preference for the latter account, I do not hope to have convinced the reader of the truth or falsity of either. I have merely presented some of the considerations telling for and against each, and some aspects of the related ongoing debates. I now set this issue aside. I think that there must be a wretchedness threshold, and I proceed as though the way to identify the wretchedness threshold were established. I assume that any theory according to which there is such a threshold identifies it in the right way (whatever that is).

Finally, the third question: which number on the wellbeing scale should represent the wretchedness threshold? I’ve assumed that we can use integers to represent wellbeings. In section 1.2 I assumed that we can represent wellbeings on an interval scale, so that a difference of some amount somewhere on the scale is worth just as much as a difference of the same amount somewhere else on the scale. A life at wellbeing 20 is worse than a life at wellbeing 25 by the same amount by which a life at wellbeing −2 is worse than a life at wellbeing 3. Now I’ve assumed that there is a special point on the scale, the wretchedness threshold. Is there any reason to assign a particular number to the wretchedness threshold? I think that there is some reason to assign 0 to the wretchedness threshold. The reason is that this makes for nice, simple presentations of moral theories. I don’t think that there is any deeper theoretical significance to this convention. (It doesn’t, for example, mean that the wellbeing scale is a ratio scale—I don’t think that we can meaningfully say that one wellbeing is twice, or a third, as high as another.) I shall therefore assign to the wretchedness threshold (if there is one) wellbeing 0.\textsuperscript{12}

\textsuperscript{11} Benatar suggests something similar when he distinguishes between two possible senses of ‘a life worth living’, and argues that the threshold for a life worth continuing to live might be lower than the threshold for a life worth starting (2006).

\textsuperscript{12} Think about the temperature analogy again. If you want to appeal to temperatures in a context in which the freezing point of water is particularly significant, you might want to use the Celsius scale, or some other scale that assigns 0 to the temperature at which water freezes (so that calculating differences from the freezing point of water, for example, is simple). If some other temperature is most significant you might want to use a scale that assigns 0 to that temperature. In neither case would you be justified in saying that one thing is twice as hot as another. (You haven’t made a ratio scale by being interested in freezing water.) Assigning wellbeing 0 to the normative-wretchedness threshold obscures a difference in the literature of which much is made, between TU and ‘critical
3.1.2 Wretchedness: the first test

My intuitions about wretchedness provide the first test for those candidate general utilitarian theories I introduced in Section 2.3: TU, AU, and HU. Whether or not there is a wretchedness threshold according to some theory depends on what that theory says about decisions of a particular kind: two-outcome decisions between creating a person and not creating a person, where things in the two outcomes are otherwise equal. TU passes this test, and AU doesn’t. HU doesn’t pass the wretchedness test either, though there are some adjustments to HU that produce slightly different theories that do pass. I’ll go through what each of the three theories say about wretchedness now in a bit more detail.

According to TU there is a wretchedness threshold, and it is wellbeing 0. To see this, consider this decision:

\[
\begin{array}{c|cc}
\text{D}_{10}: & \text{Anne} & \text{Bob} \\
A & 10 & * \\
B & 10 & -10 \\
\end{array}
\]

The only difference between the two available outcomes, A and B, is that Bob exists with wellbeing −10 in B and doesn’t exist at all in A. The total wellbeing of the population of A in A is 10, and the total wellbeing of the population of B in B is 0. Creating Bob leads to a lower total wellbeing, and thus according to TU one ought not create Bob in this decision. It is clear that this conclusion will generalise. According to TU, in a decision between creating a person with a wellbeing below 0 and not creating that person, other things being equal, one ought not create that person. And in a decision between creating a person with a wellbeing of 0 or higher and not creating that person, other things being equal, it is not the case that one ought not create that person (creating the person does not lead to a lower total wellbeing than not creating her). Thus wellbeing 0 is the wretchedness threshold according to TU.\textsuperscript{13}

According to AU there is no wretchedness threshold. It is true that in \( \text{D}_{10} \) AU agrees with TU, that one ought not create Bob. But this isn’t because Bob’s wellbeing in B is below some stable threshold wellbeing level; it is because

\textsuperscript{13} As I said, I think that using 0 to pick out the wretchedness threshold is nothing more than a useful convention. If we assigned the numbers to wellbeings slightly differently, for example by shifting them all down by seven, we could represent TU slightly differently. In that case, the wretchedness threshold would be 7, and according to TU choice-worthiness would be defined with reference not to sums of wellbeings, but to sums of wellbeings-minus-7.
3.1. WRETCHEDNESS

Bob’s wellbeing lowers the average, which is a matter of how it compares to Anne’s wellbeing in each outcome. Consider another decision:

<table>
<thead>
<tr>
<th></th>
<th>Carl</th>
<th>Dave</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>−12</td>
<td>*</td>
</tr>
<tr>
<td>(B)</td>
<td>−12</td>
<td>−10</td>
</tr>
</tbody>
</table>

In this decisions, AU says that one ought to create Bob—despite the fact that his wellbeing is −10, as in the previous decision—because doing so leads to a higher average wellbeing. As I’ve said, the idea that the wretchedness threshold is 0 wellbeing is merely conventional. It should be clear, though, that there is no wretchedness threshold according to AU, and not just that the threshold is not 0. No matter which numbers you attach to which wellbeings, there is no wellbeing below which lives are wretched, according to AU. Whether one ought to create a person, other things being equal, depends not only on the wellbeing that person will have, but also on other people’s wellbeings (even when those other wellbeings are the same in both outcomes). This is not to say that according to AU all lives are happy; rather, there are neither happy nor wretched lives. There are no wellbeings such that, other things being equal, one ought never create a person who will live such a life. There are no wellbeings such that, other things being equal, it is never the case that one ought not create a person who will live a life at that wellbeing.\(^{14}\)

According to HU there is no wretchedness threshold. Consider \(D_{10}\) again. No-one incurs any harm in outcome \(B\) (nor in \(A\)). Harms, remember, are assessed by comparing a person’s wellbeing in one available outcome with that same person’s wellbeing in each other available outcome. (See the definition of harm on page 43.) I have assumed that a person has a wellbeing in an outcome only if she exists in that outcome. So if a person exists in only one available outcome in a particular decision, she is not harmed in that decision. There is no other wellbeing with which to compare her wellbeing in the outcome in which she exists. She has by default her maximum wellbeing in the outcome in which she exists, because she has no wellbeing (no wellbeing, not wellbeing zero) in

\(^{14}\text{This result is counterintuitive in itself. It also produces some other, or more specific, counterintuitive results. For example, AU also says that whether or not one ought to create a possible person depends on the wellbeings of very distant people who aren’t affected by the decision and who don’t affect the (merely) possible people in question. For example, whether or not I ought to have a child depends on how well the ancient Egyptians fared (on this see McMahan 1981, 115, and Parfit 1984, 420).}\)
any other available outcome. This means that, not only is there no wretchedness
threshold according to HU, but in any two-way decision between creating a
person and not creating a person, other things being equal, HU says that the
two options are equally choice-worthy.

Clearly, this result relies on my assumption that a person has a wellbeing
only in outcomes in which she exists. However, most people who endorse a
harm minimising theory—including McDermott (1982b) and Roberts (2011b)—
reject this assumption. In each outcome, they attribute to each possible person
who does not exist in that outcome wellbeing zero. This, in conjunction with
the account of HU (also on page 43), produces quite a different and, I think,
more plausible theory.\footnote{Revising this assumption in this way does not
affect TU or AU at all, because I have presented them each in terms of maximising the total or average wellbeing of the populations of outcomes (and the population of an outcome consists of only the people who exist in that outcome). It would affect a version of AU according to which one ought to maximise the average wellbeing of all possible people. That version of AU would be equivalent to TU (because the number of possible people is constant, at least within a decision).} It means that a person is harmed not only when she
exists with a lower wellbeing rather than existing with a higher wellbeing, but
also when she exists with a wellbeing below zero rather than not existing at
all. So in decision $D_{10}$, for example, Bob does incur a harm in $B$. In $B$ Bob has
wellbeing $-10$ and in $A$ he has (under this revised assumption) wellbeing $0$.\footnote{Under this revision, a person is also harmed when she doesn’t exist at all rather than existing with a wellbeing above zero. However, this kind of harm does not contribute to the total harm incurred by the population of an outcome, because a person who does not exist in some outcome is not a member of the population of that outcome. (See the definition of HU on page 43.) Much more on this in the next section, where we see that NHU produces an intuitive ‘Asymmetry’ concerning wretched and happy people.}

I presented HU as a moral theory, but now we see that which theory is
picked out by my definitions of harm and Harm Utilitarianism depends on
which assumptions we make about wellbeing. I shall continue to use ‘Harm
Utilitarianism’ (and ‘HU’) to refer to the theory picked out by my definitions
under the assumption that a person has a wellbeing only in outcomes in which
she exists. I shall call the other version—the theory picked out by my definitions
under the assumption that a person has wellbeing zero in outcomes in which she
does not exist—‘Non-existence Harm Utilitarianism’, or ‘NHU’.\footnote{There are other ways of representing NHU (remembering that I identify a theory by the choice-
worthiness orderings it produces). One might maintain that a person has a wellbeing only in
outcomes in which she exists, and alter the definition of a harm to reach the conclusions that Bob
is harmed in $B$ in $D_{10}$ despite the fact that he has no wellbeing in $A$. One might motivate this
revised definition of harm by defending the existence of ‘non-comparative’ harms. Bykvist (2007)
defends the existence of such non-comparative harms in this context. Though I find NHU a more
plausible theory than HU, I find both of these representations of NHU unsatisfying. It strikes me as}
3.2. Neutrality

HU there is no wretchedness threshold; in a decision between the two, other things being equal, creating a person and not creating her are always equally choice-worthy. According to NHU there is a wretchedness threshold, and it is 0.

This intuition of wretchedness is the first of my four intuitions about variable population decisions. I have presented it in terms of a normative-wretchedness threshold. There is, I intuit, a wellbeing level such that one ought not create a person with a wellbeing below that level (other things being equal and in a decision between the two), and it is not the case that one ought not create a person with a wellbeing at or above that level (other things being equal and in a decision between the two). TU and NHU agree with this intuition; AU and HU do not. In the next section I’ll introduce my second intuition, which is an intuition about the creation of happy people. I will return to wretchedness in that section, when I address the issue of a well-known asymmetry between these two intuitions.

3.2 Neutrality

Imagine you face the decision between creating a person who will, if she exists, live a happy life (a life with a wellbeing at or above the wretchedness threshold) and not creating that person, and that no-one else is affected by the decision. Part of my intuition about wretchedness, and my definition of the wretchedness threshold, is that it is not the case that you ought not create that person. I have not so far said anything about whether you ought to create them. My intuition is that it is not the case that one ought to create a happy person, other things being equal. Creating a person who will live a happy life and not creating that person are equally choice-worthy in a decision between the two, other things being equal. I will call this the neutrality intuition: the idea is that creating a happy person is morally neutral. A decision to illustrate the neutrality

18 unnecessarily forced to try to account for what’s wrong with creating a wretched person in terms of worseness (or badness) for that person (which is what I take harms to be about).

I’m adapting the term from Broome’s ‘neutrality intuition’, which is roughly the intuition that there is some range of wellbeings such that it is neither better nor worse to create a person at a wellbeing in that range (other things being equal) (2004, 143). My neutrality intuition is that this range has no upper bound and that the lower bound is the wretchedness threshold. Broome’s focus is axiological: ‘Interpreted axiologically, in terms of goodness, the intuition is that if a person is added to the population of the world, her addition has no positive or negative value in itself.’ (2004, 145–146). My focus is again normative (though of course I expect the normative facts to be
CHAPTER 3. VARIABLE POPULATION DECISIONS

intuition is the following.

<table>
<thead>
<tr>
<th></th>
<th>Anne</th>
<th>Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{12}$:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A$</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>$B$</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

This is the decision between creating Bob (who would then live a happy life) and not creating Bob, where things are otherwise equal (in particular, inevitable Anne’s wellbeing is not affected by the decision). The intuition of neutrality is that $A$ and $B$ are equally choice-worthy in this decision. It is neither less nor more choice-worthy to create a person—even a very happy person—rather than not to create her (other things being equal, in a decision between the two).

I have a direct moral intuition of neutrality. This intuition also seems to me to be a utilitarian intuition. Here are some thoughts that seem to me to support this intuition to some extent. As utilitarians we ought to make it the case that people are as well off as possible, in some difficult-to-capture sense. Creating a happy person rather than not creating her does not make her better off in the relevant sense. In the relevant sense, it doesn’t seem to affect how well off people are at all. It merely makes it the case that there is an extra person (with a particular wellbeing) rather than not.

Another well-known idea in this vein is summarised in Narveson’s catchphrase: ‘Make people happy, not happy people’ (1967). Narveson also provides an analogy to support the neutrality intuition (1967, 66–67). Imagine that the government of the country of Fervia announces that they are going to increase the wellbeing of their citizens. Fervians rejoice. It is then revealed that the increase in wellbeing of the citizens of Fervia will be achieved by granting Fervian nationality to the members of a recently discovered flourishing population on Mars. The wellbeings of current Fervians will not be affected by the expansion of the national borders. For that matter, neither will the wellbeings of current Martians. But the sum total wellbeing of Fervian citizens will be increased (we can stipulate that the average will also rise), simply in virtue of this shifting of national borders. I would expect Fervians to be disappointed when these details are revealed. I would expect them to feel misled, to feel that the government is not making its citizens better off at all. After all, by stipulation

determined by axiology).

19 Also ‘We are in favour of making people happy, but neutral about making happy people’ (Narveson 1973, 80).
they haven’t affected any particular person’s wellbeing (nor even exchanged some people for other, better-off people). It might be that after the Martian expansion the government will be able truthfully to assert that the total (or average) wellbeing of Fervians is $x$, where $x$ is higher than the total (or average) wellbeing of Fervians currently. However, the Fervian government will not be able truthfully and transparently to claim to have increased the wellbeing of Fervians by this act. So, the analogy goes, one does not increase the wellbeing of people generally, by adding a happy possible person to their number—at least, one does not do so in the morally relevant sense.

The neutrality intuition is quite common, though not so common, I think, as the intuition of wretchedness. And among those who have the intuition of neutrality, some are persuaded that it is misguided, for one reason or another. Let me consider a couple of those reasons. Some people find their neutrality intuition shaken by some further prodding and prompting of intuitions, along the following lines: ‘Imagine you could press a button to bring into existence, on a remote and isolated planet, a population of people who would all live very happy lives. Don’t you think that it would be better if you pressed that button? Wouldn’t the world be better for containing this planet of very happy people? If so, doesn’t this mean that you should press the button?’ This kind of reasoning is not without persuasive efficacy. However, my considered response is to reject it. Narveson considers a similar line of thought (1967, 67–69) and my response is similar to his. He suggests that the inclination to say that it would be better if the button were pressed reflects a ‘matter of taste’, rather than a moral claim (1967, 68). Let me elaborate.

I think it probably would be better if I pressed the button to bring into existence a remote and happy population. I would be inclined to press the button, given the chance (and given the fairly unimaginable guarantee that doing so will have no ill effects on inevitable people and that all members of the remote population will be happy). However, I don’t think that this attitude is morally relevant, or based on any morally relevant facts. I think that it is based on my preference that complex and interesting things exist, regardless what their effects on people’s wellbeings are. I have all sorts of preferences of this kind. I prefer that beautiful flowers flourish in unvisited valleys. I

---

20One might wonder whether these preferences are rational or not. I have nothing to say on that matter.
prefer that the world is understood (scientific progress is made, etcetera), even if this doesn’t affect anyone’s wellbeing. But I think that the content of these preferences are morally irrelevant. Imagine you could press a button to bring into existence, on a remote and isolated planet, a very beautiful mountain. Don’t you think that it would be better if you pressed that button? Wouldn’t the world be better for containing this very beautiful mountain? In this case too, I think that the world would be better if I pressed the button. But I don’t think that it would be better in a morally relevant way.\textsuperscript{21} We should not be too hasty to accept as moral intuitions any compelling axiological or normative claims. Not all betterness is morally relevant betterness and not all ‘ought’ claims are ‘morally ought’ claims.

Another thing that has the potential to lead people away from the neutrality intuition is that it means that our obligations with respect to happy lives and wretched lives are puzzlingly asymmetric. This fact is much discussed under the title ‘the Asymmetry’, and is often thought to be in need of explanation or justification.\textsuperscript{22} Why should (how could) wretchedness count against creating a person if happiness doesn’t count for it, or vice versa? Whatever explains or justifies the existence of a wretchedness threshold, won’t some symmetrical justification for an obligation to create happy people be compelling? Many people seem to find the wretchedness intuition more compelling that the neutrality intuition, so in the absence of an explanation of the asymmetry between the two they feel some pressure to give up neutrality. Let me set this issue aside until I discuss how HU and NHU fare with respect to neutrality shortly. In that context it will be easier for me to show how the asymmetry might be explained.

3.2.1 Neutrality: the second test

My intuitions about neutrality provide the second test for the theories we have encountered so far. As was the case for wretchedness, the test here concerns two-way decisions of a particular kind—decisions about whether or not to create a person, other things being equal. A theory that passes this test is a theory according to which it is not the case, in any such decision, that one ought to create the person. (If the person would be wretched one ought not

\textsuperscript{21} This raises the question of whether (and how) non-moral factors might play into decision-making. I leave this entirely open.

\textsuperscript{22} See for example Parfit 1984, McMahan 2009, Roberts 2011a, and Bradley 2012a.
create her, and if she would be happy the two available outcomes are equally choice-worthy.) A theory that fails is a theory according to which it is (at least sometimes) the case that one ought to create a happy person rather than not create her, in a decision between the two when things are otherwise equal. On this test, TU fails and NHU succeeds. The status of AU and HU is a little complicated (because they failed the wretchedness test), but I conclude that AU fails and HU succeeds. Here is some further explanation of these results.

According to TU, it is almost always the case that one ought to create a happy person rather than not creating her, other things equal. Only if the happy person has wellbeing zero is this not the case. Creating a happy person with a wellbeing above zero increases the total wellbeing of the population. This should be fairly obvious. Consider decision $D_{12}$ again.

\[
\begin{array}{ccc}
\hline
D_{12}: & A & 10 & \ast \\
& B & 10 & 10 \\
\hline
\end{array}
\]

So long as Bob’s wellbeing is above zero, the total wellbeing in $B$ is higher than the total wellbeing in $A$, and so $B$ is more choice-worthy in $D_{12}$ than $A$ according to TU. If Bob’s wellbeing were zero, creating Bob and not creating him would be equally choice-worthy. (If his wellbeing were below zero of course he would be wretched, and creating him would be less choice-worthy.) TU takes a symmetrical approach to wretchedness and happiness. According to TU one ought not create a wretched person and one ought to create a happy person (other than in the special case where her wellbeing is zero).

It is a little difficult to know what to say about AU and the neutrality intuition. I have spelled out the neutrality intuition by reference to happy lives, which in turn are defined in terms of the (normative-)wretchedness threshold. But according to AU there is no such threshold. So we can’t really ask what AU says about creating happy people. I’ve made the neutrality test more general, however. My intuitions about wretchedness and neutrality combine to produce the conclusion that it is never the case that one ought to create a person rather than not, in a decision between the two, other things being equal. It might be that one ought not create the person (if her life would be particularly bad for her); otherwise, creating her and not creating her would be equally choice-worthy. AU fails this test. According to AU it is sometimes the case that one ought to create a person rather than not, in a decision between the two where
things are otherwise equal. Whether or not this is the case depends on how creating the person will affect the average wellbeing of the population. In $D_{12}$, adding Bob has no effect on average wellbeing. Adding Bob and not adding him are equally choice-worthy in that decision according to AU. If Bob’s wellbeing were above the average in $A$, however, $B$ would be more choice-worthy in $D_{12}$ than $A$. According to AU one ought to create a person iff doing so will result in a higher average wellbeing.

Now let’s consider HU and NHU. We saw above that according to HU, creating a person and not creating her are always equally choice-worthy, other things being equal, in a decision between the two. This is because, under the assumption that people have wellbeings only when they exist, a person who exists in only one available outcome in a decision has her maximum wellbeing in that outcome by default and is not harmed in that outcome in that decision. Nor is she harmed in the outcome in which she doesn’t exist. By stipulation no-one else is harmed in either outcome. Both outcomes are harm-free and thus equally choice-worthy. This meant that HU failed my test for wretchedness. HU implies the implausible conclusion that there is no wretchedness threshold. So again (as for TU), it is not quite straightforward to say how HU fares with respect to the neutrality test. However, taking the test again to be whether the theory agrees that it is never the case that one ought to create a person rather than not (in a decision between the two, other things being equal), HU passes. Like TU, HU takes a symmetrical approach to wretchedness and happiness. (This is speaking loosely. According to HU there is no wretchedness threshold, so it would be more accurate to say that according to HU there are no wretched or happy lives.) According to HU it is never the case that one ought to create a person and never the case that one ought not create a person (in the two-way, other things being equal kind of decision we’re considering).

The only one of the theories we have encountered so far that agrees with my neutrality intuition completely is NHU. According to NHU there is a wretchedness threshold, and creating a happy person is morally neutral. So NHU is asymmetric in the way I hoped a general utilitarian theory would be. We’ve already seen what accounts for the wretchedness threshold according to NHU. If people have wellbeing zero in outcomes in which they don’t exist, then a person is harmed if she exists with a wellbeing below zero rather than not existing at all (i.e. she is harmed in this way in a decision in which both of these
options are available). This harm counts against creating her. Where things are otherwise equal this harm is decisive. So, one ought not create a wretched person, other things being equal, in a decision between the two.

NHU deals symmetrically with happiness, up to a point. If people have wellbeing zero in outcomes in which they don’t exist, then a person is harmed if she does not exist instead of existing with a wellbeing above zero (see footnote 16 on page 60). In B in D$_{12}$, Bob has wellbeing 10. In A, the revised assumption is that Bob has wellbeing 0 (by not existing). Here is the decision again, adjusted to reflect this revision.

<table>
<thead>
<tr>
<th></th>
<th>Anne</th>
<th>Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>*</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

Under this revision, Bob incurs a harm of 10 in A. However, this harm doesn’t count against A, according to NHU. Bob is not a member of the population of A, and the morally relevant feature of an outcome according to NHU is the total harm incurred by the members of the population of that outcome. In D$_{12}$ the total harm incurred by the members of the population of A in A is 0; likewise for the total harm incurred by the members of the population of B in B. The two outcomes are equally choice-worthy in D$_{12}$ according to NHU, and this will be the case no matter how high Bob’s wellbeing is in B.

At this point you might have some independent qualms about NHU. Why should the harms incurred by Bob in A in D$_{12}$ be morally irrelevant? Having assumed that Bob has wellbeing zero in A, and accepted the given definition of harm, shouldn’t Bob’s harm in A count against bringing it about, when B is available? It is promising that NHU agrees with the wretchedness and neutrality intuitions, but you might wonder not only what the explanation of this is, but also how it is to be justified. I’ll spell out in a bit more detail the explanation of how and why NHU agrees with these two intuitions. I’ll also explain why I find this explanation fairly compelling. I will not attempt to justify this explanation or the Asymmetric intuitions.$^{23}$ There are two elements to the explanation. First, according to NHU there is a difference between a harm incurred by an existing person and one incurred by a non-existent (merely

$^{23}$ Thanks to John Cusbert and Matt Clark for drawing my attention to this important difference, between an explanation of how a theory produces certain conclusions, and a justification of those conclusions as true.
possible) person. Secondly, according to NHU harms are morally significant in a way that benefits are not. Let me take these two elements in turn.

Bob is harmed in $A$ in $D_{12}$. He has wellbeing 0 in that outcome, while in another available outcome he has wellbeing 10. NHU says that this harm is irrelevant because Bob does not exist in $A$. The harms that matter are the harms incurred by the members of the population of $A$. But why should this be so? We care about all possible people and their wellbeings, surely. Shouldn’t we care about all of their harms? In fact I think that it is more plausible that we should only care about harms incurred by existing people. The harm incurred by Bob in $A$ in $D_{12}$ is a harm that will, if $A$ is chosen, affect only a merely possible person. It is not a harm that any actual person will ever suffer. We should feel something like moral remorse if we see that someone has lower wellbeing than she might have had. That person is missing out! That person deserves our compassion, condolences, perhaps compensation. If we see that some merely possible person, who in fact will never exist, has lower wellbeing than she might have had, I don’t think that we should feel the same way. There is no-one who is missing out, no-one to whom we might offer compassion, condolences, and so on.

The other element of the explanation of the asymmetry produced by NHU is that benefits are morally irrelevant according to that theory. You might be persuaded that the harm incurred by non-existent Bob in $A$ in $D_{12}$ is not morally relevant. However, you might think that Bob is benefited in $B$ in $D_{12}$, and that this is morally relevant. NHU disagrees. Here too I think that NHU is compelling, though this aspect of the theory is perhaps more controversial and more difficult to defend.

Here is a plausible definition of a benefit, which fits nicely with the definition of a harm on which NHU is based.

**Benefit:** For any decision, $D_1$, person, $p_1$, and available outcome in $D_1, A$: $p_1$ incurs a benefit of amount $x$ in $A$ just in case $p_1$’s wellbeing in $A$ exceeds the minimum wellbeing of $p_1$ in any available outcome in $D_1$ by $x$.

In fixed population two-way decisions, minimising harms incurred and maximising benefits incurred amount to the same thing. Here is a decision to illustrate.
3.2. NEUTRALITY

In \( D_{13} \) \( p_2 \) is harmed in \( A \) by 4 and is benefited in \( B \) by 4. These two claims amount to the same thing, and it is difficult to know (if the question is even meaningful) whether what matters morally is that \( p_2 \) is worse off in \( A \) than in \( B \), or that she is better off in \( B \) than in \( A \). In decisions about creating people, though, the question is more pressing. In \( D_{12} \) there are no harms suffered (incurred by existing people), but there is a benefit enjoyed (incurred in \( B \) by existing Bob). Here it matters whether we focus on harms incurred by existing people, benefits incurred by existing people, or both.

The idea that we ought to both minimise harms and maximise benefits is intuitively appealing. On reflection, though, I think that the idea that our only obligation is to minimise harms is at least as compelling. It is difficult to offer a non-circular justification of this claim. Some people have attempted to argue from intuitions in fixed population cases that harms are more morally significant than benefits.\(^{24}\) It would provide some support for this idea if it were the only alternative to giving up the neutrality intuition (or the wretchedness intuition, though that seems a less population option). I won’t attempt to justify the claim that our only obligation is to minimise harms here.

3.2.2 Neutrality and overall betterness

At this point I would like to point out that the neutrality intuition seems to me to be in tension with the overall betterness requirement.\(^{25}\) Consider these three decisions:

\[
D_{14}: \begin{array}{c|c|c}
\ p_1 \ & \ p_2 \\
A & 15 & \\
B & 15 & 13
\end{array} \quad D_{15}: \begin{array}{c|c|c}
\ p_1 \ & \ p_2 \\
A & 15 & \\
C & 15 & 17
\end{array} \quad D_{16}: \begin{array}{c|c|c}
\ p_1 \ & \ p_2 \\
B & 15 & 13 \\
C & 15 & 17
\end{array}
\]

According to the neutrality intuition, \( A \) and \( B \) are equally choice-worthy in \( D_{14} \), and \( A \) and \( C \) are equally choice-worthy in \( D_{15} \). If we are to represent choice-worthiness as an overall betterness relation (or directly determined by one), it

\(^{24}\)For example Shiffrin 1999.

\(^{25}\)This point is made at length in Cusbert and Kath 2016.
CHAPTER 3. VARIABLE POPULATION DECISIONS

must be that $A$, $B$, and $C$ are all equally good overall. But then look at $D_{16}$. This is a fixed-population decision, and falls under the jurisdiction of CFPU: $C$ is more choice-worthy in $D_{16}$ than $B$. To represent this fact in terms of an overall betterness relation we would have to say that $C$ is overall better than $B$. At this point the problem should be clear; these three claims about overall betterness do not fit into a single betterness relation. For we’ve just asserted that $A$ is equally as good as $B$ and also equally as good as $C$, and that $B$ is better than $C$.

Consider also the decision between all three outcomes.

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>15</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>$C$</td>
<td>15</td>
<td>17</td>
</tr>
</tbody>
</table>

This is a variable population decision, so CFPU does not tell us what the choice-worthiness ordering in this decision should be. However, most people who have the intuitions I appealed to in the two-way decisions above have the further intuition that $A$ and $C$ are equally choice-worthy (and both permissible) and $B$ less choice-worthy (and impermissible) in this three-way decision. A natural way to explain this intuition is to say that we remain neutral about whether or not one ought to create a happy person, but if one is to create a happy person one ought to create her as happy as possible. To account for this intuition in terms of overall betterness we would have to say that $A$ and $C$ are equally good overall, and both overall better than $B$. This conflicts with what we wanted to say in the decision just between $A$ and $B$, $D_{14}$. There we wanted to say that $A$ and $B$ were equally good overall.

As well as indicating a possible tension between the neutrality intuition and the overall betterness requirement, these cases provide some initial support for the thought that choice-worthiness might be menu-dependent in interesting ways. If $A$ and $B$ are equally choice-worthy in $D_{14}$ and $A$ is more choice-worthy

\[26\] An alternative is to say that they are all incomparable. Rabinowicz defends this (2009). Broome argues that this kind of incomparability is unpalatably ‘greedy’: too much ends up incomparable and thus permissible (2004 and 2009). More on this debate between Rabinowicz and Broome in Cusbert and Kath 2016. Bader also endorses widespread incomparability of overall goodness (2015).

\[27\] Broome appeals to this intuitive idea here: ‘it does not matter morally whether we add a new person to the population (provided her life will be good), but if we do add one, we must do our best for her’ (Broome 1994, 175).

\[28\] Or, again, take the overall incomparability route; see the previous footnote.
3.2. **NEUTRALITY**

in D₁₇ than B, this is an instance of a change in the relative choice-worthiness of a pair of outcomes from one decision to another. Much more on this (including whether choice-worthiness is ‘really’ just menu-dependent overall betterness) in Sections 5.1 and 5.2.

NHU, by the way, agrees with all of the intuitions I have appealed to regarding decisions D₁₄-D₁₇. NHU seems to arrive at these conclusions solely on the basis of various relations of betterness—relations of betterness for particular people. Some people have taken the tension between the neutrality intuition and the overall betterness requirement, and the menu-dependence inherent in these neutrality-based intuitions, as evidence against the neutrality intuition. I take these things as evidence against the overall betterness requirement, and in favour of the menu-dependence of choice-worthiness.

Recall the analogy I gave in Section 2.2.3 to illustrate why I wanted to avoid the overall betterness requirement. I set up the project of developing a procedure for deciding which beach to go to. Here again is the table containing the beaches’ rankings in the three relevant respects of betterness.

<table>
<thead>
<tr>
<th></th>
<th>Ocean</th>
<th>Sand</th>
<th>Pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bondi</td>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Tamarama</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Bronte</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Gordon’s Bay</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Coogee</td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

I proposed an axiological beach decision procedure that reflected a kind of neutrality about pools. I said that it was possible that one way to make decisions on the basis of these three kinds of betterness was to focus on (i.e. minimise) worseness—ocean-worseness, sand-worseness, and pool-worseness. This might be the right approach if my attitude to pools is that I neither prefer to go to a beach with a pool than one without, nor vice versa, but if I do go to a beach with a pool I prefer to go to one with the nicest possible pool conditions. On this approach, in a decision between Tamarama and Bondi those two beaches are equally go-worthy, but in a decision between Tamarama, Bondi, and Bronte, Tamarama is more go-worthy than Bondi.

I hope that it’s now a little clearer how this analogy relates to variable population decisions. Had I embraced the overall betterness requirement, I would have ruled out moral theories that reflect neutrality about creating
happy people. Neutrality about creating happy people is one of my strongest
intuitions about variable population decisions, and it results from some general
utilitarian theories that violate the overall betterness requirement (including
NHU). I don’t want to rule out such theories.

This concludes my discussion of the neutrality intuition, and of the Asym-
metry consisting of the neutrality and wretchedness intuitions. So far, NHU is
the only theory that agrees with both of these intuitions. Now I’ll present my
third intuition about variable population decisions.

3.3 Non-identity

Imagine you face a decision about which of two possible people to create. You
could create one person who, if she exists, will live a moderately happy life;
or you could create a different person who, if he exists, will live a very happy
life. Imagine these are the only two available options, and no-one else will be
affected by your decision. Here is such a decision.

\[
\begin{array}{ccc}
P_1 & P_2 & P_3 \\
A & 10 & 10 & \star \\
B & 10 & \star & 8 \\
\end{array}
\]

A decision with this structure is known as a non-identity decision. It is a
decision between creating a person at one happy wellbeing and creating a non-
identical person at a different happy wellbeing, where things are otherwise
equal. The category also includes decisions between creating one group of
people and creating another, entirely distinct group of people. My focus in this
section is on decisions between groups of people of the same size. In the next
section I will pay some attention to a non-identity decision between groups of
different sizes.

One much discussed non-identity decision is the following. A woman who
wants to conceive a child discovers that she has an illness. If she conceives
now, her child will be born with a wellbeing-lowering disability (though she
will still live a life above the wretchedness threshold). If the woman waits two
months to undergo treatment and then conceives in good health her child will

---

29See for example Parfit 1984 and Boonin 2014. There are also decisions about which people to
create where some of the possible people would, if created, live wretched lives. I exclude such
decisions from my definition of non-identity decisions.
be born without the disability. Neither the woman’s own wellbeing, nor that of any other person will be affected by the decision. The child she could conceive now and the child she could conceive in two months are different people. Thus, the woman faces a decision between creating one happy person, and creating a different, happier person. This might be decision D₁₈ above.\(^{30}\)

I have the intuition that, in decisions like this, one ought to create the happier person rather than the less happy person. The woman ought not conceive now, but ought to wait to conceive in two months. It is better, in some morally relevant way, that one person lives a very happy life, than that a different person lives a moderately happy life.

One way of thinking of this response to non-identity decisions is in terms of the idea that non-identity (or identity) is morally irrelevant. Consider a case quite like the ill mother case above, but without the non-identity element. Imagine a woman has already conceived a child when she finds out that she has an illness. If she does nothing, her child will be born with a wellbeing-lowering disability. If she undergoes treatment her child will be born without the disability. The woman faces a decision between creating a particular person happy, or happier.\(^{31}\) Here is such a decision.

\(^{30}\)This is similar to Parfit’s young mother case (1984, 358). Boonin (2014) focusses on a case like this, in his book on how to avoid the non-identity problem (the answer being: accept its unpalatable implications). He presents the difficulty of avoiding the non-identity problem in terms of these five plausible premises and their implausible conclusion (where Wilma is a mother facing the decision between creating happy Pebbles or happier Rocks):

P₁: Wilma’s act of conceiving now rather than taking a pill once a day for two months before conceiving does not make Pebbles worse off than she would otherwise have been

P₂: If A’s act harms B, then A’s act makes B worse off than B would otherwise have been

P₃: Wilma’s act of conceiving now rather than taking a pill once a day for two months before conceiving does not harm anyone other than Pebbles

P₄: If an act does not harm anyone, then the act does not wrong anyone

P₅: If an act does not wrong anyone, then the act is not morally wrong

C: Wilma’s act of conceiving Pebbles is not morally wrong’ (2014, 27).

See also Adams (1979), who, following Leibniz, uses something like the recognition of non-identity to respond to the problem of evil; and Schwartz 1978. It is easy to think of other cases, for example involving scientists with gametes in petrie dishes.

\(^{31}\)One might question whether her decision is about creating a person at all. Doesn’t the child already exist? Or at least, isn’t it already certain that he will exist? This is true, in the simple case I have given. We could consider a case in which conception had not yet taken place, and in which not conceiving at all was an option. I don’t think that these things would make a difference to my point about the irrelevance of non-identity.
As we’ve already established, \( A \) is more choice-worthy in this kind of fixed population decision than \( B \). One ought not create a person with some wellbeing instead of creating that person with higher wellbeing, other things being equal, in a decision between the two. My intuition is that a general utilitarian theory should conclude that one ought not create a person with some wellbeing instead of creating a person with higher wellbeing, even if the decision is between creating two different people.

### 3.3.1 Non-identity: the third test

Non-identity decisions provide my third test for a general utilitarian theory. I hope to find a general utilitarian theory according to which, in a decision between creating one person at a particular happy wellbeing and a different person at a happier wellbeing, other things being equal, one ought to create the happier person. This is the conclusion reached by TU and AU. HU and NHU reach the counterintuitive conclusion that this is not the case. According to the latter two theories, creating one happy person and creating a different happier person are equally choice-worthy in a decision between the two. There are more complex kinds of non-identity decision too: decisions between creating non-overlapping groups of people, either where the groups are of the same size or where their sizes differ. I’ll address these more complex kinds of non-identity decisions, and how the theories we have in hand so far deal with them, shortly.

That TU deals intuitively with simple non-identity decisions should be fairly obvious. Personal identity is irrelevant to TU. In a decision about which person to create, the only relevant information is the wellbeing of each possible person. Creating the person with higher wellbeing will result in a population with a higher total wellbeing. So, one ought to create that person. The same is true according to AU. Whatever the effect on average wellbeing of creating a person with some particular wellbeing, creating a person with a higher wellbeing instead will result in a population with a higher average wellbeing. Depending on the details of the decision, it might be that both acts of creating a person decrease the average wellbeing of the population. Still, creating the person
3.3. NON-IDENTITY

with higher wellbeing will decrease it by less.

HU and NHU do not deal intuitively with even simple non-identity decisions. This is the well-known Achilles’ heel of person-affecting theories, and when it comes to making such decisions HU and NHU are equivalent. In a decision between creating one happy person or another, everyone in the population of each available outcome has their maximum wellbeing in that outcome. This is because they either have the same wellbeing in both outcomes, or they don’t exist in the other available outcome and thus have either no wellbeing (HU) or wellbeing zero (NHU) in that other outcome. Consider the non-identity decision of the prospective mother again, \(D_{18}\) above. It is clear that HU and NHU deliver the counterintuitive conclusion in that decision, that it is not the case that the woman ought to wait to conceive the child who will live a happier life. Regardless whether the woman waits or conceives now, no-one who exists will be harmed. It is true according to NHU that each possible child is harmed in the outcome in which she/he does not exist. However, these harms are incurred by merely possible people and thus do not count against those outcomes. (We saw in the previous section that this is what enables NHU to agree with the neutrality intuition.) According to both HU and NHU, \(A\) and \(B\) are equally choice-worthy in \(D_{18}\).

This concludes my main point in this section: TU and AU deal intuitively with simple non-identity decisions, and HU and NHU do not. Before I move on to my fourth and final intuition about variable population decisions, I’ll just briefly consider some of the more complex kinds of non-identity decisions I mentioned earlier. Here, for example, is a case (not particularly realistic, but clear) involving a decision between creating one group of people, and creating an entirely distinct group of people of the same size. Imagine the planet Earth is about to be destroyed. We Earthlings have made Mars habitable and arranged things so that as our last act we can send a set of gamete-pairs safely to Mars where they will become persons and develop a new society. We face a decision about which set of gamete-pairs to send: a healthy set, or a set with a genetic disorder that will mean the people they produce have lower wellbeing. Here is the decision.

\[
D_{20}:
\begin{array}{cccc}
& x \text{ billion Earthlings} & y \text{ billion Martians} & y \text{ billion Martians} \\
A & 80, 80, \ldots, 80 & 80, 80, \ldots, 80 & * , * , \ldots, * \\
B & 80, 80, \ldots, 80 & * , * , \ldots, * & 60, 60, \ldots, 60
\end{array}
\]
CHAPTER 3. VARIABLE POPULATION DECISIONS

My non-identity intuition—the intuition that one ought to create happier rather than less happy, non-identical people—is just as strong, if not stronger, in this case. Here we might make it the case not just that one person exists at a lower wellbeing rather than a different person existing at a higher wellbeing, but that many people exist at a lower wellbeing rather than different people existing at a higher wellbeing. The accounts I gave of how TU, AU, HU, and NHU deal with simple non-identity decisions extend straightforwardly to decisions of this kind, between groups of people of the same size. TU and AU agree with my intuition; HU and NHU do not.

A slight variation of this case is a case in which our decision will affect not only which people will come to exist and what wellbeings they will have, but also how many people will come to exist. Imagine (as is perhaps more plausible) that the decision of which pair of gametes we send is of this more complex kind. It is a decision between creating one group of people with some (uniform) wellbeing, and creating a different number of non-overlapping people with a different (but also uniform) wellbeing. Imagine it is this decision.

\[
D_{21}: \begin{array}{c|c|c}
   & x \text{ billion Earthlings} & y \text{ billion Martians} \\
\hline
A & 80, 80, \ldots, 80 & 80, 80, \ldots, 80 \\
B & 80, 80, \ldots, 80 & *, *, \ldots, * \\
\end{array}
\]

HU and NHU deal the same way with this more complex non-identity decision as they do with the others. Interestingly, I am unable to say how AU and TU will make this kind of decision in general. I can say that according to AU, however \(x, y,\) and \(z\) are filled in, \(A\) is more choice-worthy in \(D_{21}\) than \(B\). However, if the Earthlings’ wellbeings were sufficiently higher, and \(y\) sufficiently greater than \(z\), this choice-worthiness ordering according to AU would be reversed. This is because sometimes adding fewer people at a lower wellbeing decreases the average wellbeing by less, than adding more people at a higher wellbeing. As for TU: what TU would say about \(D_{21}\) is unclear until the values of \(x, y,\) and \(z\) are filled in. In particular, if \(z\) is sufficiently larger than \(y\), TU will conclude that \(B\) is more choice-worthy in \(D_{21}\) than \(A\). This is because the sum of many smaller amounts can be greater than the sum of fewer greater amounts. This leads neatly to my final intuition, which concerns this kind of decision.

This concludes my discussion of my non-identity intuition, and how the general utilitarian theories we have encountered so far measure up against it. Now I’ll introduce my final intuition about variable population decisions.
3.4 Repugnance

Imagine you’re in the first generation of people (perhaps you’re Adam or Eve; perhaps you’re a member of a larger initial population), and you face a decision between two outcomes. In outcome $A$ the population of the world increases at a moderate rate over time until, at time $t$, the Earth is hit by a meteor and everyone dies instantly. The population of $A$ is large—say, a trillion—and everyone in the population has high wellbeing in $A$—say, wellbeing 100. In the other available outcome, $B$, the population of the world increases at a rapid rate over time until, at time $t$, the Earth is hit by the meteor and everyone dies instantly. The population of $B$ is humongous—many, many times the size of the population of $A$—and everyone in the population has just barely positive wellbeing in $B$—say, wellbeing 1. My intuition is that, faced with this decision, you ought not bring about $B$, regardless how large the population of $B$ is. The conclusion that you ought to bring about $B$ (or even that it is permissible to bring about $B$) is, I think, repugnant.

‘Repugnance’ is a strong word. I use it because it is associated with this kind of decision in the population ethics literature. I don’t mean to imply that this conclusion is more counterintuitive than those I’ve already discussed, concerning wretchedness, neutrality, and non-identity. The use of ‘repugnance’ in this context originates with Parfit, whose ‘Repugnant Conclusion’ is this:

For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better, even though its members have lives that are barely worth living (1984, 388).

Parfit focusses here on an axiological claim about which population is better

---

32 The reason for setting this decision in the first generation of people is to make it plausible that the decision affects the wellbeings of everyone who ever lives. I could stipulate that no-one has yet died, to emphasise that no-one’s wellbeing is yet set (though some people think that a person’s wellbeing can still be affected after they have died—for example if their will is not carried out or their loved ones fare poorly).

33 When pressed on the question of the significance of each issue recently, I said that I was probably most reluctant to give up my intuitions about wretchedness and repugnance, perhaps less strongly attached to my intuitions about neutrality and non-identity. I don’t think this is a particularly useful line of thought, however, and my attitudes have been known to shift on the basis of argument and what the possible theories seem to be. In the not-too-distant past I thought that, on balance, giving up my intuition about wretchedness was the best option.
CHAPTER 3. VARIABLE POPULATION DECISIONS

than which. My focus is again on decisions of a particular kind, and how they ought to be made. The intuition I endorse is that it is always the case that, in a two-way decision between an outcome in which many people exist with uniformly high wellbeing and an outcome in which many more people exist with uniformly much lower wellbeing, other things being equal, one ought to choose the former (the former is more choice-worthy in such decisions than the latter). I shall refer to decisions like this as ‘repugnant decisions’ (without, of course, meaning to imply that there is anything objectionable about merely facing a decision of this kind). Though anti-repugnance intuitions are widespread, mine is quite a strong version of the intuition.

In my description of the case above I didn’t stipulate to what extent the populations of A and B overlap. As I set the case up, the populations must overlap at least to some extent: at the very least you the decision-maker exist no matter what. The population of B might include all of the members of the population of A, in which case the decision looks like this.

\[
D_{22}: \begin{array}{c|cc}
1 \text{ trillion people} & x \text{ trillion people} \\
\hline
A & 100, 100, \ldots, 100 & \ast, \ast, \ldots, \ast \\
B & 1, 1, \ldots, 1 & 1, 1, \ldots, 1 \\
\end{array}
\]

If we can imagine a decision being made by some non-person agent outside the moral calculus—a god, perhaps—we can imagine a decision in which there is no overlap of the two populations at all. This would be a non-identity decision and look like this.

\[
D_{23}: \begin{array}{c|cc}
1 \text{ trillion people} & (x + 1) \text{ trillion people} \\
\hline
A & 100, 100, \ldots, 100 & \ast, \ast, \ldots, \ast \\
B & \ast, \ast, \ldots, \ast & 1, 1, \ldots, 1 \\
\end{array}
\]

My focus will be on decisions like the former, but I’ll also say something about decision of this latter kind.

Once again, my intuitions are widely but not unanimously shared. And among those who share my intuitions about repugnant decisions, some have found themselves nonetheless compelled to abandon them and accept a moral

---

34 He is implicitly talking about overall betterness here, I think.

35 At least, it is strong in one way. My intuition is that in all repugnant decisions, no matter how small the difference in wellbeing in question, one ought to choose the outcome in which individuals have higher wellbeings. However, my intuition is rather weak in that it concerns only two-way decisions (like my other core intuitions). I will address some more complex decisions, later in this section and in Section 5.3, but they aren’t part of my anti-repugnance intuition as such.
3.4. REPUGNANCE

theory that implies some repugnant conclusions. Let me survey some of the reasons given in favour of accepting repugnant conclusions.

Some have argued that the wretchedness threshold is much higher than people are initially inclined to assume. Mackie argues that the ‘level that is really marginally better than nonexistence must already constitute a high degree of flourishing, and beyond this little further improvement is possible’ (1985, quoted in Tännsjö 2002, 345). According to Tännsjö the lives of affluent western people today are at approximately this level. Tännsjö says that he can’t conceive of human lives significantly higher in wellbeing than ‘our own’. He restricts the scale of human wellbeing to two categories: ours—affluent western, just happy; or worse—wretched. Repugnant decisions thus become impossible. There is no scope for a decision between people at one happy wellbeing, and people at a higher happy wellbeing. This is the extreme version of the general claim that the wretchedness threshold is high and the range of happy wellbeings small.

My first objection to this kind of claim is that it seems implausible and at best contingently true. As such, it is an unsatisfying response to the problem at hand. As Tännsjö acknowledges, many people can imagine their wellbeing substantially decreasing and yet remaining above the wretchedness threshold. Some even claim that this has happened to them over time. Many of us (even affluent westerners!) are also inclined to believe that it would be possible for our lives to be substantially improved. Tännsjö claims that we are all mistaken. If our lives are at or above the wretchedness threshold (if it would not be impermissible to create someone to live our life, other things being equal), then our lives are as good as it gets. Even if this were true, it would surely be contingently true. Surely I can at least imagine a life much better than the best actual life. In which case, I think we should want to know what our moral obligations would be with respect to such better lives. If the range of happy lives were great (as we intuitively, and Tännsjö claims mistakenly, believe): what then?

My second objection is that this response, in any but its most extreme form, does not seem to facilitate avoidance of repugnant conclusions completely. This indicates that the response misidentifies where the repugnance of repugnant conclusions lies. Hypothesising a narrower range of happy lives means that the difference in wellbeing possible in repugnant decisions is less. If the repugnance
CHAPTER 3. VARIABLE POPULATION DECISIONS

of the conclusion is roughly that a decrease in quality of life can be compensated for by an increase in quantity of life, this approach lessens the repugnance but does not eliminate it. Only if there is one single level of happy wellbeing is this phenomenon eliminated entirely (and even then, it is eliminated by the unsatisfying method of denying the possibility of the hypothesised scenario). In this I reveal, perhaps, that my conception of repugnance sets a higher bar than, say, Parfit’s. I consider repugnant the conclusion that it is ever the case that substantial decreases in wellbeing can be outweighed by great enough increases in population, in two-way decisions. I think that the repugnance of the claim that \( B \) is more choice-worthy in \( D_{22} \) than \( A \) lies in this general feature, rather than in any details regarding what lives at wellbeing 100 and lives at wellbeing 1 are like. (Remember, I haven’t said much at all about the wellbeing scale. The difference between wellbeing 100 and wellbeing 1 might be made by a pin prick or by ninety-nine years of torture, for all I’ve said so far.)

A number of people have also proposed some possible reasons not to trust intuitions against repugnance. Tännström points out that we should be wary lest such intuitions are influenced by attitudes to population growth presently, in our actual world (2002). It does not follow immediately from accepting a repugnant conclusion that we should increase our population as much as and whenever possible. No plausible moral theory says that a larger less well-off population is always more choice-worthy than a smaller better-off population, in a decision between the two. The decisions we currently find ourselves facing—whether to have children, or which public policies to adopt—may well be decisions in which creating less people is more choice-worthy. Prior conviction that this is the case would not provide any support for the anti-repugnant intuition I adopt. I agree that this is an important point. However, I am confident (as confident as I can be) that my anti-repugnance intuitions are not influenced by real-world concerns in this way.

Another proposed reason for mistrusting our intuitions about repugnance, endorsed for example by Broome (2004), is that they are intuitions about large numbers (of people) and we are notoriously bad at thinking about large numbers. In my experience, a lot of people find this thought compelling. The case

\[36\text{See Holmg 2004 and Arrhenius 2004 for similar arguments. In response to a similar point Tännström retreats to the claim that even if the repugnant conclusion as he understands it is false, it is not repugnant in the sense of being obviously false. This does not seem to me to be a solution to anything. The problems remain even if it is no longer repugnant but merely false.}\]
3.4. REPUGNANCE

I gave deals with trillions of people. The repugnance cases discussed in the literature often mention very large numbers. Arguably, people lack the capacity to properly imagine such large populations, and to judge the significance of differences between very large numbers. If so, the argument goes, we shouldn’t trust our intuitions about repugnance. I don’t think this argument works, at least against intuitions of the kind I have, because I don’t think those intuitions rely on the large numbers in these cases. I have consistently anti-repugnance intuitions in smaller number cases. I also have an intuition that a general anti-repugnance principle is true, which takes the form of a universal claim and does not, I think, reply on my being able to imagine large populations.

It may help at this point to provide an analogy regarding fixed population decisions. Consider a fixed population decision between giving higher or lower wellbeings to each of five people, other things being equal. I have a strong intuition that one ought to give those people the higher wellbeings. I have this intuition when the number of people concerned rises from five to six, to ten, to 150, and so on. I have this intuition to the limits of my capacity to imagine the number of people involved, and beyond. I also have the intuition regardless how small the difference in wellbeing concerned is. I have an intuition of some general, universal claim, to the effect that one ought to give higher rather than lower wellbeings to any number of people, other things being equal. D_{24} might be any of these decisions.

\[
D_{24} : \begin{array}{|c|c|}
\hline
x \text{ people} \\
\hline
A & 58, 58, \ldots, 58 \\
B & 57, 57, \ldots, 57 \\
\hline
\end{array}
\]

I have the intuition that A is more choice-worthy in D_{24} than B, regardless what x is, and even if x is some vast number that I cannot properly imagine.

Now consider a repugnant decision on a small scale.

\[
D_{25} : \begin{array}{|c|c|c|c|}
\hline
p_1 & p_2 & 100 \text{ people} \\
\hline
A & 50 & 50 & * , * , \ldots , * \\
B & 1 & 1 & 1 , 1 , \ldots , 1 \\
\hline
\end{array}
\]

I have the intuition that A is more choice-worthy in D_{25} than B. I have this intuition when the number of people at wellbeing 50 in A rises from two to three, to ten, to 150, and as the number of people in B rises from 100 to 200, to 1000, and so on. I have this intuition to the limits of my capacity to imagine the
number of people involved, and beyond. I also have the intuition regardless how small the difference in wellbeing concerned is.\textsuperscript{37} I have the intuition of a general, universal claim to the effect that one ought not create larger, less happy populations rather than smaller, happier populations, other things being equal in a decision between the two.

Some people, I realise, have less robust anti-repugnance intuitions than I have.\textsuperscript{38} In any case, I think that these reasons to be suspicious of our anti-repugnance intuitions should at most make us less confident in those intuitions; make them a better target for revision than other intuitions, if a choice has to be made. In this dissertation I propose a theory that is consistent with the intuition of anti-repugnance without conflicting with any more secure intuitions, so this role is unnecessary.

Finally, before I discuss how TU, AU, HU, and NHU fare with respect to repugnance, a note on the relationship between repugnance and neutrality. Neutrality and anti-repugnance seem in some sense to go together—at least within the welfarist axiological consequentialist framework. Abandoning one seems to threaten losing the other. If one ought to create a happy person (if we give up neutrality), it is hard to avoid the conclusion that one sometimes ought to add many happy people at the cost of a reduction in wellbeing in repugnant ways. Seen from the other direction: if one sometimes ought to choose an outcome with a larger less happy population rather than one with a smaller more happy population, it is hard to account for this in a way that doesn’t imply that one ought (at least sometimes) to create a happy person rather than not.\textsuperscript{39}

### 3.4.1 Repugnance: the fourth test

Repugnant decisions provide my fourth and final test of general utilitarian theories. I hope to find a theory according to which it is always the case that, in a decision between an outcome with a larger population of uniformly less

\textsuperscript{37}Again, this is one way in which my intuition may be stronger than others’.

\textsuperscript{38}I think that one reason people tend to discuss repugnance cases involving very large numbers is that some people have less clear intuitions in small number cases, like D\textsubscript{25}. They think that perhaps the extinction of humans is bad in some morally relevant way, or that there are things that are good in a morally relevant way that can’t be produced by small numbers of people. They may therefore be unable to imagine that things are otherwise equal, in the relevant cases. I discuss some of these issues in Section 5.3.1. In brief: I don’t think they are utilitarian concerns.

\textsuperscript{39}These things are hard but perhaps not impossible to reconcile.
happy people and an outcome with a smaller population of uniformly happier people, the latter is more choice-worthy than the former. TU fails this test, and AU passes. HU and NHU fail the test.

That TU reaches repugnant conclusions should be obvious. I pointed out in the previous section, in terms of the non-identity decision $D_{21}$, why this is the case. Sometimes, the sum of very many smaller numbers is greater than the sum of fewer larger numbers. Parfit provides the analogy that a large number of bottles with one drop of milk each may contain the greatest mass of milk (1984). This is the case for example in the small scale repugnant decision $D_{25}$: the total wellbeing of the population of $A$ in $A$ is 100, whereas the total wellbeing of the population of $B$ in $B$ is 102. According to TU, the outcome with the larger, less well-off population is more choice-worthy in that decision that the outcome with the smaller, better-off population. It is clear that this will be the case in some decisions involving very large numbers of people too.\footnote{The question of what TU says (and what other theories say) in cases involving populations of infinite size is an interesting one, but not one I will address. See for example Vallentyne 1993 and Arntzenius 2014.}

There have been a series of proposals of, as I see them, ways of amending TU in response to its counterintuitive conclusions in repugnant decisions. I shall discuss three kinds of theories produced in this way: perfectionist (or lexical value) theories, critical level theories, and variable value theories. Each of these approaches attempts to reduce the scale of repugnance, rather than eliminate it entirely. As such, I find them all fairly unsatisfying.

Parfit’s ‘perfectionist’ suggestion (2004, 18) is that ‘the best things in life’, as well as contributing significantly to people’s wellbeing, are valuable in themselves in such a way that no amount of any less-good things could be worth as much as the least amount of these best things.\footnote{Recent similar proposals include Thomas’s (2015) and Beard’s (2016).} If the best thing in life is Mozart, then any world with Mozart in it is better than every world without Mozart, regardless how much Muzak those worlds contain. The best things in life are lexically superior to the next-best things. Parfit appeals to this idea to explain why, at least often, one ought to choose an outcome with a population with very high wellbeing rather than an outcome with a population with low wellbeing. The population with very high wellbeing is likely to contain the best things in life, and the population with low wellbeing is likely not to contain those things. One difficulty with this approach, in the present context, is that
it violates welfarism. It makes evaluations of outcomes depend on something other than wellbeings, such as whether a small amount of Mozart or a large amount of Muzak produces those wellbeings. The effectiveness of this proposal also depends on the correlation of the best things in life with differences in wellbeing levels. It does not address cases of repugnance in which wellbeings vary without any loss of the best things (or the second-best things, and so on). A final difficulty is to see how the very special things will be identified.

The next proposal is to introduce a critical level of wellbeing. This approach is a little difficult to describe in my terminology and framework, due to a difference in method, and in particular the somewhat idiosyncratic way I have defined the wretchedness threshold. I defined a normative wretchedness threshold, as the level of wellbeing below which one ought not create a person. I (fairly arbitrarily) assigned that wellbeing zero on the wellbeing scale. I remained agnostic about how this normative wretchedness threshold is to be identified: is it the level of wellbeing such that a life at that level is equally as good for a person as non-existence; the level of wellbeing such that it is better for a person to live the shortest possible truncation of a life at that level? An alternative method is to settle (at least temporarily) on one of these ways of identifying a significant wellbeing, and assign to it both wellbeing zero and the title ‘wretchedness threshold’. Using that terminology, it is possible to debate the moral significance of the wretchedness threshold. One can meaningfully say that it is not always the case that one ought not create a wretched person, other things being equal; or that there is some critical level of wellbeing, above the wretchedness threshold, such that one ought not create a person below that level (and ought to create a person above that level, in keeping with TU). This is how critical level theories have been presented.

---

42 Another possibility, which would overcome this particular difficulty, is that the perfectionism operates at the level of determining wellbeings.

43 Again, defending a correlation of wellbeings and the best things in life would address this. That would be a significant claim and difficult to defend. As I wish to refrain from commitment to any particular accounts of wellbeing I won’t address this possibility further.

44 See Blackorby et al. 2004, Broome 2004, and Mulgan 2004. Mulgan proposes that the critical level is context dependant. Any given society will develop (according to Rawlsian contractualist methods) a wretchedness threshold, depending on current levels of wellbeing in that society.
As I’ve set things up, TU and critical level theories are equivalent. Or rather, TU as I’ve defined it just is the ‘correct’ critical level theory. In my terminology, the difference between TU and any other critical level theory is a matter of where the (normative) wretchedness threshold is with respect to the level of wellbeing at which a life is ‘worth living’. I’ve remained agnostic on this question in my definition of TU, and defined TU as the theory that maximises the total wellbeing of populations, where wellbeings are numerically represented in such a way that zero is the (actual, normative) wretchedness threshold. This should explain why the critical level approach is not an effective response to my anti-repugnant intuitions. The intuitive idea of the approach is meant to be that quantity of life can outweigh quality only down to some level of wellbeing at which this is intuitively less repugnant. In the absence of a lot of information about where the wretchedness threshold is, what lives at that level are like, and so on, this is an ineffective response. Also, this approach leads to some counterintuitive results in cases involving wretched people.45

The final amendment to TU that I will discuss is the proposal that the value of a life (or of the wellbeing of the person living a life) varies not only according to the level of wellbeing of that life, but also according to population size. So, for example, the ten-millionth person with some particular wellbeing might still be a very valuable addition to a population, but the 100-millionth person with the same wellbeing will be a much less valuable addition, approaching a value of zero. If the value decreases asymptotically, there will be a maximum value for any population of people at a given wellbeing. Reductions in ‘quality’ of wellbeing for the sake of ‘quantity’ will be limited by these maximum values.46

Once again, the variable value response is a limited response to repugnance. It also seems very strange that the value of an additional life should depend on the existence of seemingly unrelated factors, such as how many people have lived lives at that wellbeing in the distant past, or will do so in the distant future.

This concludes my very brief sketch of some of the proposals for adjustments to TU in response to its repugnant conclusions. They all seek to reduce rather than eliminate repugnant conclusions. There are more sophisticated defences

---

45See Arrhenius 2000 on the ‘sadistic conclusion’. The higher the ‘critical level’ (normative-wretchedness threshold), the less repugnant conclusions but the more sadistic conclusions are produced.

46See Hurka 1983 and Ng 1989 on variable values.
of these alterations in the literature, but I don’t have time to go into them here. Now I’ll move on to the other familiar theories.

AU passes the repugnance test with flying colours. According to AU, quality of wellbeing ought not be compromised for quantity of wellbeing. In any decision between an outcome with a (uniformly) less well-off happy population and an outcome with a (uniformly) better-off population, AU says that the outcome with the better-off population is more choice-worthy, regardless how large the two populations are.

HU and NHU are equivalent when it comes to repugnant decisions. They coincide with my anti-repugnant intuitions only in decisions in which the populations of the two available outcomes overlap. So HU and NHU say that A is more choice-worthy in D_{25} than B, because in B p_1 and p_2 are harmed, while no member of the population of A is harmed in A. The total harm in B is 49 \times 2 = 98 and the total harm in A is 0. We can see that this will also tell against repugnance even in a decision in which there is less than full overlap in the available populations; for example, in this decision.

\[
\begin{array}{c|c|c}
D_{26}: & p_1 & p_2 \\
& 50 & 50 \\
& 1 & \ast \\
A & \ast, \ast, \ldots, \ast \\
B & 1, 1, \ldots, 1 \\
\end{array}
\]

This decision is like D_{25} except that instead of p_2 existing in B with wellbeing 1, some other person exists in B with wellbeing 1. HU and NHU still conclude that A is more choice-worthy in D_{26} than B, but it seems that the moral significance of the decision is less. In this decision the total harm in B is only 49 (and the total harm in A is still 0). The next step is a non-identity decision.

\[
\begin{array}{c|c|c}
D_{27}: & p_1 & p_2 \\
& 50 & 50 \\
& \ast & \ast \\
A & \ast, \ast, \ldots, \ast \\
B & 1, 1, \ldots, 1 \\
\end{array}
\]

According to HU and NHU, A and B are equally choice-worthy in D_{27}. The total harm in each outcome is 0. Of course we knew already that HU and NHU make counterintuitive conclusions in non-identity decision, so this should come as no surprise.
3.4. REPUGNANCE

3.4.2 Mere addition

So far in this section I’ve focussed on repugnant decisions directly. In the
literature on repugnance there is also a much-discussed argument from non-
repugnant decisions to repugnance. This is the mere addition argument. The conclusion of the mere addition argument is a repugnant one. It is often
considered to present a general problem (not just for TU): one premise must be
rejected to avoid repugnance, but each premise seems compelling.

I’ll set out the argument in terms of choice-worthiness, though it is often,
I think, intended to be made in terms of overall betterness. I’ll explain why
the difference matters shortly. The argument proceeds by reference to three
decisions. Here is the first decision. (Of course the argument needn’t use
exactly these three decisions, just three with the same structure.)

\[
\begin{array}{ccc}
D_{28}: & 10 \text{ million people} & 3 \text{ million people} \\
A & 100,100,\ldots,100 & *,*,\ldots,* \\
A^+ & 100,100,\ldots,100 & 70,70,\ldots,70 \\
\end{array}
\]

The first premise of the argument is that \(A^+\) is not less choice-worthy than \(A\). This premise is based on intuition. \(A^+\) differs from \(A\) only by the mere
addition (hence the name of the argument) of some happy people. It is in the
spirit of neutrality that this does not make it less choice-worthy, in a decision
between the two, than \(A\). Here is the second decision.

\[
\begin{array}{ccc}
D_{29}: & 10 \text{ million people} & 3 \text{ million people} \\
A^+ & 100,100,\ldots,100 & 70,70,\ldots,70 \\
B & 95,95,\ldots,95 & 95,95,\ldots,95 \\
\end{array}
\]

The second premise is that \(B\) is more choice-worthy than \(A^+\). This is a
fixed population decision and the second premise is entailed by CFPU (and, of
course, by each of the general utilitarian theories we have encountered so far).
Finally, here is the third decision.

---

48 There is also a ‘benign addition’ version of this argument. In that version the wellbeings of the
members of the population of \(A\) are slightly higher in \(A^+\) than they are in \(A\). I address that version
of the argument in Section 5.3.
49 There is inequality of wellbeing in \(A^+\) and not in \(A\). One might think that this tells against
\(A^+\), making \(A^+\) less choice-worthy in this decision than \(A\). However, I think that we should expect
equality to be irrelevant according to a general utilitarian theory. After all, it is irrelevant according
to CFPU.
The conclusion of the argument is that $B$ is more choice-worthy in $D_{30}$ than $A$. And if the argument goes through, then the stepping down process can be repeated to reach the conclusion that a very large population of people with barely happy lives is more choice-worthy than $A$, in a decision between the two.\footnote{At least, it’s not obvious why it couldn’t be repeated in this way.}

The conclusion \textit{would follow} from the premises if we presented the argument in terms of overall betterness, rather than choice-worthiness. If $A^{+}$ is at least as good overall as $A$ (i.e. not worse than, assuming completeness of the overall betterness relation), and $B$ is better overall than $A^{+}$, then $B$ must be better overall than $A$. We cannot account for the three very compelling normative intuitions concerning these three decisions in terms of overall betterness. I think that this provides further reason to question the overall betterness requirement.

The conclusion does not follow from the premises as I presented them, in terms of choice-worthiness. HU, for example, provides an axiological justification of the two premises of the mere addition argument and the negation of its conclusion. $A^{+}$ is no less choice-worthy in $D_{28}$ than $A$ because the total harm in each in that decision is 0. $B$ is more choice-worthy in $D_{29}$ than $A^{+}$ because the total harm in $B$ in that decision is less than the total harm in $A^{+}$. And $B$ is not more choice-worthy in $D_{30}$ than $A$ because the total harm in $B$ in that decision is greater than the total harm in $A$.\footnote{I don’t meant to claim that these are the correct justifications, nor that HU gets things right in all mere addition cases. I just mean to illustrate that there is a consistent axiological explanation of why the conclusion doesn’t follow from the premises of the version of the mere addition argument that I have presented; albeit one that does not make use of, and is not able to be captured in, an overall betterness relation.} This is an advantage of HU over TU and AU, and a desirable feature of a general utilitarian theory. I’ll discuss the mere addition argument again in Sections 4.3 and 5.1.

To summarise my findings in this section: AU agrees with my anti-repugnance intuition, while HU, NHU, and TU do not. The possibilities for altering TU that I have discussed—the perfectionist, critical level, and variable value methods—go some way toward reducing the extent of the counterintuitive conclusions, but don’t eliminate them entirely. Also, they each face novel problems of their own. The overall goodness version of the mere addition argument is valid but
possibly unsound (the overall betterness premises are not compelling enough to justify accepting the conclusion); and the choice-worthiness version of the argument is unsound (though it has very intuitively compelling premises).

Conclusion

In this chapter I have presented four of my strongest intuitions about variable populations decisions. These intuitions, or versions thereof, are widely shared. They are as follows. There is a normative wretchedness threshold: a wellbeing level such that one ought not create a person who will live a life below that wellbeing, other things being equal. Creating a happy person is morally neutral: it is not the case that one ought to create a happy person (nor that one ought not create her), other things being equal. One ought to create a happier person rather than a different less-happy person, other things being equal. And it is never the case that one ought to create a larger less happy population rather than a smaller happier one. My core intuitions concern the relevant two-way decisions, though I have also endorsed intuitions about some more complex decisions.

I hope to find a general utilitarian theory that agrees with all four of these intuitions. None of the theories that we have encountered so far does so, and it has been established elsewhere that no moral theory that meets the overall betterness requirement does so. In the next chapter I will propose Shortfall Utilitarianism, a general utilitarian theory that does agree with all four intuitions. This provides a reason to favour SU over the other theories we’ve considered, a reason to reject the overall betterness requirement, and a reason to further explore the class of general utilitarian theories.
Chapter 4

Shortfall Utilitarianism

I shall now propose Shortfall Utilitarianism (SU). This chapter has four sections. In Section 4.1 I set out SU as clearly as I can, while also offering some motivation for it. In Section 4.2 I show that SU is a general utilitarian theory: that it is an extension of CFPU, and is consequentialist, axiological, and welfarist. By way of showing that the theory is axiological, I spell out exactly what the axiology of the theory is. I also show that SU does not meet the overall betterness requirement. In Section 4.3 I show that SU agrees with my intuitions about wretchedness, neutrality, non-identity, and repugnance. This third section includes much of my support for SU (with further support and defence to follow in Chapter 5).

4.1 Introducing SU

Let me begin this section with a very brief sketch of SU. This will give the reader some idea where we’re headed; the details should become clear throughout the rest of the section.

SU identifies two kinds of worseness among outcomes, and says that if a first outcome is worse than a second in one of these ways, that counts against choosing the first rather than the second (i.e. choosing the first when the second is also available). When making a decision, one ought to choose an available outcome that is worse than any single other available outcome in these ways by the least amount. This is an available outcome such that things count against choosing it (rather than some alternative), by no more than things count against
choosing any of those alternatives. More generally, in any decision, the worse some available outcome is than a single other available outcome in these two ways, the less choice-worthy that outcome is in that decision.

The two kinds of worseness reflect two concerns. First, we want to make sure that the same people don’t exist worse off rather than better off. So, when comparing a pair of outcomes we should look to the people who exist in both. To the extent that those people are worse off, as a group, in one of those outcomes than in the other, that outcome is intersectionally worse than the other. Second, we also want to make sure that worse-off people don’t exist instead of different, better-off people. So, when comparing a pair of outcomes we should also look to the people who exist in either one of those outcomes but not both. How do the people who exist in the first outcome but not in the second fare, as a group, in the first outcome? How do the people who exist in the second outcome but not the first fare, as a group, in the second outcome? To the extent that the first group are worse off in the first outcome, than the second group are in the second outcome, the first outcome is complementarily worse than the second.1

Intersectional and complementary worseness are what I call two kinds of worseness (reflecting two ways in which one outcome can be worse than another). They are not two respects of worseness, or two worseness (or betterness) relations. The axiology of SU is more complex than this. I won’t spell out the axiology of SU fully until Section 4.2, because I can present SU more simply by way of its choice-worthiness output, without appealing to its axiology (just as I presented CFPU initially and most simply without appealing to axiology in Section 2.1). However, here is a sketch of the axiology of SU. There are many morally relevant betterness relations underpinning intersectional and complementary worseness. There is a relation of betterness with respect to each pair of identical sets of possible people, and with respect to each pair of non-overlapping sets of possible people.2 A first outcome is better than a second, with respect to a particular pair of sets of people, iff the first set of people

---

1 Thanks to Gustav Arrhenius for pointing out that SU bears some similarities to his ‘Harm Comparativism’ (2016) or ‘Soft Comparativism’ (2009a), which he defines as follows: ‘The value of an outcome is a function of both the value of the welfare and the value of the comparative harm in the outcome’ (2016, 257). Arrhenius himself cashes this out as a ‘harm-adjusted version of Total Utilitarianism’ (2016, 258) (a sort of TU-HU hybrid). However, on a broad enough interpretation of his definition, SU might be considered another, rather different, version of Harm Comparativism.

2 That is, a relation of betterness with respect to each pair of sets of people that share either all or none of their members.
are better off in the first outcome, than the second set of people are in the second outcome. Each relation is very incomplete, holding only between outcomes in which particular people exist. Which of the relations (which respects of betterness) are relevant to intersectional and complementary worseness between a given pair of outcomes also depends on which people exist in those outcomes. My point here is just to indicate the complexity of SU’s axiology. I’ll set axiology aside again now, until the next section.

That was a rough sketch of SU. Here is how I’ll set the theory out more fully. First I’ll define intersectional and complementary worseness. Then I’ll explain how intersectional and complementary worseness determine the ‘shortfall’, and thus the choice-worthiness, of each available outcome in a decision. The theory, remember, is identified by this choice-worthiness output.

4.1.1 Intersectional worseness

Intersectional worseness takes its name from the set of people who exist in both of a pair of outcomes: this set is the intersection of the populations of the pair of outcomes. I said that intersectional worseness reflects the concern that the same people don’t exist worse off rather than better off; and that one outcome is intersectionally worse than another to the extent that the people who exist in both are worse off as a group in the one, than they are in the other. My immediate goal now is to make sense of this notion of comparing outcomes in terms of how the members of the intersection of their populations fare, as a group; and to use this notion to define the intersectional worseness of one outcome than another, for any two outcomes.

First, consider how this group-wise comparison notion differs from, and how it resembles, familiar kinds of comparison. We are familiar—from Harm Utilitarianism (HU), for example—with comparisons with respect to an individual: comparing how a particular person fares in one outcome with how she fares in another; how much better or worse one outcome is than another for her. We are also familiar—from Total Utilitarianism (TU) and Average Utilitarianism (AU)—with comparisons of overall goodness of outcomes, which we might think of as comparisons with respect to whole populations: comparing how good it is overall that this population exists with these wellbeings, with how

---

3A population is a set of people, and the intersection of a pair of sets is the subset consisting of all and only members of both original sets (the ‘overlapping’ bit of a two-circle Venn diagram).
good it is overall that *that* population exists with *those* wellbeings. Group-wise comparisons are a sort of conceptual middle ground between these two familiar kinds of comparison. Like familiar comparisons of overall goodness, they involve some aggregation of elements—they make a single assessment based on all of the wellbeings of the people in the relevant group in each outcome. Like comparisons with respect to individuals, their scope is limited: they take into account only some features of outcomes (the wellbeings of the people in the group) and not others.\(^4\)

So, how do we make these group-wise wellbeing comparisons? In particular, how do we compare two outcomes with respect to the set of people who exist in both? My proposal is that we do this by reference to the average wellbeings of the members of the set in each of the relevant outcomes. Take any two outcomes, \(A\) and \(B\). Call the people who exist both in \(A\) and in \(B\)—the members of the intersection of the population of \(A\) and the population of \(B\)—the \(A\)-and-\(B\) people. The \(A\)-and-\(B\) people fare worse as a group in \(A\) than in \(B\) just in case the average wellbeing of the \(A\)-and-\(B\) people in \(A\) is lower than the average wellbeing of the \(A\)-and-\(B\) people in \(B\). Further, if the \(A\)-and-\(B\) people fare worse as a group in \(A\) than in \(B\), they fare worse by a particular amount: the difference between the average wellbeing of the \(A\)-and-\(B\) people in \(A\) and the average wellbeing of the \(A\)-and-\(B\) people in \(B\), multiplied by the number of \(A\)-and-\(B\) people.\(^5\)

Intersectional worseness is based on this kind of group-wise comparison. It is what we might call strict group-wise worseness.\(^6\) For any pair of outcomes, \(A\) and \(B\), \(A\) is intersectionally worse than \(B\) by some amount greater than or equal to zero. If the \(A\)-and-\(B\) people fare worse in \(A\) than in \(B\) (in the way just described) then \(A\) is intersectionally worse by the amount by which the \(A\) people fare worse in \(A\) than in \(B\), which will be positive. If the \(A\)-and-\(B\) people do not fare worse in \(A\) than in \(B\) (if they fare equally well in \(A\) and \(B\), or if they

---

4Sometimes group-wise comparisons coincide with comparisons with respect to an individual, or comparisons with respect to whole populations, or even both. Consider two outcomes in which only one person, Sally, exists. Comparing these outcomes with respect to the one-person group consisting of Sally amounts to the same thing as comparing them with respect to Sally-the-individual and comparing them with respect to their whole populations.

5There are other, equivalent ways of presenting this comparison; for example, in terms of total wellbeings. More on these options shortly.

6One thing is *strictly* worse than another just in case it is worse than the other thing by some positive amount. This is, I think, the most common everyday use of the term ‘worse’. I could define a symmetrical notion of (strict) intersectional betterness. I don’t do so because that notion doesn’t come into play in SU.
fare better in $A$ than in $B$) then $A$ is intersectionally worse than $B$ by zero. Here, in summary, is how to calculate by how much $A$ is intersectionally worse than $B$, for any $A$ and $B$:

**Intersectional worseness:** If the average wellbeing of the $A$-and-$B$ people in $A$ is less than the average wellbeing of the $A$-and-$B$ people in $B$, then $A$ is intersectionally worse (worse with respect to the $A$-and-$B$ people) than $B$ by the difference between those averages, multiplied by the number of $A$-and-$B$ people. If the average wellbeing of the $A$-and-$B$ people in $A$ is not less than the average wellbeing of the $A$-and-$B$ people in $B$, then $A$ is intersectionally worse than $B$ by zero.\(^7\)

This definition is pretty directly inspired by CFPU. After all, this is just the kind of comparison we are intuitively relying on when we make a fixed population decision: we want to compare how the members of the fixed population fare, as a group, in the various available outcomes. Of course, as was the case for CFPU, I could present intersectional worseness in terms of sums of wellbeings rather than averages; or in terms of differences between total harms, or in terms of some other calculation. I don’t think it matters which of these equivalent definitions we adopt. I’ve gone with a presentation in terms of average wellbeings here because I think that it might make the transition to talking about complementary worseness easier (because in that context averages, and not totals, play an important role).

Here are some outcomes to illustrate intersectional worseness.

```
<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>6</td>
<td>9</td>
<td>9</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>10</td>
<td>8</td>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>
```

The $A$-and-$B$ people are $p_1$, $p_2$, and $p_3$. Their average wellbeing in $A$ is 8; in $B$ it is 9. So, $A$ is intersectionally worse than $B$ by $1 \times 3 = 3$ (the difference in averages multiplied by the number of people in the group), and $B$ is intersectionally worse than $A$ by 0.

\(^7\)If there are no $A$-and-$B$ people (if the intersection of the populations of $A$ and $B$ is empty), then the average wellbeing of the $A$-and-$B$ people in $A$ is undefined and, I assume, not less than (nor greater than, nor equal to) the average wellbeing of the $A$-and-$B$ people in $B$ (which is also undefined). So in such a case also $A$ is intersectionally worse than $B$ by zero.
A note about what is going on here axiologically. The claim that $A$ is intersectionally worse than $B$ (by such-and-such amount) is an axiological claim. It asserts that there is some morally relevant betterness relation according to which $A$ is strictly worse than $B$ by that amount. However, (as I pointed out above) there is no single relation of intersectional betterness. Rather, there are a series of betterness relations, each being the relation of betterness with respect to a particular pair of sets of people. ‘$A$ is intersectionally worse than $B$’ is shorthand for ‘$A$ is strictly worse with respect to the pair of (identical) sets, ($A$-and-$B$ set; $A$-and-$B$ set), than $B$’.\footnote{In the context of intersectional worseness it is simpler, and unproblematic, to think of it as the relation of betterness with respect to a single set (the $A$-and-$B$ set). The pairs become important later.} It is a claim about worseness in terms of the particular relation of betterness with respect to this pair of sets of people. Other relations of better- and worseness between $A$ and $B$ hold too. (This will become more obvious when I spell out the full axiology in the next section.) But those relations are irrelevant to intersectional worseness between $A$ and $B$.

4.1.2 Intersectional Utilitarianism

4.1.2 Intersectional Utilitarianism

At this point I would like to pause my presentation of SU briefly, to acknowledge another general utilitarian theory. I’ll call this other theory Intersectional Utilitarianism (IU). IU is a lot like SU, but whereas SU acknowledges two kinds of worseness—intersectional and complementary—IU acknowledges only intersectional worseness. One reason I want to pause to examine this simpler theory is that this theory is very similar to, but I think in one way an improvement on, HU. People who are strongly attracted to HU or Non-existence Harm Utilitarianism (NHU), and not perturbed by the non-identity problem, might find IU more appealing than the more complex SU. (Of course I am perturbed by the non-identity problem, and that’s one of the reasons I favour SU over IU.) Another reason to make this detour is that it might help the reader to become familiar with some aspects of the normative structure that IU and SU share, before I complicate things by introducing complementary worseness.

Here is how IU works. For any pair of outcomes, we can say by how much each one is intersectionally worse than the other. If a first outcome is positively intersectionally worse than a second (i.e. worse by some amount greater than zero), this tells against choosing the first instead of the second. In a decision,
we can make this kind of comparison between each pair of available outcomes, and one ought to choose the available outcome that is intersectionally worse than any single other available outcome by the least amount. More generally, in any decision, a first available outcome is more choice-worthy than a second just in case the first is intersectionally worse than any single available outcome by less than is the second.

**Intersectional Utilitarianism (IU):** For any decision, $D_1$, one available outcome, $A$, is at least as choice-worthy in $D_1$ as a second available outcome, $B$, just in case the maximum amount by which $A$ is intersectionally worse than any available outcome in $D_1$ is no greater than the maximum amount by which $B$ is intersectionally worse than any available outcome in $D_1$.

Consider what this theory says about a decision between the two outcomes we compared above.

<table>
<thead>
<tr>
<th>$D_{31}$</th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>6</td>
<td>9</td>
<td>9</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>10</td>
<td>8</td>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>

$A$ is intersectionally worse than $B$ by 3 and $B$ is intersectionally worse than $A$ by 0. Every outcome is intersectionally worse than itself by 0. So, the maximum amount by which $A$ is intersectionally worse than any available outcome in $D_{31}$ is 3, and the maximum amount by which $B$ is intersectionally worse than any available outcome in $D_{31}$ is 0. $B$ is at least as choice-worthy in $D_{31}$ as $A$ and $A$ is not at least as choice-worthy in $D_{31}$ as $B$: in other words, $B$ is more choice-worthy in $D_{31}$ than $A$.

As it happens, in $D_{31}$ IU and HU (and NHU) deliver the same choice-worthiness ordering. That is not always the case. In some decisions I think IU delivers more plausible choice-worthiness orderings than these harm minimising theories. Here is such a decision.

<table>
<thead>
<tr>
<th>$D_{32}$</th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>5</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>$C$</td>
<td>10</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>$D$</td>
<td>10</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>
Each of the available outcomes in $D_{32}$ is intersectionally worse than each other available outcome by zero. This means that all four outcomes are equally choice-worthy in this decision according to IU. On the other hand, in each of $B$, $C$, and $D$ (but not in $A$) some harm is incurred. So according to HU and NHU, $A$ is more choice-worthy in $D_{32}$ than $B$, $C$, and $D$. This is an instance of the general implication of HU and NHU, that one ought not create people, other things being equal, unless one can create everyone with their maximum available wellbeing.\(^9\)

I find this implication of HU and NHU counterintuitive. My intuition is that in $D_{32}$ all available outcomes are equally choice-worthy. My reasoning proceeds roughly like this. Creating people is morally neutral, so nothing tells against $A$. If you create people, you ought to create people so that they are as well off as possible. Now, in each of $B$, $C$, and $D$, some individual or other is not as well off as possible. But this doesn’t seem to me to mean that people are not as well off as possible, in the relevant sense. Say I’m considering $B$, in which $p_1$ is not as well off as possible. It doesn’t seem right to me to say that people are not as well off as possible in $B$ because $p_1$ incurs a harm there. If people are not as well off as possible in $B$ there should be some other available outcome in which people are better off, but here this doesn’t seem to be the case. In each other outcome either no-one exists, or some other individual incurs a harm of the same amount: it would seem very odd to me to say that people are better off in one of those outcomes than in $B$.

To me, $D_{32}$ seems morally indistinguishable from this slightly different decision, which HU and NHU treat differently.

\[
\begin{array}{c|ccc}
 & p_1 & p_2 & p_3 \\
\hline
A & * & * & * \\
B & 5 & 10 & 10 \\
B & 5 & 10 & 10 \\
B & 5 & 10 & 10 \\
\end{array}
\]

HU and NHU agree with IU (and with my intuitions) in $D_{33}$. There is no

---

\(^9\)This problem (as it seemed to them and to me) for HU was first put to me by Michael McDermott; and persuasively conveyed by Martin Pickup and John Cusbert. Ross (2015, 443) calls it ‘the problem of improvable life avoidance’, provides some great cases to illustrate it, and argues convincingly that it is a problem. He also suggests a revision of person affecting theories (in particular, Temkin’s) not unlike the shift from HU to IU. Note that the problem does not arise only in decisions in which one of the available outcomes has the empty population. This just makes for the simplest decisions to illustrate the problem.
harm incurred in any of the available outcomes in \( D_{33} \), and so all available outcomes are equally choice-worthy.\(^{10}\) Each individual is as well off in each available outcome in which she exists as she is in any other available outcome in which she exists. According to HU and NHU this makes all the difference. My suggestion is that this makes no difference.

I don’t expect to have convinced everyone who is partial to a harm minimising theory, that IU is an improvement on those theories. I haven’t even convinced myself that this is the case. For one thing, IU is a theory according to which there is no wretchedness threshold, which I think makes it in one way less appealing than NHU. However, there may be scope for amending IU in this regard (as HU is amended to produce NHU). And I think that it is useful to have in mind possible variations on the theme of SU; after all, my secondary goal in this dissertation is to encourage further exploration of the class of general utilitarian theories. Now, though, I’ll return to spelling out SU; and in particular to defining complementary worseness.

### 4.1.3 Complementary worseness

Complementary worseness takes its name from the set of people who exist in one of a pair of outcomes, but not in the other: this set is the \textit{complement} of the population of one outcome in another.\(^{11}\) I said that complementary worseness reflects the concern that worse-off people don’t exist instead of different, better-off people; and that one outcome is complementarily worse than another to the extent that the people who exist in the former outcome but not the latter fare worse (as a group, in the former) than the people who exist in the latter outcome but not the former do (as a group, in the latter). (Describing this comparison can get a bit confusing, but the idea is pretty simple, really.) My immediate goal now is to make sense of this notion of comparing outcomes in terms of how the members of the complementary sets of people fare, as groups; and to

---

\(^{10}\)This is the first decision I’ve discussed in which multiple available outcomes are identical. Remember that presenting decisions as sets of outcomes is a handy abbreviation we can make in the context of consequentialism. Decisions are more properly sets of acts, so \( D_{33} \) is a set of four different acts, three of which will lead to the same outcome, \( B \). I could also make the same point using a more complex decision in which the three available \( B \)'s are replaced by three different non-identical outcomes (distinguished perhaps by the existence of a different but equally happy person in each one).

\(^{11}\)A population is a set of people, and the (relative) complement of a first set in a second is the subset consisting of all and only the members of the second set that are not also in the first set (the ‘non-overlapping’ part of one of the circles in a two-circle venn diagram).
use this notion to define the complementary worseness of one outcome than another, for any two outcomes.\textsuperscript{12}

We’re familiar by now, from my discussion of intersectional worseness, with group-wise wellbeing comparisons of a particular kind: comparing how well the members of a particular set of people fare, as a group, in two outcomes in which they all exist. Now, we want to do something less straightforward. We want to make interpersonal (or inter-set) group-wise comparisons: to compare how the members of one set of people fare as a group in one outcome, with how the members of a different, non-overlapping set of people fare in another outcome. I think that the details of intersectional worseness were fairly straightforward and intuitive, in the context of trying to spell out utilitarianism. It is much less obvious, I think, when we should say that one group of people fares better or worse than an entirely distinct group of people (which may be of a different size). The conclusion I arrive at is a little complicated. I shall therefore build up to my definition of complementary worseness by considering some relatively simple cases. I’ll consider four kinds of comparison: first, between a pair of outcomes such that all members of both of the relevant complementary sets are happy, and those sets are of the same size; second, between a pair of outcomes such that all members of both of the relevant complementary sets are happy, and those sets are of different sizes; third, between a pair of outcomes such that all members of both of the relevant complementary sets are wretched (regardless of the size of those sets); and finally, between a pair of outcomes such that some of the members of the relevant complementary sets are happy, and some wretched. Then I’ll give the general formula for calculating complementary worseness.

For any pair of outcomes, $A$ and $B$, call the people who exist in $A$ but not in $B$ the $A$-not-$B$ people; and the set of those people (the complement of the population of $B$ in the population of $A$) the $A$-not-$B$ set. Now, imagine two outcomes, $A$ and $B$, such that all of the $A$-not-$B$ people are happy (have

\textsuperscript{12}A potential concern about terminology here. I talk of complementary worseness in terms, for example, of $A$ being worse for a group of people who exist in $A$, than $B$ is for a group of different people who exist in $B$. I go on to say that $A$ is complementarily worse than $B$ if the population of $A$ consists of just one wretched person, and the population of $B$ is the empty population. I seem therefor to be saying that $A$ might be worse for some wretched person who exists there, than $B$ is for... no-one. However, the idea of how things are for no-one is an odd one, to say the least. It reminds me of the claim, about which I expressed doubt in Section 3.1, that existence can be either better or worse for a person than non-existence. I think that my talk of one outcome being better than another for a group of people should be taken less literally than this. It’s just a way of talking about comparing some of the wellbeings in one outcome with what happens instead in another.
positive wellbeings) in $A$, and all of the $B$-not-$A$ people are happy in $B$. Imagine further that the number of $A$-not-$B$ people is the same as the number of $B$-not-$A$ people. My suggestion is that we should make this kind of interpersonal group-wise comparison in much the same way that we made group-wise comparisons concerning just one set of people above. In such a case: if the average wellbeing of the $A$-not-$B$ people in $A$ is less than the average wellbeing of the $B$-not-$A$ people in $B$, then $A$ is complementarily worse than $B$ by the difference between those averages, multiplied by the number of $A$-not-$B$ people (which is also the number of $B$-not-$A$ people). If the average wellbeing of the $A$-not-$B$ people in $A$ is not less than the average wellbeing of the $B$-not-$A$ people in $B$, then $A$ is complementarily worse than $B$ by zero.

This proposal is based in part on my intuitions about non-identity decisions (see Section 3.3), and in part on my assumption of the inter-personal comparability of wellbeings (see Section 1.2). I assumed that the life lived by one person with a lower wellbeing is worse for her, than the life lived by a different person with a higher wellbeing is for him. This is related to my intuition about non-identity decisions: roughly, that it is worse in some morally relevant way that one person exists with lower wellbeing, than that a different person exists with a higher wellbeing instead. That the former person’s life is worse for her than the latter person’s life is for him is surely part of the explanation for this morally relevant worseness. My claim here is that we should use averages to generalise this kind of inter-personal comparison to group-wise comparisons (as we did for intra-personal group-wise comparisons).

Let me illustrate using these three outcomes.

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
<th>$p_5$</th>
<th>$p_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>6</td>
<td>9</td>
<td>9</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>10</td>
<td>8</td>
<td>9</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$C$</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>10</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

$A$ is intersectionally worse than $B$ because in $A$ three people exist with particular happy wellbeings, and in $B$ those people exist with (on average) happier wellbeings. Those people fare worse, as a group, in $A$ than in $B$. My suggestion is that $A$ is complementarily worse than $C$ for a very similar reason: in $A$ three people exist with particular happy wellbeings, and in $C$ three different people exist instead, with (on average) happier wellbeings. The three people who exist in $A$ fare worse in $A$ as a group, than the three people who exist in $B$ fare
in B. A is complementarily worse than C by the same amount by which it is intersectionally worse than B, I think.

Of course, as I’ve pointed out, there are other ways of describing intersectional worseness; other possible explanations of why A is worse than B, and other assessments of the degree of worseness. My decision to present intersectional worseness in terms of average wellbeings was partly motivated by the way that it fits with my assessment of complementary comparisons, so I don’t mean to lean on that presentation to support this assessment. I’ll now turn to a complementary comparison of a slightly different kind, which I think does provide some support for the averagist aspect of my proposal.

Imagine two outcomes, A and B, such that all of the A-not-B people are happy in A and all of the B-not-A people are happy in B, but the number of A-not-B people is different from the number of B-not-A people. Here I think we should begin, again, by calculating the average wellbeing of the A-not-B people in A, and the average wellbeing of the B-not-A people in B. If the former is not less than the latter, A is complementarily worse than B by zero. If the former is less than the latter, A is complementarily worse than B by the difference between the two averages multiplied by the number of people in the smaller of the A-not-B and B-not-A sets. Why the number of people in the smaller of the two sets? Because this is the number of people who exist at the lower wellbeing (on average) instead of someone else existing at a higher wellbeing (on average) instead.

Here are some outcomes to illustrate.

<table>
<thead>
<tr>
<th></th>
<th>p₁</th>
<th>p₂</th>
<th>p₃</th>
<th>p₄</th>
<th>p₅</th>
<th>p₆</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>B</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

How should we make the complementary comparison between A and B? How much worse do the A-not-B people fare in A than the B-not-A people fare in B? The average of the A-not-B people in A is 1. The average of the B-not-A people in B is 10. So they fare worse by 9, on average. There are four A-not-B people at this lower (average) wellbeing, so it is tempting to think that we should multiply 9 by 4 to find the amount by which A is complementarily worse than B. My claim is that we should rather multiply 9 by 2, which is the number of people in the smaller of the A-not-B and B-not-A sets (in this case, the number of B-not-A people): A is complementarily worse than B by 9 × 2 = 18. For we
want to capture the idea that it is worse that someone exists at a lower wellbeing than that someone else exists at a higher wellbeing instead, and only two of the $A$-not-$B$ people are ‘replaced’ in $B$ (though not any specific two). Of the third and fourth $A$-not-$B$ people it is not the case that someone else exists happier in $B$ in their stead; rather, no-one exists in their stead in $B$ (which is something we are supposed to be neutral about). If the number were reversed and there were more $B$-not-$A$ people than $A$-not-$B$ people something similar would be true: there would be extra $B$-not-$A$ people not replacing anyone in $A$.

I think this kind of case shows that complementary comparisons should not be based on total wellbeings, or on some interpersonal analogue to harms. Consider the totalist option first. If we were to say that the $A$-not-$B$ people fare worse in $A$ than the $B$-not-$A$ people fare in $B$ if the total of the $A$-not-$B$ people’s wellbeings in $A$ is less than the total of the $B$-not-$A$ people’s wellbeing in $B$, this would conflict fairly directly with the neutrality intuition. This approach would sometimes lead to the conclusion that the $A$-not-$B$ people fare worse in $A$ than the $B$-not-$A$ people fare in $B$, even though the wellbeing of every $A$-not-$B$ person is higher than the wellbeing of any $B$-not-$A$ person. This approach favours larger (happy) populations, and seems likely to leads us towards repugnance, at least in non-identity decisions (thought at this point I am only considering when one outcome is complementarily worse than another, and not how such an axiology might affect choice-worthiness).

Consider the option of appealing to some interpersonal analogue to harms. We might want, in this kind of case, to ‘match up’ one $A$-not-$B$ person with one $B$-not-$A$ person, and compare their wellbeings directly. We could then say that some $A$-not-$B$ person incurs an interpersonal harm in $A$ if her wellbeing in $A$ is lower than her counterpart’s wellbeing in $B$ (that is, if $A$ and $B$ are the available outcomes in a decision). And then we could proceed to make complementary comparisons on the basis of those interpersonal harms. In a case in which the number of $A$-not-$B$ people and the number of $B$-not-$A$ people is the same this approach is plausible, because it doesn’t matter how you match people up (the balance of harms will be the same). Considering a case in which the number of $A$-not-$B$ people and the number of $B$-not-$A$ people differ, I think it is evident that there is no good way to match people up. Which people should be matched with which, and which people in the larger group should be left out? These questions matter, and I can’t think of a good reason for any particular answer.
So much for complementary comparisons when everyone is happy. How should we make such comparisons when some of the people concerned are wretched? First, what if all of the people concerned are wretched? Imagine two outcomes, A and B, such that all of the A-not-B people are wretched in A and all of the B-not-A people are wretched in B. I think that this kind of comparison should proceed a little differently, to reflect my asymmetric intuitions about happiness and wretchedness. I am neutral about a happy person existing instead of not existing, but I am not neutral about a wretched person existing instead of not existing. It is worse, in some morally relevant way, that a person exists wretched than that she doesn’t exist. This difference warrants a shift from averages to sums. If all of the A-not-B people are wretched in A and all of the B-not-A people are wretched in B, and the sum of the wellbeings of the A-not-B people in A is less than the sum of the wellbeings of the B-not-A people in B, A is complementarily worse than B by the difference between those two sums. If the sum of the wellbeings of the A-not-B people in A is not less than the sum of the wellbeings of the B-not-A people in B, A is complementarily worse than B by zero.

Here are some outcomes to illustrate.

<table>
<thead>
<tr>
<th></th>
<th>p₁</th>
<th>p₂</th>
<th>p₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>−8</td>
<td>−6</td>
<td>*</td>
</tr>
<tr>
<td>B</td>
<td>*</td>
<td>*</td>
<td>−9</td>
</tr>
</tbody>
</table>

The A-not-B people are p₁ and p₂, and the one B-not-A person is p₃. The sum of the wellbeings of the A-not-B people in A is −14 and the sum of the wellbeings of the B-not-A people in B is −9. So, A is complementarily worse than B by the difference between the two, which is −9 − −14 = 5. B is complementarily worse than A by zero. This reflects how much more wretchedness there is in the A-not-B set in A than in the B-not-A set in B, and vice versa.

Now I’ve spelled out how I think complementary comparisons should proceed if all of the relevant people are happy, and if all of the relevant people are wretched. The tricky thing is putting these two elements together. How should a complementary comparison between A and B proceed if some of the A-not-B people are wretched and some happy in A, and some of the B-not-A people are wretched and some happy in B? Let me try to explain how I think we should make such comparisons. First, we should make a wretchedness-based comparison. How much more wretchedness is there among the A-not-B people
in A than among the B-not-A people in B? We do this by summing the wretched wellbeings in each set, and finding the difference: call this the wretchedness element of complementary worseness. Having dealt with these wretched wellbeings, we should replace each one by wellbeing zero. Using these ‘adjusted wellbeings’ we should make a happiness-based comparison. To what extent are people in one outcome replaced by happier people in the other outcome? We do this by calculating the average of the adjusted wellbeings in each set, finding the difference, and multiplying the difference by the number of people in the smaller of the two sets. Call this the happiness element of complementary worseness. We should sum the wretchedness element and the happiness element for each outcome. If the sum of these elements for A is greater than the sum of these elements for B, then A is complementarily worse than B by the difference between the sums. If not, A is complementarily worse than B by zero. This describes the general approach to complementary comparisons according to SU.

Take any two outcomes, A and B. Here is how to calculate by how much A is complementarily worse than B:

**Complementary worseness:** If the sum of the wretched wellbeings of the A-not-B people in A is lower than the sum of the wretched wellbeings of the B-not-A people in B, A is complementarily worse than B due to wretchedness by the difference between those sums. If the sum of the wretched wellbeings of the A-not-B people in A is not lower than the sum of the wretched wellbeings of the B-not-A people in B, A is complementarily worse than B due to wretchedness by zero. Replace all wretched wellbeings in A and B with wellbeing 0. Now, if the average adjusted-wellbeing of the A-not-B people in A is less than the average adjusted-wellbeing of the B-not-A people in B, then A is complementarily worse than B due to happiness by the difference between those averages, multiplied by whichever is smallest of the number of A-not-B people and the number of B-not-A people. If the average adjusted-wellbeing of the A-not-B people in A is not less than the average adjusted-wellbeing of the B-not-A people in B, then A is complementarily worse than B due to happiness by zero. Take the sum of the amounts by which A is complementarily worse than B due to wretchedness and due to
happiness, and the sum of the amounts by which \( B \) is worse than \( A \) due to wretchedness and due to happiness. If the former is greater than the latter, \( A \) is complementarily worse than \( B \) by the difference between the two sums. If the former is not greater than the latter, \( A \) is complementarily worse than \( B \) by zero.\(^{13}\)

This is all much less complicated than it sounds, I think. Here are some outcomes to illustrate complementary worseness:

\[
\begin{array}{c|cccc}
A & p_1 & p_2 & p_3 & p_4 & p_5 \\
\hline
B & 2 & 0 & * & + & * \\
& * & * & 9 & -2 & -1
\end{array}
\]

The \( A \)-not-\( B \) people are \( p_1 \) and \( p_2 \), and the \( B \)-not-\( A \) people are \( p_3, p_4, \) and \( p_5 \). First the wretchedness element. \( B \) is complementarily worse than \( A \) due to wretchedness by 3 (the difference between the sum of wretched wellbeings in the \( B \)-not-\( A \) set in \( B \) and the sum of wretched wellbeings in the \( A \)-not-\( B \) set in \( A \)). \( A \) is complementarily worse than \( B \) due to wretchedness by 0. We now replace those wretched wellbeings with zeros, and calculate averages. The average adjusted wellbeing of the \( A \)-not-\( B \) people in \( A \) is 1. The average adjusted wellbeing of the \( B \)-not-\( A \) people in \( B \) is 3. The difference between these is 2, and there are two people in the smaller of the two sets, so \( A \) is complementarily worse than \( B \) due to happiness by 2 \( \times \) 2 = 4. \( B \) is complementarily worse than \( A \) due to happiness by 0. We sum the worseness due to wretchedness and happiness for each outcome and find the difference to conclude: \( A \) is complementarily worse than \( B \) by 1 and \( B \) is complementarily worse than \( A \) by 0.\(^{14}\)

\(^{13}\) If there are no \( A \)-not-\( B \) people, then the sum of the wellbeings of the \( A \)-not-\( B \) people in \( A \) is 0 and (as above) the average adjusted wellbeing of the \( A \)-not-\( B \) people in \( A \) is undefined (likewise for the \( B \)-not-\( A \) people). This reflects mathematical convention: the sum of the members of an empty set being 0, and its average (or anything divided by 0) being undefined. More importantly, though, it also reflects both the view that it is worse that people exist wretched than that no-one exists at all, and the view that it is not worse that no-one exists at all than that people exist happy.

\(^{14}\) Note that the formula for calculating complementary worseness is equivalent to the formula for calculating intersectional worseness, when applied to non-empty intersectional sets. So I could have presented intersectional worseness, or indeed CFPU, by this formula. Take two sets of integers of the same size. Calculate the difference between the sums of the negative integers in each set. Replace those negative integers with zeros and calculate the difference between the sums of the members of each revised set (which is equivalent to the difference between their average multiplied by the number of members of each set, and also equivalent to the difference between the sums of the integers greater than or equal to zero in the original sets). Sum the two differences. This is equivalent to simply finding the difference between the sums of the entire original sets, or to finding the difference between the averages of the original sets multiplied by the number of members in each set.
Another note on what is going on here axiologically. The claim that A is complementarily worse than B by some amount is an axiological claim. It asserts that there is some morally relevant betterness relation according to which A is strictly worse than B by that amount. However, there is no single relation of complementary betterness. Rather, there are a series of betterness relations, each being the relation of betterness with respect to a pair of sets of people. ‘A is complementarily worse than B’ is shorthand for ‘A is worse with respect to the pair of sets, (A-not-B set, B-not-A set)’.

4.1.4 Total worseness

These are the two significant kinds of worseness according to SU: intersectional worseness, and complementary worseness. To assist me in explaining how worseness of these two kinds determine choice-worthiness in a decision, let me introduce another piece of terminology. For any pair of outcomes, A and B, we can now say by how much A is intersectionally worse than B, and by how much A is complementarily worse than B. Call the sum of these amounts the ‘total worseness’ of A than B. The total worseness of any outcome than any other outcome is greater than or equal to zero. Here is the definition of total worseness for any pair of outcomes, A and B.

Total worseness: The total worseness of outcome A than outcome B is the sum of the amount by which A is intersectionally worse than B and the amount by which A is complementarily worse than B.

Total worseness is not a third kind of worseness. It is not overall (or all-things-considered) worseness. It is simply a handy piece of terminology for referring to the sum of the amounts by which one outcome is intersectionally and complementarily worse than a second outcome. It enables my discussion of what we do with these kinds of worseness to be a little bit simpler and more concise. Instead of having constantly to speak of the sum of those two amounts I can simply speak of total worseness. In summing intersectional and complementary worseness we are not balancing one against the other. We are summarising how much things tell against choosing one particular outcome rather than another. If you need convincing that total worseness is not overall worseness, note that the total worseness of A than B and the total worseness of B than A might both be positive. This will be the case if A is intersection-
ally worse than $B$ by some positive amount, and $B$ is complementarily worse than $A$ by some positive amount (or vice versa). It would be a very strange overall betterness relation indeed, that accommodated this kind of reflexivity of worseness.\footnote{There is a presentational or terminological element here. Some people are happy to wreak havoc with the structure of betterness relations. I hold fixed to the standard structure, set out in Section 2.2.2. There are other presentational options. I discuss some of them in Section 5.2.}

Here are some outcomes to illustrate total worseness.

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>4</td>
<td>1</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>6</td>
<td>*</td>
<td>-2</td>
<td>-2</td>
</tr>
</tbody>
</table>

Here, $A$ is intersectionally worse than $B$ by 2, and $B$ is intersectionally worse than $A$ by 0. $A$ is complementarily worse than $B$ by 0, and $B$ is complementarily worse than $A$ by 5. Thus, the total worseness of $A$ than $B$ is $2 + 0 = 2$ and the total worseness of $B$ than $A$ is $0 + 5 = 5$.

### 4.1.5 Shortfalls and choice-worthiness

So far (apart from my detour through IU) I have just been talking about comparisons between pairs of outcomes: picking out the kinds of worseness between one outcome and another that are morally significant according to SU. Now I shall return to the question of how these kinds of worseness determine choice-worthiness, according to SU. This part shouldn’t come as much of a surprise by now. It is very similar to the way that intersectional worseness alone determines choice-worthiness according to IU.

Imagine some decision, $D_1$. For each pair of available outcomes in $D_1$, $A$ and $B$, there is a total worseness of $A$ than $B$. This is a summary of how much things tell against choosing $A$ rather than $B$. Call the greatest total worseness of $A$ than $B$, where $A$ and $B$ are both available outcomes in $D_1$, the *shortfall* of $A$ in $D_1$.

*Shortfall:* The shortfall of available outcome $A$ in decision $D_1$ is the greatest total worseness of $A$ than $B$, where $B$ is an available outcome in $D_1$.

Here is a decision to illustrate shortfalls. Beside the decision table, the total worsenesses relevant to the decision are represented in a separate table. Read
the number in a field as the total worseness of the outcome to the left than the outcome above (i.e. the first number in the second row is the total worseness of \( B \) than \( A \)). The shortfalls are marked in **bold**.

\[
\begin{array}{cccc|ccc}
 & p_1 & p_2 & p_3 & p_4 & A & B & C \\
\hline
D_{34}:
A & 4 & 2 & * & * & A & 0 & 2 & 0 \\
B & 6 & * & -2 & -1 & B & 5 & 0 & 3 \\
C & * & 1 & * & * & C & 1 & 0 & 0 \\
\end{array}
\]

The shortfall of \( A \) in \( D_{34} \) is 2; the shortfall of \( B \) in \( D_{34} \) is 5; the shortfall of \( C \) in \( D_{34} \) is 1.

Finally, shortfalls determine choice-worthiness according to SU. Faced with any decision, one ought to choose the available outcome with the lowest shortfall in that decision. This is the outcome such that things tell against choosing it rather than some (single) available alternative by the least amount. More generally, the lower the shortfall, the more choice-worthy the outcome.

**Shortfall Utilitarianism:** For any decision, \( D_1 \), one available outcome, \( A \), is at least as choice-worthy in \( D_1 \) as a second available outcome, \( B \), just in case the shortfall of \( A \) in \( D_1 \) is no greater than the shortfall of \( B \) in \( D_1 \).

In \( D_{34} \), then, one ought to bring about \( C \), which has a lower shortfall and is thus more choice-worthy in \( D_{34} \) than \( A \), both of which are more choice-worthy in \( D_{34} \) than \( B \).

Here is one more decision to illustrate the theory in action.

\[
\begin{array}{cccc|ccc}
 & p_1 & p_2 & p_3 & p_4 & A & B & C \\
\hline
D_{35}:
A & 8 & 2 & * & * & A & 0 & 3 & 2 \\
B & 6 & * & 6 & 4 & B & 2 & 0 & 0 \\
C & * & 4 & 4 & -6 & C & 12 & 14 & 0 \\
\end{array}
\]

I’ll go through the calculations of each positive amount of intersectional or complementary worseness of one of the available outcomes than another (the others are all 0). \( A \) is complementarily worse than \( B \) by 3: one person exists at wellbeing 2 in \( A \) instead of someone else existing at (on average) wellbeing 5 in \( B \). \( A \) is intersectionally worse than \( C \) by 2: \( p_2 \) exists at wellbeing 2 in \( A \) instead of at wellbeing 4 in \( C \). \( B \) is intersectionally worse than \( A \) by 2: \( p_1 \) exists at wellbeing 6 in \( B \) instead of at wellbeing 8 in \( A \). \( C \) is complementarily worse
than \( A \) by 12: someone exists at wellbeing \(-6\) in \( C \) instead of not existing at all in \( A \); and, further, someone exists with an average (adjusted) wellbeing of \( 2 \) in \( C \) instead of someone else existing with a wellbeing of \( 8 \) in \( A \). \( C \) is intersectionally worse than \( B \) by 12: \( p_3 \) and \( p_4 \) exist at (on average) wellbeing \(-1\) in \( C \) instead of at (on average) wellbeing \( 5 \) in \( B \). \( C \) is also complementarily worse than \( B \) by 2: someone exists at wellbeing 4 in \( C \) instead of someone else existing at wellbeing 6 in \( B \). So, the shortfalls of \( A, B, \) and \( C \) in \( D_{35} \) are \( 3, 2, \) and \( 14 \), respectively. \( B \) is the available outcome most choice-worthy in \( D_{35} \), followed by \( A \) and then \( C \).

That is Shortfall Utilitarianism. I think that it is a better version of utilitarianism than any I have encountered so far. In the next section I show that SU is indeed a general utilitarian theory, as defined in Chapter 2. In the following section I show that it agrees with the four intuitions about variable population decisions that I discussed in Chapter 3.

### 4.2 SU is a general utilitarian theory

SU is a general utilitarian theory: it is a consequentialist, axiological, welfarist extension of CFPU. Apart from reassuring the reader that this is the case, I hope that this section will shed some further light on the axiological and normative structure of SU.

#### 4.2.1 SU extends CFPU

CFPU is the restricted moral theory according to which in any fixed population decision a first available outcome is more choice-worthy than a second just in case the total wellbeing of the first is higher than the total wellbeing of the second (see page 20). An extension of CFPU is a theory that agrees with the choice-worthiness orderings of CFPU in all fixed population decisions, but also provides choice-worthiness orderings for all variable population decisions (and thus for all decisions). SU is such a theory. It is self-evident, I hope, that SU is fully general: it provides choice-worthiness orderings for all decisions, whether the population is fixed or variable. (A reminder that an outcome with the empty population always has shortfall zero, and is thus in the choice-worthy ordering for any decision in which it is available.) It remains to be shown that SU agrees with CFPU in fixed population decisions.
The available outcomes in any fixed population decision will be ordered for choice-worthiness in that decision by SU, as by CFPU, according to the total wellbeing of their population—the higher the total wellbeing, the more choice-worthy the outcome. Here is how we can know this. Consider any fixed population decision, $D_1$. We can begin by noting that for any pair of available outcomes in $D_1$, $A$ and $B$, $A$ is complementarily worse than $B$ by zero. In a fixed population decision exactly the same people exist in each available outcome; and one outcome is complementarily worse than another by zero unless there are some people who exist in that outcome but not in the other. This means that in fixed population decisions intersectional worseness alone determines the total worseness of each available outcome than each other available outcome, and thus the shortfall and choice-worthiness of each available outcome in the decision.

Now, among the available outcomes in $D_1$, an available outcome with the highest total wellbeing is also an outcome with the highest average wellbeing, and thus intersectionally worse than each available outcome by zero. Such an outcome has a total worseness than each other available outcome in $D_1$ of zero, and consequently a shortfall in $D_1$ of zero. This is the lowest shortfall an available outcome can have in a decision. (Such an outcome is therefore permissible in the decision.) Say such an outcome in $D_1$ is $A$. If there is any available outcome $B$ in $D_1$ with a lower total wellbeing than $A$, $B$ is intersectionally worse than $A$ by the difference between the total wellbeings of $A$ and $B$ (which is of course equivalent to the difference between the average wellbeings in $A$ and in $B$ multiplied by the number of people in the population). The amount by which $B$ is intersectionally worse than $A$ is also the shortfall of $B$ in $D_1$.

To summarise: the shortfall of any available outcome in a fixed population decision is zero if that outcome has the highest total wellbeing of any available outcome in the decision; otherwise, it is the amount by which the total wellbeing of the outcome falls short of the highest total wellbeing of any available outcome in the decision. SU, like CFPU and its other extensions, orders the available outcomes in fixed populations decisions for choice-worthiness according to the total wellbeing of their (fixed) population.

---

16There is a minor disconnect between average wellbeings and total wellbeings, in the case of fixed population decisions where the fixed population is the empty population. An empty population has a total wellbeing (0), but its average wellbeing is undefined. The available outcomes in such a decision all have shortfalls of 0, according to SU, and are thus equally choice-worthy. So SU agrees with CFPU regarding these decisions also.
4.2.2 SU is consequentialist

SU is a consequentialist theory, according to my definition of consequentialism on page 24. Here is that definition again.

*Consequentialism:* The choice-worthiness relation in any decision, $D_1$, is fully determined by the available outcomes in $D_1$.

I proposed that a consequentialist theory can be represented as making decisions between available outcomes, rather than between available acts each of which will lead to an outcome; and that a non-consequentialist theory cannot be represented in this way. Clearly, SU can be represented in this way. I have been able to present the theory without mentioning acts at all. Having set this kind of consequentialism as a requirement, I was able to drop talk of decisions between available acts that lead to particular outcomes, in favour of the simpler talk of decisions between available outcomes. The only input given to SU, the only information required to apply it, is which outcomes are available in the decision at hand. It does not matter, for example, what kind of act it is that leads, if performed, to any of those outcomes. It also does not matter who is making the decision in question—SU is an agent neutral theory.

4.2.3 SU is axiological

SU is an axiological consequentialist theory, according to my definition of axiological consequentialism on page 28. Here is that definition again.

*Axiological consequentialism:* The choice-worthiness relation in any decision, $D_1$, is fully determined by the morally relevant betterness relations among the available outcomes in $D_1$.

I proposed that an axiological consequentialist theory consists of an axiology (an account of all of the morally relevant betterness relations among outcomes) and a normative component (a procedure or function that determines choice-worthiness from the axiology). My presentation of SU was to some extent focussed on the normative component of the theory. I briefly discussed what the complete axiology of SU must look like, but defined intersectional and complementary worseness without fully specifying the various respects of betterness underpinning them. I haven’t yet given a full account of all of the
morally relevant betterness relations among outcomes, according to SU. I’ll do that now.\footnote{That I was able to present SU without appealing to its axiology should not cast doubt on the axiological nature of the theory. Remember, I was able to present CFPU without appealing to its axiology too (see page 20).}

There are many morally relevant betterness relations according to SU (and underpinning intersectional and complementary worseness). Each is the relation of betterness with respect to a particular pair of sets of possible people. There is a relation of betterness with respect to each pair of identical sets of possible people, and a relation of betterness with respect to each pair of non-overlapping (entirely distinct) sets of possible people.\footnote{One might in one sense simplify the axiology by including a relation of betterness with respect to each partially-overlapping set of possible people. This would mean that there was a relation of betterness with respect to every pair of sets of possible people. I don’t include these relations because they don’t play any role in SU: SU is identified by its choice-worthiness output, and adding superfluous betterness relations to the axiology wouldn’t (in this case) change that output. I think it more informative to stick to representing only the necessary or efficacious (‘morally relevant’) axiology of SU.} The former underpin intersectional worseness, and the latter underpin complementary worseness. Each relation is standardly structured (as I assumed in Section 2.2.2).

Label all subsets of possible people, \(S_1, S_2, S_3, \ldots\). Then, label the relation of being at least as good with respect to a pair of sets of people, \((S_1, S_2), R_{(S_1,S_2)}\). We can represent these relations with the symbol \(\geq_{(S_1,S_2)}\).\footnote{And, as usual, the other related symbols too: for example, \(\succ_{(S_1,S_1)}, \prec_{(S_1,S_2)}, \simeq_{(S_1,S_1)},\) and \(\preceq_{(S_1,S_2)}\).} These relations reflect the notion that I have been appealing to of a group of people in one outcome being better or worse off than a group of people (either all the same, or all different people) in another outcome. We can read ‘\(A \geq_{(S_1,S_2)} B\)’ as ‘\(A\) is at least as good with respect to the pair of sets of people, \((S_1, S_2)\), as \(B\)’; or as ‘the members of \(S_1\) fare at least as well as a group in \(A\) as the members of \(S_2\) do in \(B\)’.

I can fully specify the axiology of SU by giving the necessary and sufficient conditions for each of these relations to hold between a pair of outcomes, as well as the magnitudes of the relation in each case (\textit{how much} worse one outcome is than another in a respect). There is not quite a relation for each possible pair or sets of people. There is a relation for each pair of identical sets of possible people, and a relation for each pair of entirely distinct (non-overlapping) sets of possible people. Let me take these two classes of relation in turn, and describe the relevant relations of betterness in those respects.

First, there is a relation of betterness with respect to each pair of sets of
people, \((S_1, S_2)\), such that \(S_1 = S_2\). The first necessary condition for some outcome, \(A\), to be at least as good as another outcome, \(B\), with respect to such a pair of identical sets of people is that the members of the relevant set exist in \(A\) and in \(B\). That is, \((S_1 = S_2) \in \mathcal{P}_A\) and \((S_1 = S_2) \in \mathcal{P}_B\). These respects of betterness can be further specified by reference to the definition of intersectional worseness (on pages 94–95). Take two outcomes, \(A\) and \(B\), and a pair of sets of people, \((S_1, S_2)\), such that \(S_1 = S_2\), and \((S_1 = S_2) \in \mathcal{P}_A\) and \((S_1 = S_2) \in \mathcal{P}_B\). Then, \(A\) is at least as good as \(B\) with respect to \((S_1, S_2)\) iff the average wellbeing in the set \(S_1 = S_2\) in \(A\) is at least as high as the average wellbeing in the set \(S_1 = S_2\) in \(B\).^{20} The magnitude of betterness between \(A\) and \(B\) with respect to \((S_1, S_2)\) is the difference between the average wellbeing in that set in \(A\) and the average wellbeing in that set in \(B\), multiplied by the number of members in the set (equivalently, the difference in total wellbeings in the set in each outcome).

A relation of this first kind, \(R_{(S_1, S_2)}\) where \(S_1 = S_2\), orders all outcomes in which all of the members of \(S_1 = S_2\) exist according to the average (equivalently, total) wellbeing among the members of that set in those outcomes. Other outcomes (i.e. in which not all members of \(S_1\) exist) do not ‘participate’ in this relation: they are not related in this respect to any other outcome. Each such relation is therefore very incomplete.

Second, there is a relation of betterness with respect to each pair of sets of people, \((S_1, S_2)\), such that \(S_1 \neq S_2\) and there is no possible person, \(p\), such that \(p \in S_1\) and \(p \in S_2\). The first necessary condition for some outcome, \(A\), to be at least as good as another outcome, \(B\), with respect to such a pair of non-overlapping sets of people is that the members of one of the relevant sets exist in \(A\) and the members of the other set exist in \(B\). That is, \(S_1 \in \mathcal{P}_A\) and \(S_2 \in \mathcal{P}_B\). These respects of betterness can be further specified by reference to the definition of complementary worseness (on page 105). Take two outcomes, \(A\) and \(B\), and a pair of sets of people, \((S_1, S_2)\), such that \(S_1 \neq S_2\); and there is no possible person, \(p\), such that \(p \in S_1\) and \(p \in S_2\); and \(S_1 \in \mathcal{P}_A\) and \(S_2 \in \mathcal{P}_B\). Then, \(A\) is at least as good as \(B\) with respect to \((S_1, S_2)\) iff the sum of the following is greater than or equal to zero: a) the difference between the amount of wretchedness among the members of \(S_1\) in \(A\) and the amount of wretchedness among the members of \(S_2\) in \(B\); and b) the difference between the average adjusted wellbeing among the members of \(S_1\) in \(A\) and the average adjusted wellbeing among the members of \(S_2\) in \(B\).^{20}

---

^{20}Strictly, also \(A\) is always at least as good as \(B\) with respect to \((\emptyset, \emptyset)\).
4.2. SU IS A GENERAL UTILITARIAN THEORY

of $S_2$ in $B$, multiplied by the number of members in the smaller of $S_1$ and $S_2$. The magnitude of betterness between $A$ and $B$ with respect to $(S_1, S_2)$ is also determined by this amount.

The full axiology of SU consists of all of these relations of betterness with respect to pairs of sets of people. These are all of the morally relevant betterness relations. They determine choice-worthiness by determining intersectional and complementary worseness. Intersectional and complementary worseness are always a measure of strict worseness according to one or another of these relations of betterness. Intersectional worseness is a measure of the strict worseness of one outcome than another with respect to the pair of identical sets of people, each the intersection of the populations of those two outcomes. Complementary worseness is a measure of the strict worseness of one outcome than another with respect to the pair of sets of outcomes, the complements of the population of each outcome in the population of the other. According to SU, one ought to minimise the strict worseness of a chosen outcome than another available outcome, in the two respects thus picked out as relevant to the comparison between the two.

This means that not all of the morally relevant betterness relations are directly morally relevant in every decision. Sometimes, not even all of the relations that hold between available outcomes in a decision are relevant to that decision. The relations that are morally relevant in a decision are those relations picked out by the definitions of intersectional and complementary worseness in the previous section. The intersectional worseness of $A$ than $B$, for example, is a matter of the relation of betterness with respect to the pair of identical sets of people: (the $A$-and-$B$ set, the $A$-and-$B$ set). It may be that a relation of betterness with respect to a different pair of identical sets of people (being some subset of the $A$-and-$B$ set) also holds between $A$ and $B$. The moral relevance of this relation in this comparison is in a sense ‘subsumed’ by the former relation. Similarly, the complementary worseness of $A$ than $B$ is a matter of the relation of betterness with respect to the pair of non-overlapping sets of people: (the $A$-not-$B$ set, the $B$-not-$A$ set). If relations of betterness with respect to other pairs of non-overlapping outcomes hold between $A$ and $B$ this is not directly morally relevant; those other relations are subsumed in the relevant relation.

This concludes my account of the axiology of SU. At this point I’d like

\[21\] Treating the empty population as for complementary worseness calculations.
to defuse a possible concern about the transitivity of the worseness relations appealed to by SU. Each of the worseness relations posited in this axiology is transitive. One might have wondered whether this is true. It seems as though the following might be true: A is intersectionally worse than B, B is intersectionally worse than C, and C is intersectionally worse than A. One might say that about these three outcomes, for example.

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8</td>
<td>*</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>8</td>
<td>*</td>
</tr>
<tr>
<td>C</td>
<td>*</td>
<td>10</td>
<td>8</td>
</tr>
</tbody>
</table>

Don’t be alarmed. Contrary to appearances, there is no intransitivity in any of the worseness relations involved. What is going on here is that three worseness relations are being run together: three different relations are appealed to as ‘intersectional worseness’. Expressed more fully, the true claim made above would be: A is worse with respect to the pair of sets ($\{p_1\}, \{p_1\}$) than B, B is worse with respect to the pair of sets ($\{p_2\}, \{p_2\}$) than C, and C is worse with respect to the pair of sets ($\{p_3\}, \{p_3\}$) than A. Worseness with respect to a particular pair of sets of people is transitive. ‘Intersectional worseness’ is not worseness in a particular respect, and should not (indeed, cannot) be represented by a single betterness relation.

It is possible, though probably less tempting, to make a similar mistake regarding a more familiar kind of worseness: worseness for a particular person (or harm). It is true of the outcomes above, though ambiguously expressed, that A is worse for a particular person than B, B is worse for a particular person than C, and C is worse for a particular person than A. Nonetheless, we should not be tempted to think that any of the relations of betterness with respect to a particular person is intransitive. The appearance of intransitivity is due to the suppression of which particular person is relevant in each comparison. The mistake in this context is fairly easy to spot, and probably rarely made. It therefore serves as a useful analogy to the possible mistake in the context of the axiology of SU.
4.2.4 SU and the overall betterness requirement

SU seems to me to be a consequentialist theory and an axiological theory. I don’t just mean that it fits my definition of an axiological consequentialist theory (though of course I’ve just shown that it does so). I mean that I am inclined to apply these terms to SU based on my intuitive, pre-definition grasp of what they mean. This is in fact part of the reason that I gave the definitions that I gave. SU makes decisions on the basis of the outcomes of the available acts and nothing else, which seems to me to make it a consequentialist theory. SU makes decisions more specifically on the basis of the relations of betterness among the available acts and nothing else, which seems to me to make it an axiological consequentialist theory. However, SU does not meet the overall betterness requirement. It does not produce a single morally relevant betterness relation among outcomes. So although it might have been tempting to define consequentialism or axiology in terms of an overall betterness relation, in light of the existence of theories like SU (and HU, NHU, and IU) I adopted more inclusive definitions.

That SU does not meet the overall betterness requirement is evident in the fact that choice-worthiness is menu-dependent according to SU, in interesting ways. So, for example, in the following decision, creating a moderately happy person (choosing outcome B) and not creating that person (choosing outcome A) are equally choice-worthy.

\[
\begin{array}{ccc}
D_{36}: & p_1 & p_2 & p_3 \\
A & 8 & 8 & *
\end{array}
\]

\[
\begin{array}{ccc}
B & 8 & 8 & 8
\end{array}
\]

Whereas in \(D_{37}\) creating a moderately happy person (choosing the same outcome B) is less choice-worthy than not creating that person (choosing the same outcome A), because the option of creating a very happy person instead (choosing outcome C) is also available.

\[
\begin{array}{cccc}
D_{37}: & p_1 & p_2 & p_3 & p_4 \\
A & 8 & 8 & * & *
\end{array}
\]

\[
\begin{array}{cccc}
B & 8 & 8 & 8 & *
\end{array}
\]

\[
\begin{array}{cccc}
C & 8 & 8 & * & 10
\end{array}
\]

We can’t account for this kind of shift in choice-worthiness by appeal to a single overall betterness relation. To account for their equal choice-worthiness in \(D_{35}\)
we would have to say that \( A \) and \( B \) are equally good overall, and to account for their unequal choice-worthiness in \( D_37 \) we would have to say that \( B \) is less good overall than \( A \). These two claims are incompatible.\(^{22}\) I conclude that there is no overall betterness relation among outcomes according to SU.\(^{23}\)

Some readers may have qualms about my assertion that there is no overall betterness relation according to SU. They might find the assertion that there is no overall betterness relation counterintuitive on its face, or they might be concerned that the lack of an overall betterness relation will have counterintuitive consequences. Or, their qualms might be more presentational. They might be inclined to think that choice-worthiness plays the role of overall goodness in SU, and therefore deserves that name; that rather than saying that there is no overall betterness relation according to SU, I should say that according to SU the overall betterness relations has a non-standard structure or non-standard relata. I respond to these qualms in Section 5.2, and ask the reader to set them aside until then.

### 4.2.5 SU is welfarist

SU is a welfarist axiological consequentialist theory, according to my definition of welfarist axiological consequentialism on page 39. Here is that definition again.

\textit{Welfarist axiological consequentialism}: The choice-worthiness relation in any decision, \( D_1 \), is fully determined by the morally relevant betterness relations among the available outcomes in \( D_1 \), which in turn are fully determined by people’s wellbeings in those outcomes.

According to such a theory, wellbeings are the only features of the available outcomes that determine choice-worthiness. This is very much like the argument Broome gives against neutrality, to the effect that it can’t be understood axiologically, by which he means in terms of a single betterness relation (see Section 3.2 and Broome 2004). For more on how multiple betterness relations allow one to capture such things axiologically see Cusbert and Kath 2016.

\(^{22}\)One might at first be tempted to think that there is an overall betterness relation according to SU, but that this relation is incomplete, and less closely connected to choice-worthiness than one might have expected. After all, SU agrees with CFPU, and it is tempting to say that an overall betterness relation holds among outcomes with the same population, and straightforwardly determines choice-worthiness in those decisions. Though this is to some extent a representational and terminological issue, I find that this idea does significant violence to my concept of overall betterness. It leads, for example, to the conclusion that it is sometimes the case that some outcome \( A \) is overall better than another outcome \( B \), but in some decisions \( B \) is more choice-worthy than \( A \). I think that it better represents the theory to say that there is no overall betterness relation, or universal overall incomparability. More on alternative representations in Section 5.2.
outcomes in a decision that play a part in determining choice-worthiness in 
that decision, and they do so by determining the relevant betterness relations 
among outcomes. Wellbeings are the only features of outcomes that I have used 
in my definitions of the various morally relevant betterness relations according 
to SU, so the latter are fully determined by the former.

4.3 SU and variable population decisions

SU is a general utilitarian theory. In this section I’ll show that SU also agrees 
with the four intuitions about variable population decisions that I endorsed in 
Chapter 3. I’ll apply the tests I developed in that chapter to SU, to show that SU 
is asymmetric about creating wretched people and creating happy people; that 
it deals intuitively with non-identity decisions; and that it avoids repugnant 
conclusions.

4.3.1 SU and wretchedness

My first intuition was that there is a wretchedness threshold of wellbeing: a level 
of wellbeing such that one ought not create a person with a wellbeing below that 
level, other things being equal. My hope was to find a general utilitarian theory 
according to which there is such a threshold. I remained agnostic about most 
of the details of the wretchedness threshold. I didn’t commit to any particular 
account of wellbeing, or of how to identify the wretchedness threshold, so I 
couldn’t say which lives are wretched and which are not. Fairly arbitrarily, 
I assigned to the wretchedness threshold (assuming there is one) the number 
zero on the wellbeing scale. We saw in Section 3.1 that TU is a theory according 
to which there is a wretchedness threshold, and that AU is not; that HU is 
not, but that NHU is. SU is another theory that agrees with my wretchedness 
intuition.

Here is $D_{10}$ again, from Section 3.1.

<table>
<thead>
<tr>
<th></th>
<th>Anne</th>
<th>Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{10}$</td>
<td>A</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>10</td>
</tr>
</tbody>
</table>

If there is a wretchedness threshold, and it is wellbeing zero, then $B$ is less 
choice-worthy in this decision than $A$. More generally, in any decision between
creating a person with a negative wellbeing and not creating that person, other things being equal, creating the person is less choice-worthy than not doing so. This is the case according to SU, in virtue of the complementary worseness of an outcome in which a person exists wretched, than an outcome in which that person does not exist. So in D_{10}, B is complementarily worse than A by 10, because someone exists at wellbeing \(-10\) instead of no-one existing at all. A is of course complementarily worse than B by 0, and each outcome is intersectionally worse than the other by 0. So the shortfall of B in this decision is 10, the shortfall of A is 0, and B is less choice-worthy in D_{10} than A. According to SU wellbeing 0 is the wretchedness threshold.

In my discussion of the wretchedness threshold in Section 3.1 I touched on the issue of whether existence (at some particular lifetime wellbeing or other) can be better or worse for a person than non-existence. I brought this up in the context of a possible account of the wretchedness threshold (though not one that I favour), as the level of wellbeing such that it is worse for a person that she lives a life at any wellbeing below that level, than that she does not exist at all. SU is silent on this issue, and I don’t think that settling it one way or the other would affect the plausibility of SU.\(^24\) It is potentially interesting to consider whether this kind of worseness for a person in some sense underpins, explains, or justifies (in part) the relations of complementary worseness. However, I’m not sure that this would be more than a terminological debate: you say ‘worse for \(p\)’, I say ‘worse with respect to but not for \(p\)’.\(^25\)

SU agrees with my wretchedness intuition, and thus passes my first test.

\(^24\) Whereas it does, I think, affect the plausibility of a harm minimising theory, making the difference between HU and NHU. According to the latter there is a wretchedness threshold; according to the former there is not. See page 60–61.

\(^25\) Assume, for the sake of argument, existence comparitivism: that it is worse for a person to exist wretched than not to exist at all. Then the explanation of complementary worseness in D_{10} is relatively straightforward: B is worse for \(p_2\) than A. If I face this decision and bring about B instead of A, you can point to \(p_2\) and object to my having brought about the state of affairs that is worse for him. Instead, assume non-comparitivism: that it is not worse for a person to exist wretched than not to exist at all. Then the complementary worseness in D_{10} cannot be based on the worseness of existence than non-existence for \(p_2\). It can, however, still be based in some way on the poor quality of life of \(p_2\) in B. It is worse in some morally relevant way that a person exist wretched than that they don’t exist at all, because their existence is so unpleasant for them. It need not be worse for them than non-existence. Note that there is a similar issue for the happiness element of complementary worseness. (Is the explanation that it is better for a happy person to exist than not to exist, but better by more for a happier person to exist?) Complementary worseness seems to me to be morally significant, regardless whether comparitivism is true. See also footnote 11 on page 99.
4.3.2 SU and Neutrality

My second intuition was that creating a happy person is morally neutral: in a decision between creating a happy person and not creating that person, other things being equal, it is not (ever) the case that one ought to create the happy person. (It is part of the intuition of wretchedness that it is also not the case that one ought not create that happy person.) A happy person is a person who lives a life at or above the wretchedness threshold, wellbeing zero. I hoped to find a general utilitarian theory according to which creating a happy person is morally neutral. We saw in Section 3.2 that TU fails and NHU succeeds straightforwardly, and that AU fails and HU succeeds in a less clear sense. SU is another theory that straightforwardly passes the neutrality test; according to which creating a happy person is morally neutral.

Here is $D_{12}$ again, to illustrate.

\[
\begin{array}{c|cc}
\text{D}_{12}: & \text{Anne} & \text{Bob} \\
\hline
A & 10 & \ast \\
B & 10 & 10
\end{array}
\]

This is a decision between creating a happy person and not creating him, other things being equal. If creating a happy person is morally neutral, $A$ and $B$ are equally choice-worthy in this decision (and however high we make the wellbeing of Bob in $B$). This is so according to SU. $A$ and $B$ are each intersectionally and complementarily worse than the other by 0. This means that in this decision the outcomes both have a shortfall of 0, and are equally choice-worthy.

SU has these two features that together constitute ‘the Asymmetry’: anti-wretchedness and neutrality about creating happy people. Because these two features are often thought to be in tension with one another, I shall spend some time discussing how they are reconciled in SU. This discussion will be somewhat speculative. SU is probably consistent with a number of explanations or justifications of the Asymmetry, and I don’t pretend to be sure which is the right one. For my purposes it is enough that SU reflects the intuitively compelling Asymmetry, without contradiction. Still, the issue is important enough to warrant some discussion.

The Asymmetry is reflected in the axiology of SU. I can describe how by reference to these three outcomes (not necessarily thinking of the decision
CHAPTER 4. SHORTFALL UTILITARIANISM

between them at the moment).

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>10</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>10</td>
<td>−10</td>
</tr>
<tr>
<td>$C$</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

According to SU, $B$ is complementarily worse than $A$ by 10. This is underpinned by the fact that $B$ is strictly worse than $A$ in some morally relevant respect; specifically, that $B$ is strictly worse than $A$ with respect to the pair of sets of people, $(\{p_2\}, \emptyset)$. It is worse in this respect that $p_2$ exists with wellbeing $-10$ than that no-one exists instead (other things being equal). On the other hand, according to SU, $A$ is complementarily worse than $C$ by 0. This is underpinned by the fact that $A$ is not strictly worse than $C$ with respect to the pair of sets of people, $(\emptyset, \{p_2\})$. In fact, $A$ is not strictly worse than $C$ in any of the respects of betterness in the axiology of SU. In any of the respects in which they are related at all, they are equally good.\(^{26}\) It is not worse that no-one exists than that $p_2$ exists with wellbeing 10 instead.

This suffices to show that SU passes the second test, and to explain why it does so. One might, however, think that this explanation requires further justification. Why think that $B$ is (strictly) worse with respect to $(\{p_2\}, \emptyset)$ than $A$, and $C$ not (strictly) better with respect to $(\emptyset, \{p_2\})$ than $A$? Why should negative wellbeings affect the morally relevant betterness relations in one way (via their sums), and and positive wellbeings affect them in a different way (via their averages)? It is promising that SU agrees with strong moral intuitions on these matters, but one might want further independent support for the way that it does so. I raised this issue of justification previously, in the context of the explanation of the Asymmetry according to NHU (on page 67). Again without attempting to provide a thorough justification of the Asymmetry (which would be too big a task), I’ll revisit this interesting issue in the context of SU.

According to SU, our moral obligation as utilitarians is to ensure that the wellbeings of the people who exist (as always, ever exist) do not fall short of what is available; that, having made our choice, it cannot be said of any existing person’s wellbeing that we might have brought about something better instead. (Of course we can’t always ensure this absolutely; in that case our obligation is to minimise the extent to which this occurs.) The reason $B$ above is strictly

\(^{26}\)These are only betterness with respect to the pair $(\{p_1\}, \{p_1\})$, the pair $(\emptyset, \emptyset)$, and the pair $(\{p_2\}, \emptyset)$. 
worse in a morally relevant respect than $A$—equivalently, $A$ is strictly better in a morally relevant respect than $B$—is that $p_2$’s wellbeing in $B$ ($-10$) falls short of what happens instead in $A$ (which is that no-one exists).\footnote{Again, this seems plausible to me independent of whether or not $B$ is worse for $p_2$ than $A$. See footnote 24 on page 120.} If I choose $B$ rather than $A$, in a decision between the two, then it can be said of some existing person’s wellbeing, that I could have brought about something better instead. If I choose $A$ rather than $B$, in that decision between the two, of no existing person’s wellbeing can it be said, that I could have brought about something better instead. People’s wellbeing in $A$ (the wellbeing of people who exist in $A$) are, in comparison with what goes on instead in $B$, unimpeachable. The reason $C$ above is not strictly better in any morally relevant respect than $A$, is that the wellbeing in $A$ of existing people are similarly unimpeachable in comparison with what goes on instead in $C$. If I choose $A$ rather than $C$, in a decision between the two, it cannot be said of any existing person’s wellbeing that I might have brought about something better instead. That’s also true if I choose $C$ in this decision, which is why neither outcome is better or worse than the other in a morally relevant respect.

In this context, the justification of the Asymmetry must include a justification of this kind of ‘unimpeachability’ of existing people’s wellbeing as the utilitarian moral goal. This is a difficult task. It is some support for SU that taking this to be our goal produces a theory that aligns with common utilitarian intuitions about how we ought to make decisions (variable population decisions in particular). Ideally there should be further, independent support for this goal. Some support may come from the fact that this goal seems to be related to two common sense ideas: 1) that we ought to be concerned with existing rather than merely possible people; and 2) that we ought to minimise ‘harm’. The first idea has been (I think mis-) used to produce ‘actualist’ moral theories: theories according to which only actual people matter.\footnote{See Parsons 2002 and Parsons 2003 for proposals and defence of actualist theories. For criticisms see Hare 2007, Arrhenius 2016 (especially Chapter 10), and Cusbert 2015.} SU is not an actualist theory. According to SU all possible people matter. However, whether they exist on a particular outcome affects how they matter, for the moral assessment of that outcome. The second idea is one that I might seem already to have dismissed. I have rejected the harm minimising versions of utilitarianism, HU and NHU. However, I think that there is a compelling idea at the heart of the
common sense idea that one ought to ‘do no harm’, which does not relate to harm specifically at all. That is the idea that I have tried to spell out by appeal to the notion of unimpeachability. It is the idea that our focus should in some sense be on worseness, rather than betterness, of existing people’s wellbeings; not personal worseness, as is the case according to HU and NHU, but worseness of whatever is the morally relevant kind.

These common sense ideas do not, perhaps, justify the Asymmetry, or SU’s explanation thereof, by themselves. However, it seems promising to me that further analysis of these ideas may provide some such justification. For the time being, I am satisfied that SU agrees with my Asymmetric intuitions, and in particular passes the neutrality test.

4.3.3 SU and non-identity

My third intuition was that one ought to create a happier person rather than a less happy person, other things being equal, even if the happier person and the less happy person are not identical (and similarly for decisions between creating one group of people and another). My hope was to find a general utilitarian theory according to which this is the case; a theory that deals intuitively with ‘non-identity decisions’. We saw in Section 3.3 that TU and AU are such theories, and that HU and NHU are not. (This is the well-known ‘non-identity problem’ for person-affecting theories like HU and NHU.) SU is another theory that deals intuitively with non-identity decisions.

A non-identity decision is a decision between two outcomes, such that some of the members of the population of each available outcome do not exist in the other available outcome: for example, a decision between outcomes A and B, such that there are some A-not-B people and also some B-not-A people. Here again to illustrate is decision $D_{18}$.

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>10</td>
<td>*</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>*</td>
<td>8</td>
</tr>
</tbody>
</table>

My intuition is that $A$ is more choice-worthy in $D_{18}$ than $B$, and that the reason this is so is that the $B$-not-$A$ person ($p_3$) is worse off in $B$ than the $A$-not-$B$ person ($p_2$) is in $A$. Not everyone shares this intuition, and one theory that repudiates it is HU. No-one who exists in $A$ is harmed in $A$, and no-one who exists in $B$...
is harmed in $B$. In neither of the available outcomes is any harm incurred; the outcomes are equally permissible, according to HU (and NHU). This strikes me as implausible, and my goal was to find a theory that avoids such conclusions.

SU agrees with my intuitions in non-identity decisions. In $D_{18}$ (which might be the decision between conceiving a disabled child while ill or a healthy child later, for example) $B$ is complementarily worse than $A$ by 2 and $A$ is complementarily worse than $B$ by 0. Each outcome is intersectionally worse than the other by 0. So, the shortfalls of outcomes $A$ and $B$ in this decision are 0 and 2, respectively. $A$ is more choice-worthy in $D_{18}$ than $B$; one ought to choose $A$ and not $B$ when faced with this decision. The key to this conclusion is the axiological fact that $B$ is worse than $A$ with respect to the pair of sets of people, $(\{p_3\}, \{p_2\})$. The members of $\{p_3\}$ fare worse in $B$ than the members of $\{p_2\}$ fare in $A$.

That was the simplest possible kind of non-identity decision: a decision between creating one person or another. It illustrates the basic motivation for the happiness element of complementary worseness and its role in SU. However, non-identity decisions can be more complex, and complementary worseness is more subtle than this simple case allows me to illustrate. By working through some more complex cases I will try to explain why I think that complementary worseness, rather than some near relative, is what matters.

Here is another decision I discussed in Section 3.3.

<table>
<thead>
<tr>
<th>$D_{20}$:</th>
<th>$x$ billion Earthlings</th>
<th>$y$ billion Martians</th>
<th>$y$ billion Martians</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>80, 80, ..., 80</td>
<td>80, 80, ..., 80</td>
<td>*, *, ..., *</td>
</tr>
<tr>
<td>$B$</td>
<td>80, 80, ..., 80</td>
<td>*, *, ..., *</td>
<td>60, 60, ..., 60</td>
</tr>
</tbody>
</table>

Instead of being a decision between creating one person or another, this is a decision between creating a large number of people, or the same number of entirely distinct people. SU makes this decision in much the same way as it made the one above. The calculations are very similar, except that the amount by which $B$ is complementarily worse than $A$, and thus $B$’s shortfall in $D_{20}$, is $20y$ billion (instead of 2). According to SU, and in line with the intuitions I appealed to in Section 3.3, there is more at stake in this large-scale decision than there is in the decision between two possible individuals above. Say I’m faced with this second decision—which might represent a broad social policy decision, or a decision about which people to send to colonise Mars—and I bring about $B$. Then, you can point to the $y$ million people at wellbeing 60 and
complain to me that instead of bringing those people into that kind of existence, I could have done something better instead: I could have brought some other people into a better kind of existence. The same kind of criticism you could make of p3’s wellbeing in D18, you can make of each of the y billion B-not-A people in D20 (and the criticism is greater in magnitude, in the sense that the difference in wellbeing is greater).

That was a case in which the number of A-not-B people and the number of B-not-A people is the same. I think that there is a very intuitive sense in such cases that the A-not-B people exist in A instead of the B-not-A people existing in B (and vice versa). If I bring about the complementarily worse outcome, then I have created some number (in this case y billion) of particular people instead of the same number of different people. What if the choice is between groups of people of different sizes? Here is the final decision from Section 3.3.

<table>
<thead>
<tr>
<th>D21:</th>
<th>x billion Earthlings</th>
<th>y billion Martians</th>
<th>z billion Martians</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>80, 80, . . . , 80</td>
<td>80, 80, . . . , 80</td>
<td>* , * , . . . , *</td>
</tr>
<tr>
<td>B</td>
<td>80, 80, . . . , 80</td>
<td>* , * , . . . , *</td>
<td>60, 60, . . . , 60</td>
</tr>
</tbody>
</table>

The choice here is between some number of people at wellbeing 80, and some different number of people at wellbeing 60. My intuition is that B is less choice-worthy than A in D21, regardless of what x, y, and z are. SU agrees. The interesting thing in this case is the question of by how much B is complementarily worse than A. This amount is the difference between the average wellbeing of the B-not-A people in B and the average wellbeing of the A-not-B people in A, multiplied by the number of members in the smaller of those two sets; either 20y billion or 20z billion. Why should we multiply the difference in wellbeing by the number of people in the smaller of the two groups? Because that is the number of people who exist at a lower wellbeing instead of someone else existing at a higher wellbeing.

Because complementary worseness matters, according to SU, SU agrees with my intuitions in non-identity decisions. I think that complementary worseness is the best way to capture what it is that is worse about some people existing at lower welbeings rather than other people existing at higher welbeings. I think that SU takes account of this genuine morally relevant kind of worseness in the best way.29

29 A note on how SU is situated with respect to Boonin’s (2014) non-identity argument (see foot-
4.3.4 SU and Repugnance

My fourth and final intuition, in Section 3.4, was that one ought to choose an outcome with a smaller, happier population rather than an outcome with a larger, less happy population, other things being equal and in a decision between the two. My hope was to find a general utilitarian theory according to which this is the case; a theory that avoids ‘repugnant conclusions’. We saw in that section that AU is such a theory, and that TU, HU, and NHU are not. (The latter two, remember, fail in the case of non-identity repugnant decisions.) SU is another theory that avoids repugnant conclusions. According to SU, in any two-way decision between an outcome with a larger, uniformly less happy population and an outcome with a smaller, uniformly happier population, one ought to choose the latter.

Let me go through the decisions I discussed in Section 3.4, and show how SU deals with them. Here is decision D_{22}.

\[
D_{22}: \begin{array}{c|cc}
 & 1 \text{ trillion people} & x \text{ trillion people} \\
A & 100,100,\ldots,100 & *,*,\ldots,* \\
B & 1,1,\ldots,1 & 1,1,\ldots,1 \\
\end{array}
\]

This is a decision between one outcome with a large population of happy people, and a second outcome with a larger population of people who are less happy, but still all have positive wellbeings. I haven’t yet said how much larger the second population is, but we know it is larger just because it contains all of the members of the first population and then some additional people. I call this a case of repugnant growth because the population of outcome B is an extension of that of outcome A. This is the kind of case in which I think many people find repugnant conclusions the most disturbing. This might represent, for example, a decision between a policy that will deliver moderate population growth and high wellbeing for all of the people who ever exist, and an alternative policy that will deliver high population growth and lower wellbeing for all of the people who ever exist.\footnote{Of course no real decision is ever so neat. Complete equality of wellbeing is a simplification.}

My intuition is that A is more choice-worthy in D_{22}.

\footnote{I think that the best way to think of SU is as rejecting P5 of that argument: ‘If an act does not wrong anyone, then the act is not morally wrong’. Boonin considers some consequentialist ways of rejecting that premise, but concludes (rightly, I think) that those he considers fail. However, the options he considers are limited to versions of consequentialism that appeal to overall goodness of outcomes; in my terminology, consequentialist theories that meet the overall betterness requirement. This only serves to bolster my claim that only a moral (consequentialist) theory that does not meet the overall betterness requirement will agree with my four intuitions.}
than B, regardless what \( x \) is.

Here is how SU deals with \( D_{22} \). \( A \) is intersectionally and complementarily worse than B by 0, so the shortfall of \( A \) in \( D_{22} \) is 0. \( B \) is complementarily worse than \( A \) by 0, but is intersectionally worse than \( A \) by 99 trillion. The members of the \( A \)-and-\( B \) set (which is the entire population of \( A \)) fare worse in \( B \) than they fare in \( A \). They fare worse by, on average, 99 units of wellbeing each, and there are a trillion of them. So the shortfall of \( B \) in \( D_{22} \) is 99 trillion. \( A \) is more choice-worthy in \( D_{22} \) than \( B \). What \( x \) is—how many extra people exist in \( B \)—affects this not at all.

This seems to me to be the right result for the right reason. The people common to both available outcomes are worse off in \( B \) than in \( A \), which is a reason not to choose \( B \). If I bring about \( B \) you can point to the wellbeings of \textit{those} people and criticise me for not having brought about something better instead—the same people existing better off. Of the rest of the \( B \) people (the \( B \)-not-\( A \) people) this is not true; nothing better replaces their wellbeings in \( A \). On the other hand, I can’t fault the wellbeings of any of the \( A \) people in \( A \); there is no available alternative in which I could have done something better instead. As we have seen, in the only other available alternative the wellbeings of the \( A \) people are worsened. So, I have no reason not to choose \( A \). I have a reason not to choose \( B \), and no reason not to choose \( A \). Thus, \( A \) is more choice-worthy in this decision than \( B \).

Next consider decision \( D_{23} \).

\[
\begin{array}{c|cccc}
\hline
\text{D}_{23}: & 1 \text{ trillion people} & (x + 1) \text{ trillion people} \\
\hline
A & 100, 100, \ldots, 100 & \ast, \ast, \ldots, \ast \\
B & \ast, \ast, \ldots, \ast & 1, 1, \ldots, 1 \\
\hline
\end{array}
\]

Assume that \( x > 0 \). This is a non-identity version of \( D_{22} \). No-one exists in both of the available outcomes in \( D_{23} \). I call this a case of repugnant substitution because the \( A \) people are wholesale replaced in \( B \) by different people (instead of being added to, as they are in a repugnant growth case). My intuition that quality ought not be sacrificed for quantity here combines with my intuition that it is worse if people exist with lower rather than higher wellbeings, even if the people with the higher wellbeings would be different people. My intuition here is that \( A \) is more choice-worthy in \( D_{23} \) than \( B \), regardless what \( x \) is.

\footnote{It is also difficult to imagine a situation in which the wellbeings of all past people will be affected (even if you think that it is possible to affect a person’s wellbeing after their death).}
SU agrees with this intuition. Here is how SU deals with $D_{23}$. $A$ is intersectionally and complementarily worse than $B$ by 0, so the shortfall of $A$ in $D_{23}$ is 0. $B$ is intersectionally worse than $A$ by 0, but is complementarily worse than $A$ by 99 trillion. This is the difference between the average wellbeing of the $A$-not-$B$ people in $A$ and the average wellbeing of the $B$-not-$A$ people in $B$, multiplied by the number of people in the smaller of those two sets of people (the $A$-not-$B$ set). So, the shortfall of $B$ in $D_{23}$ is 99 trillion. $A$ is more choice-worthy in $D_{23}$ than $B$. ($D_{22}$ and $D_{23}$ are morally very similar, even to the degree of morally relevant worseness involved.) What $x$ is does not affect this at all (so long as it is positive).\[31\]

### 4.3.5 Mere addition

In Section 3.4.2 I discussed the mere addition argument. How does SU account for the compellingness of that argument, but avoid its conclusion? To recap the argument, here are the three relevant decisions again.

<table>
<thead>
<tr>
<th>$D_{28}$:</th>
<th>10 million people</th>
<th>3 million people</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>100, 100, … , 100</td>
<td>$<em>$, … , $</em>$</td>
</tr>
<tr>
<td>$A^+$</td>
<td>100, 100, … , 100</td>
<td>70, 70, … , 70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$D_{29}$:</th>
<th>10 million people</th>
<th>3 million people</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A^+$</td>
<td>100, 100, … , 100</td>
<td>70, 70, … , 70</td>
</tr>
<tr>
<td>$B$</td>
<td>95, 95, … , 95</td>
<td>95, 95, … , 95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$D_{30}$:</th>
<th>10 million people</th>
<th>3 million people</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>100, 100, … , 100</td>
<td>$<em>$, … , $</em>$</td>
</tr>
<tr>
<td>$B$</td>
<td>95, 95, … , 95</td>
<td>95, 95, … , 95</td>
</tr>
</tbody>
</table>

The first premise of the (choice-worthiness version of the) argument is that $A^+$ is not less choice-worthy in $D_{28}$ than $A$. The second premise is that $B$ is more choice-worthy in $D_{29}$ than $A^+$. The conclusion of the argument is that $B$ is more choice-worthy in $D_{30}$ than $A$. SU agrees with both of the premises, but not with the conclusion. I argued in Section 3.4.2 that the conclusion follows from the premises in the overall betterness version of the argument, but does not follow in the choice-worthiness version of the argument (unless one accepts

\[31\] Even if $x$ is negative, the story is similar. But this is less controversial and makes $D_{23}$ no longer a repugnant decision.
the overall betterness requirement). This is a partial explanation of how it is that SU agrees with the premises, but not the conclusion, of the mere addition argument.

Here is how SU makes each decision. First, $A^+$ is not less choice-worthy in $D_{28}$ than $A$. In fact $A^+$ and $A$ are equally choice-worthy in $D_{28}$ according to SU. Each of $A^+$ and $A$ is both intersectionally and complementarily worse than the other by 0. Second, according to SU $B$ is more choice-worthy in $D_{29}$ than $A^+$. This is because $A^+$ is intersectionally worse than $B$ by some positive amount (the difference between the average wellbeing in $A^+$ and the average wellbeing in $B$, multiplied by 13 million). The shortfall of $A^+$ in $D_{29}$ is 25 million, and the shortfall of $B$ in $D_{29}$ is 0. Finally, $B$ is not more choice-worthy in $D_{30}$ than $A$. In fact, $A$ is more choice-worthy in $D_{30}$ than $B$. This is because $A$ is intersectionally and complementarily worse than $B$ by 0, while $B$ is complementarily worse than $A$ by 0 but intersectionally worse than $A$ by 50 million ($5 \times 10$ million).

The overall betterness version of the mere addition argument is valid. According to SU, however, the premises of that argument are false. That version of the argument succeeds in showing that the intuitively compelling moral judgments regarding these three decisions can’t be accommodated within a single (overall) betterness relation. We can account for them in terms of multiple respects of betterness; HU and SU both do so. Allowing for multiple morally relevant respects of betterness means that what at first might look like intrinsivity, need not be. Speaking sloppily, the axiological claims that lead to SU’s conclusions in these decisions are: $A^+$ is worse than $B$, $A^+$ is not worse than $A$, and $B$ is worse than $A$. These three claims are inconsistent if they refer to single relation of betterness. According to SU they don’t. More fully (though still informally), the claims are: $A^+$ is worse with respect to the $A^+$-and-$B$ people than $B$, $A^+$ is not worse than $A$ in any respect relevant to $D_{28}$, and $B$ is worse with respect to the $B$-and-$A$ people than $A$. These claims are all perfectly consistent.

What does SU say about the decision between all three outcomes?

<table>
<thead>
<tr>
<th>D_{37}</th>
<th>10 million people</th>
<th>3 million people</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>100,100,\ldots,100</td>
<td><em>,</em>\ldots,*</td>
</tr>
<tr>
<td>$A^+$</td>
<td>100,100,\ldots,100</td>
<td>70,70,\ldots,70</td>
</tr>
<tr>
<td>$B$</td>
<td>95,95,\ldots,95</td>
<td>95,95,\ldots,95</td>
</tr>
</tbody>
</table>

We’ve already established all of the relevant amounts of worseness. In this three-way decision the shortfalls of $A$, $A^+$ and $B$ are, respectively, 0, 25 million,
4.3. SU AND VARIABLE POPULATION DECISIONS

and 50 million. $A$ is more choice-worthy in $D_{37}$ than $A^+$, which in turn is more choice-worthy in $D_{37}$ than the worst option, $B$. We can reach all of these conclusions only by giving up the overall goodness assumption, and embracing menu-dependence of choice-worthiness. $A$ is more choice-worthy in the three-way decision than $A^+$ despite the fact that it the two are equally choice-worthy in a two-way decision.

The way that SU deals with each of these decisions seems right to me. At this point you might be concerned about how SU would apply to a series of decisions. What if I face the decision between $A$ and $A^+$ and choose $A^+$ (as SU seems to allow); and then later face a decision between $A^+$ and $B$ and choose $B$ (as SU seems to require)? Wouldn’t I then have made a series of permissible choices leading to a repugnant outcome? This is an instance of a more general concern about menu-dependence and stepping through a series of decisions to bad results, which I address in Section 5.1. My response is roughly that series of decisions of this kind are incompatible with the consequentialist framework of decisions under certainty that I have assumed.

That is the SU response to the mere addition argument. There is a similar argument, the benign addition argument. The first step in the mere addition argument concerns, roughly, the mere ‘addition’ of some happy people, with no change to the wellbeings of the ‘original’ people. The first step of the benign addition argument concerns the addition of some happy people while also raising the wellbeing of the original people. The way SU deals with the benign addition argument is a little more complex, and in some cases may be somewhat unintuitive. I shall address this argument in Section 5.3.3.

Conclusion

In this chapter I have proposed SU, and shown that it is a general utilitarian theory that agrees with four intuitions about variable population decisions that I find particularly compelling. I think that this makes SU the most promising general utilitarian theory I am aware of. It also confirms that there are general utilitarian theories that agree with all four of these intuitions. It has been shown that no moral theory that meets the overall betterness requirement agrees with all four of these intuitions. This provides some reason, I think, to reject the overall betterness requirement and explore the broader class of
general utilitarian theories more carefully. In the next chapter I defend SU from some possible objections. Some readers may conclude that these (or other) objections strengthen the motivation to explore the class of general utilitarian theories, rather than accepting SU itself.

To recap SU. We should assess each available outcome in a decision by comparing it with each other available outcome. Our focus should be the wellbeings of the people who exist in the outcome to be assessed, which we should compare with what happens instead in each other available outcome. The wellbeings of the people who exist in a particular outcome can be worse than what replaces them in another outcome in two ways. They are intersectionally worse if they are replaced by higher wellbeings among the same people. They are complementarily worse if they are replaced by higher wellbeings among different people; or, in some cases involving wretchedness, if they are replaced by no-one at all. In any decision, we should bring about an outcome that is worse than any single other available outcome in these ways by the least total amount. This is an outcome with the lowest shortfall in the decision.
Chapter 5

Defending Shortfall Utilitarianism

Shortfall Utilitarianism (SU) is a general utilitarian theory that is asymmetric concerning the creation of wretched and happy people, deals intuitively with non-identity decisions, and avoids repugnant conclusions. This provides some reason to favour the theory (at least for someone who shares my intuitions on these matters). In this chapter I address some possible objections to SU.

I divide the challenges into four kinds. First, in Section 5.1 I consider some concerns about the menu-dependence of choice-worthiness according to SU. I dismiss one version of these concerns as incoherent. I also do my best to clarify a coherent version of the concerns that may arise when it comes to extending SU to deal with risky decisions. In Section 5.2 I return to the issue of overall betterness. I defend my claim that there is no overall betterness relation among outcomes according to SU, but countenance an overall betterness relation among outcomes of a finer grain. I also address Arrhenius’s normative impossibility results. In Section 5.3 I consider the conclusions of SU in specific variable population decisions, which some people may find counterintuitive. I defend those conclusions, and in some cases also suggest possible adjustments to SU. Finally, in Section 5.4, I propose a way to extend SU to make risky decisions.
5.1 Menu-dependence

According to SU choice-worthiness is menu-dependent in interesting ways—what one ought to choose depends on what the available outcomes are. This raises a number of questions and possible concerns. Will SU lead to cyclical decision-making? Would that be problematic? When is an outcome available? Can one change the choice-worthiness of one outcome by affecting the availability of others? In this section I’ll address some of these concerns.

I’ll begin by spelling out the varieties of menu-dependence of choice-worthiness (from now on, just ‘menu-dependence’) entailed by SU. We have encountered some of them already. The rough idea of menu-dependence is that what one ought to choose depends on what the available options are. In one sense this is true according to all of the theories I have considered. Take Total Utilitarianism (TU) for example. According to TU, in any decision one ought to choose the available outcome with the highest total wellbeing. Say that in some decision, $D_1$, the available outcome with the uniquely highest total wellbeing is outcome $A$. In $D_1$ $A$ is highly (maximally) choice-worthy and ought to be chosen. But we can create another decision, $D_2$, just by making available another outcome with a higher total wellbeing than $A$—say, $B$. $B$ is more choice-worthy in $D_2$ than $A$; $A$ is not maximally choice-worthy and ought not be chosen in that decision.\footnote{One might say that the choice-worthiness ‘score’ of $A$ is the same in any decision, is just the total wellbeing of the population of $A$ in $A$, say. Still, whether $A$ ought to be chosen is menu-dependent in the way described.} This is not what I mean by ‘interesting’ menu-dependence. I mean something less common: when the relative choice-worthiness of two available outcomes changes from one decision to another.

One way this might happen is if a pair of outcomes are equally choice-worthy in one decision, and one of them is more choice-worthy in a second decision than the other. In Section 3.2 I argued that the neutrality intuition provides some motivation for accepting this kind of menu-dependence, by appealing to intuitions about decisions $D_{14}$ and $D_{17}$. Here are those decisions again.

$D_{14}$:

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>15</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>15</td>
<td>13</td>
</tr>
</tbody>
</table>

$D_{17}$:

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>15</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>$C$</td>
<td>15</td>
<td>17</td>
</tr>
</tbody>
</table>
According to the neutrality intuition, A and B are equally choice-worthy in D_{14}. I have the further intuition that A is more choice-worthy in D_{17} than B. SU agrees with these intuitions, and so we know that choice-worthiness exhibits this kind of menu-dependence according to SU. It is clear that any theory that captures these intuitions must be menu-dependent in this way.\(^2\) If there is a problem with this kind of menu-dependence, it is a problem not just for SU but for any theory that captures these compelling intuitions.

Menu-dependence might take another, more surprising form. It might be that for some pair of outcomes, A and B, A is more choice-worthy in one decision than B, and B is more choice-worthy in a different decision than A. According to SU choice-worthiness exhibits this kind of menu-dependence too. Here are some decisions to illustrate.

\[
\begin{array}{c|c|c|c}
D_{38}: & p_1 & p_2 & \text{Shortfall} \\
A & 10 & 8 & 0 \\
B & 8 & * & 2 \\
\end{array}
\hspace{1cm}
\begin{array}{c|c|c|c}
D_{39}: & p_1 & p_2 & \text{Shortfall} \\
A & 10 & 8 & 4 \\
B & 8 & * & 2 \\
C & 10 & 12 & 0 \\
\end{array}
\]

According to SU, A is more choice-worthy in D_{38} than B. However, B is more choice-worthy in D_{39} than A.\(^3\) Once again, this seems to me to be the correct conclusion in this case. Here is an intuitive explanation. In D_{38} something counts against choosing B, and nothing counts against choosing A. In D_{39}, something still counts against B, but something also counts against A to a greater degree (without also counting against B, which is perhaps the surprising thing).

Choice-worthiness is highly menu-dependent according to SU. Is this a problem? I’ll now address some possible concerns. Just before I do, I want point out a possible confusion about menu-dependence of choice-worthiness. It might sound, at first, as though this entails a strange kind of instability of betterness (or goodness, if you prefer). This is not the case. Betterness in each morally relevant respect is fixed, stable, once-and-for-all determined.

\(^2\)Broome (2004, 2009) shows that it is not possible to reconcile these intuitions within the overall betterness framework. He claims thus to have shown that it is not possible to make sense of them in terms of betterness at all, but in Cusbert and Kath 2016 we show that they can be accommodated in a framework of multiple respects of betterness. In this thesis I have proposed a particular moral theory in which they are thus accommodated.

\(^3\)In these simple cases relative permissibility is not menu-dependent. That would be the case if, for example, while A is permissible and B impermissible in D_{38}, B were permissible and A impermissible in D_{39} (in fact, of course, they’re both impermissible in D_{39}). Later, in Section 5.3, we’ll see that permissibility is also menu-dependent in some more complex cases according to SU.
Likewise, choice-worthiness in each decision. As I said in Section 1.2, choice-worthiness can also be represented in terms of betterness relations, and ‘A is more choice-worthy in $D_1$ than $B$’ can also be expressed as ‘it is morally better to choose A when faced with $D_1$ than to choose B when faced with $D_1$’. Each of the betterness relations, ‘being morally better to choose when faced with $D_1$’, is fixed, stable, once-and-for-all determined. What is at issue here is how choice-worthiness-in-one-decision differs from choice-worthiness-in-another-decision; or how the relation of being morally better to choose when faced with one decision differs from the relation of being morally better to choose when faced with another decision. This seems important to me, as I have difficulty accepting the idea that betterness (or goodness) might be menu-dependent in the sense of changing from one context to the next.

### 5.1.1 Series of decisions

The first kind of concern is that a theory according to which choice-worthiness is menu-dependent might deliver plausible conclusions in particular decisions, which prove implausible or problematic over a series of decisions. A person who applies such a theory might be susceptible to some kind of ‘money pump’; or she might be led to choose one outcome in a single decision, and a different outcome over a series of decisions in which all and only the same outcomes are available. I’ll illustrate these concerns using just three outcomes, and the various possible decisions between them. Here are the outcomes.

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>19</td>
<td>*</td>
<td>10</td>
</tr>
<tr>
<td>$B$</td>
<td>22</td>
<td>10</td>
<td>*</td>
</tr>
<tr>
<td>$C$</td>
<td>*</td>
<td>12</td>
<td>*</td>
</tr>
</tbody>
</table>

And here are the decisions. Under each decision I have included the ordering of the available outcomes for choice-worthiness in that decision, according to SU.

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{40}$:</td>
<td>*</td>
<td>10</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>19</td>
<td>10</td>
<td>*</td>
</tr>
<tr>
<td>$C$</td>
<td>*</td>
<td>12</td>
<td>*</td>
</tr>
</tbody>
</table>

$B >_{D_{40}} A$  
$C >_{D_{41}} B$
A quick summary of how SU makes each of these decisions. B is more choice-worthy in \( D_{40} \) than A because \( p_1 \) fares worse in A than in B. C is more choice-worthy in \( D_{41} \) than B because \( p_2 \) fares worse in B than in C. A is more choice-worthy in \( D_{42} \) than C because \( p_3 \) fares worse in C than \( p_1 \) and \( p_2 \) do, on average, in A. In \( D_{43} \) the reasoning is as follows. Something counts against B: that \( p_2 \) fares worse in B than in C by 2. Something counts even more strongly against C: that \( p_3 \) fares worse in C than \( p_1 \) and \( p_2 \) do, on average, in A by 2.5. And something counts still more strongly against A: that \( p_1 \) fares worse in A than in B by 3. So, although B is not impeccable in \( D_{43} \), it is more choice-worthy in \( D_{43} \) than C, which in turn is more choice-worthy in \( D_{43} \) than A.

I think that the conclusions of SU in these decisions are quite compelling. I’ve deliberately chosen decisions in which this is the case because I want to focus here on the issue of the menu-dependence of choice-worthiness, and its possible consequences for series of decisions. I hope that the reader also finds the conclusions in each decision intuitively acceptable; if not, however, I ask her to set her other concerns aside. Now, let me present as best I can some of the possible concerns about menu-dependence itself.

First, the apparent ‘money pump’ problem.\(^4\) Imagine you face decision \( D_{40} \). According to SU you ought to choose B and not A. Perhaps you ought to be willing to pay some money to bring about B rather than A.\(^5\) Imagine you subsequently face decision \( D_{41} \). According to SU you ought to choose C and not B. Perhaps you ought to be willing to pay some money to bring about C rather than B. Finally, imagine you go on to face decision \( D_{42} \). According to SU you ought to choose A and not C. Perhaps you ought to be willing to pay some money to bring about A rather than C. Now it seems that SU has led you through a series of decisions, paying at each step, to get to A; which was

\(^{4}\)The money pump idea is first suggested in Ramsey 1931 and developed in Davidson et al. 1955. For a recent treatment see Gustafsson 2010.

\(^{5}\)It seems simplest to use money here to describe the problem. However, in the best version of the objection these ‘payments’ would probably be made in wellbeing. After all, it isn’t so obviously problematic to pay a lot of money to fulfil one’s moral obligations, money being morally irrelevant in itself. Thanks to John Cusbert for useful discussion on this point.
available in the first decision, and which you chose (perhaps paid) to avoid. And it seems this process could go on and on. You might face an endless cycle of decisions: \(D_{40}, D_{41}, D_{42}, D_{40}, D_{41}\), and so on, choosing \(B\) and then \(C\) and then \(A\) and then \(B\) and then \(C\). This seems concerning.

A second, similar problem is the following. Imagine you face decision \(D_{41}\). According to SU you ought to choose \(C\). Imagine you subsequently face \(D_{42}\). According to SU you ought to choose \(A\). If you follow SU, you bring about \(A\) despite the fact that both \(C\) and \(B\) were available to you at some point. But according SU, in the three-way decision \(D_{43}\) both \(B\) and \(C\) are more choice-worthy than \(A\). So according to SU, if you face the three-way decision you ought to do one thing (choose \(B\)), but if you face the two pairwise decisions you ought to do something quite different: choose \(C\) and then, finally, \(A\). This also seems odd.

I will respond to these two apparent problems in two steps. My first response is to point out that the problems as described are inconsistent with the framework in which SU operates. SU is a theory about how one ought to make decisions between complete possible worlds (or at least worlds with their populations and wellbeings fixed) under certainty. The idea of a series of such decisions is incoherent. Any objection to SU on the basis of the conclusions of SU in series of decisions is also incoherent and ineffective. My second response is to ask whether a similar, but coherent concern may threaten extensions of SU to deal with risky decisions. As I’ve said before, the assumption of certainty must be temporary. In a framework that allows decisions under risk, series of decisions make sense (indeed I assume we face such series of decisions all the time). Will an extension of SU be susceptible to money pumps or the like? If so, would that be problematic? Answering these questions would be much too vast an undertaking to attempt here. However, I hope that my discussion will go some way towards clarifying what the important questions are.

Let me now spell out my first response more fully. The crux of this response is that the apparent problems for SU are incoherent, because they involve series of decisions between complete possible worlds under certainty. In each of the problems I just described, it was suggested that one might face first \(D_{41}\) (between \(B\) and \(C\)) and then \(D_{42}\) (between \(C\) and \(A\)), so I’ll focus on that particular short series of decisions in my explanations.

The problem is that, if one of these decisions is faced, the other cannot
5.1. MENU-DEPENDENCE

be. First, assume that someone in fact faces $D_{41}$. This means that this person must choose either $B$ or $C$. If she chooses $B$, then the actual world will, with certainty, be the $B$-world; and if she chooses $C$, then the actual world will be the $C$-world.\(^6\) Say she makes $D_{41}$ by choosing $C$.\(^7\) This determines with certainty who will ever exist, and what wellbeings they will have. It determines, among much else, that $p_1$ will not exist, and that $p_2$ will exist with wellbeing 12. So far, so good. Here comes the problem. We’re asked to consider a scenario in which this person goes on to face another decision about who will ever exist and what wellbeings they will have: $D_{42}$, between $C$ and $A$.\(^8\) But we’ve just assumed that $C$ has been fixed as the actual world; that it has been determined that $p_1$ will never exist and that $p_2$ will. Now, we’re asked to suppose that who exists is up for grabs again; that it is possible that in fact the actual world is $A$; that this person can undo her previous decision and make it the case that $p_1$ will exist and that $p_2$ will not. This kind of revisability is inconsistent with the world-fixing picture of the first decision (and of decisions in general) that we assumed. Having chosen $C$ in some decision, one can’t find oneself again facing a decision about whether or not to bring about $C$. There can be no available outcome other than $C$; and a decision must have more than one available outcome.

Consider things from the other direction. Assume that someone faces $D_{42}$, between $C$ and $A$. This means that $A$ and $C$ are both available to this person; that it is possible for her to choose to bring it about that $C$ is the actual world, and also possible for her to choose to bring it about that $A$ is the actual world. It cannot be that this person (or any other) has already chosen $C$ in a decision, thereby fixing $C$ as the actual world. Assuming that $A$ genuinely is available to some person is inconsistent with $C$ having already been chosen. Outcomes $A$ and $C$ are different complete possible worlds. It is not possible to bring about one and (then) the other. The model of decisions that I have assumed—decisions under certainty between possible worlds—is incompatible with series of decisions. SU is a theory for making decisions under certainty between possible worlds. As such, it is not a decision for making series of decisions. The claim that

---

\(^6\)I say ‘the $C$-world’, but because the world is only specified by population and wellbeings, it is more properly a $C$-world.

\(^7\)My response does not depend on her choosing $C$. It is if anything more stark if we assume she chooses $B$. She must choose one or the other, by hypothesis (that is what it means for these to be the only two available outcomes; see Section 1.2).

\(^8\)My response does not depend on it being the same person facing the two decisions.
SU makes series of decisions in such-and-such a way is mistaken, and cannot constitute an effective objection to SU.

I do not mean to suggest that people do not in fact face series of decisions. Of course, we do. My point is that it is not possible to represent series of decisions within the simplifying framework that I have adopted (and which, remember, is often adopted in the population ethics literature). I adopted this framework because it enables me to focus on some issues in particular, and bracket others (such as risk and rights). It is unforeseen, but I think not necessarily problematic, that this framework also brackets issues concerning series of decisions. This is due, I think, to two elements of the framework: the use of complete possible worlds as the alternatives in decisions, and the assumption of certainty.

Take the first element. It was crucial in my response above that the decisions hypothesised were decisions between complete possible worlds. Series of decisions under certainty between some less inclusive (‘narrower’) alternatives are perfectly possible. Consider decisions about making exchanges: for example, exchanging this banana for that apple, or exchanging that apple for this pear. We could think of exchanges as events, or temporal sections of possible worlds, or decisions between states at a time. I can make the decision at one time to exchange this banana for that apple (rather than not, say), and then make the decision at a later time to exchange that apple for this pear. Having brought about one event (fixed some temporal part of the world), I can later bring about another (fix a different part of the world). So, it is possible that I might make a series of cyclical decisions, exchanging this banana for that apple, that apple for this pear, this pear for this banana, and so on. I could not, however, make a series of decisions each of which determined exactly which fruits I ever own (which would be analogous to a series of decisions that each determine which people ever exist and with which wellbeings).

I’m not sure that there is much scope for altering this element of the framework. Possible worlds seem to me to be good candidates for the objects of consequentialist (and in particular utilitarian) concern. We are concerned with the consequences of our actions throughout the world, with no temporal or spacial boundaries (though with, perhaps, restriction to people and their well-

---

10Strictly, the alternatives in decisions are acts and possible worlds are the outcomes of acts.
11See Hansson 1993 for a similar case involving swapping stamps.
beings). Focussing on decisions between less inclusive alternatives would not enable a proper representation of a consequentialist moral theory.

The other important element of the theory, which I think we can more readily alter, is the assumption of certainty. Series of decisions under uncertainty between prospects (possible worlds and their associated probabilities) are perfectly possible. Let me illustrate by proposing one coherent series of decisions that shares some similarities with the incoherent series, \( D_{41} \) followed by \( D_{42} \). Imagine someone faces a risky decision between two acts, \( b \) and \( c \), with the following prospects. Act \( b \), if chosen, would lead with probability 1 to outcome \( B \). Act \( c \), if chosen, would lead with probability 0.7 to outcome \( C \), and with probability 0.3 to outcome \( A \). Imagine the person chooses to perform act \( c \). Imagine she then faces a second decision, between acts \( a \) and \( d \), with the following prospects. Act \( a \) will lead with probability 1 to outcome \( A \). Act \( d \) will lead with probability 1 to outcome \( C \). (This second decision is a decision under certainty: in fact it is just \( D_{42} \).) The first decision differs from \( D_{41} \) in the introduction of uncertainty about one of the available acts’ outcome. Instead of a decision under certainty between \( B \) and \( C \), we have a decision between \( B \)-with-certainty, and risky-\( A \)-or-\( C \). This element of uncertainty might in fact be due to the second decision. At the time of the first decision it might be that the probabilities that you would choose \( C \) and \( A \) in the second decision, if you faced it, are 0.7 and 0.3 respectively.\(^{12}\)

There seems to me to be more scope for adjustment regarding this element. In fact, (as I said in Section 1.2.2) we must eventually give up the assumption of certainty and extend SU to deal with risky decisions, in order for the theory to be practically useful. In Section 5.4 I propose a way of doing this. Here, I’ll just consider the question of whether some version of these series-based problems might arise when we do so. I can think of two versions of the problem that might arise. One is based on mistakes about which outcomes are available, and whether they are certain. The other is a potential problem for an extended moral theory for making risky decisions.

Here is a version of the money pump problem I described above that is consistent with the framework of SU. Imagine you believe, mistakenly, that you

\(^{12}\)Though there is debate about whether one can have credences about one’s own action with respect to a decision about which one is presently deliberating, it is widely accepted that one can have credences about one’s more distant future actions. See Rabinowicz 2002 and Hájek Forthcoming.
face decision $D_{40}$, the decision under certainty between $A$ and $B$. According to SU you ought to choose $B$ when faced with $D_{40}$, so you choose what you think is $B$-with-certainty. In fact you did not face decision $D_{40}$, as you discover when you subsequently find yourself facing another decision. This decision you believe, again mistakenly, to be decision $D_{41}$, the decision under certainty between $B$ and $C$. According to SU you ought to choose $C$ when faced with $D_{41}$, so you choose what you think is $C$-with-certainty. In fact you did not face decision $D_{41}$, as you discover when you subsequently find yourself facing another decision. This decision you believe, perhaps correctly, to be decision $D_{42}$, the decision under certainty between $C$ and $A$. According to SU you ought to choose $A$ when faced with $D_{42}$, so you choose (what may in fact be) $A$. Now it seems that SU has led you through a series of decisions to get to $A$, which was (we can assume) available in the first decision, and which you chose (perhaps paid) to avoid. And it seems this process could go on and on.

It is true that following the guidance of SU under a misapprehension that one faces a decision under certainty will sometimes lead to unappealing results. I don’t think that this is a problem for SU, or for the phenomenon of menu-dependent choice-worthiness. It should not be surprising that a theory for making decisions under certainty might, given such false input, return an undesirable result. This is the case even for theories that are not menu-dependent, such as Average Utilitarianism (AU). Say I think that I face a decision under certainty between two outcomes with the same average wellbeing, when in fact one of those outcomes is a risky prospect and ends up leading to a lower average wellbeing. It is not a fault with AU that misapplying it in this case might lead me to choose the risky prospect, and in the end the less choice-worthy outcome (less choice-worthy by the lights of AU itself). We should only expect a theory about how one ought to make decisions under certainty to provide good guidance when applied to actual decisions under certainty.

That was my first response to these concerns about series of decisions. My second response is to acknowledge the possibility that a coherent version of these concerns will arise when we give up the assumption of certainty—as I think we must. How will an extension of SU make series of risky decisions? Might such a theory lead to cyclical or otherwise surprising patterns of decision-making? If so, would that be problematic? My answer on all counts is: I’m not sure. Clearly, the answer to the first question depends on which extension of SU
we consider. I propose one way of extending SU to risky decisions in Section 5.4, but there are many other possible ways. In any case, it is beyond the scope of my proposal here to establish what kinds of patterns of decision-making even that one extension of SU produces. I must also defer any attempt to answer the latter questions: whether, for example, cycles in risky decision-making would be problematic. There is a vast literature on this kind of problem in series of decisions. I hope that by pointing out the incoherence of series of decisions under certainty between possible worlds, I have also helped to clarify which concerns may be legitimate when it comes to assessing moral theories about how one ought to make risky decisions.

This concludes my discussion of objections to menu-dependence on the basis of series of decisions. First, these concerns are misdirected. SU is a theory about how one ought to make decisions under certainty between complete possible worlds. Series of such decisions are incoherent. Thus, concerns about how SU will make such series of decisions is misdirected. Series of decisions under certainty between parts of possible worlds, or some other narrower alternatives, are possible but don’t adequately reflect consequentialist concerns. Series of risky decisions between prospects are possible, and SU must eventually be extended to deal with them. However, the implications of mis-applications of SU to risky decisions should not concern us. And finally, whether an extension of SU would lead to cyclical decision-making, and whether that should concern us if it did, is beyond my scope here.

### 5.1.2 Manipulating availability

Another concern about menu-dependent choice-worthiness seems to arise in the context of a single decision. If the choice-worthiness of outcomes depends on the decision in which they are encountered, one might wonder whether it is possible for a person to manipulate her decision context, and thereby manipulate the choice-worthiness (and even permissibility) of the outcomes available to her. I can illustrate this concern, too, using these by now familiar decisions.

---

13For example, Joyce 1992 and Rabinowicz 1995.
14Thanks to Leon Leonteyev bringing this to my attention, and for useful discussion of this and related issues.
CHAPTER 5. DEFENDING SHORTFALL UTILITARIANISM

\[
\begin{array}{c|cc}
D_{14}: & p_1 & p_2 \\
A & 15 & * \\
B & 15 & 13 \\
\end{array}
\]

\[
\begin{array}{c|cc}
D_{17}: & p_1 & p_2 \\
A & 15 & * \\
B & 15 & 13 \\
C & 15 & 17 \\
\end{array}
\]

Says SU: A and C are equally choice-worthy (and both permissible) in D_{17}, and more choice-worthy in D_{17} than B (which is impermissible in that decision). A and B are equally choice-worthy (and both permissible) in D_{14}. The question is: faced with D_{17}, can I somehow make C unavailable, thus transforming my decision into D_{14} and making B equally as choice-worthy (in D_{14}) as A (and permissible, rather than impermissible). Can I manipulate the choice-worthiness of some outcomes by manipulating the availability of others?

There are two versions of this idea. The first version involves the decision-facer doing something to remove the option of bringing C about. Imagine someone facing the available options, A, B, and C. She has a preference for B, but recognises that B is impermissible in this decision. She also realises that if C were not available, B would be permissible (she would then face D_{14} instead of D_{17}). So she conceives of, and carries out, some way of making C unavailable. (Perhaps she throws some crucial medicine overboard.) Such an act has an air of irreproachability. There are two equally choice-worthy, permissible options in her decision, A and C, and so not bringing about C is also permissible. By doing something to rule out C she still leaves herself with a permissible option: A. However, having ruled out C she seems to change the decision that she faces, into one in which B is also permissible. She can then, it seems, choose B without having done anything wrong.

This version of the idea involves a series of decisions, and so (for the reasons discussed at length in the previous section) misses its target. My best coherent reconstruction of this scenario is the following. Our agent faces first a four-way, risky decision between A-with-certainty, B-with-certainty, C-with-certainty, and a risky prospect with some chance of A and some chance of B. This last option is the prospect of the act that, if she performs it, makes C no longer available. This act will likely lead to B (given the agent’s assumed reasons for carrying it out and preference for B over A), but may with some small probability lead to A (in case the agent changes her mind, or make a mistake). Call this four-way risky decision D_{17}. If she chooses the risky prospect, in D_{17}, the agent will then face D_{14}, between A and B.
SU doesn’t tell us how one ought to make \( D_{17} \). However, I think we should expect an extension of SU to conclude that in this particular risky decision, the choice-worthiness ordering is as follows: \( A \)-with-certainty and \( C \)-with-certainty are most choice-worthy in \( D_{17} \), followed by the risky prospect (probably \( B \), but with some chance of \( A \)), followed by \( B \)-with-certainty. The air of irreproachability of the risky option is dispelled on reflection. It would be impermissible to perform this \( C \)-precluding act, because that would likely lead to \( B \). It is true that, having made this impermissible choice, it would then be permissible for the agent to bring about \( B \). However, she will not have done so ‘without having done anything wrong’.\(^{15}\)

That was one version of the idea: doing something to alter the availability of outcomes. The other version of the idea doesn’t involve acting, or altering availability of outcomes at all. Rather, it concerns the process of thinking or reasoning through a decision. One might think that someone facing decision \( D_{17} \) should be able to reason their way through the decision like this: ‘Although \( C \) is one of the most choice-worthy available outcomes, it is permissible for me not to choose \( C \). Not choosing \( C \) still leaves me with the equally choice-worthy and permissible option of choosing \( A \). So surely I can decide not to choose \( C \). Fine, I decide not to choose \( C \). Having decided not to choose \( C \), shouldn’t I now treat \( C \) as unavailable, and consider myself to be facing decision \( D_{14} \)? In which case, isn’t \( B \) now equally as choice-worthy in the decision I face as \( A \), and permissible?’

I understand this version of the idea of the malleability of availability to constitute a (partial) alternative to SU. What it seems to suggest is that it is in fact not the case that one ought to make decisions by minimising shortfalls (in the way set out by SU); rather, one ought to make decisions in a different way—a way that allows one to treat permissible outcomes as unavailable in the proposed way. My response to this proposal is to argue that it constitutes a worsening, rather than an improvement, of SU. I think that the original objector would agree. The original point was that this alternative might be supported, or at least allowed, by SU. In fact, this version of the idea conflicts with SU. SU

\(^{15}\)For a defence of act consequentialism appealing to a similar thought see Braddon-Mitchell 2007. Braddon-Mitchell argues that cultivating certain inabilities (perhaps just mental) might be the right thing to do, because it might have better overall consequences to be that kind of person, than to be the kind of person who follows utilitarianism applied to a wider range of available outcomes. The point, I think, is that a version of rule utilitarianism is entailed by act utilitarianism (and certain empirical psychological facts).
reaches the right conclusion in $D_{14}$, I think, and does so for compelling reasons. According to SU, one ought not ‘treat $C$ as unavailable’ when one faces $D_{17}$. The fact that $C$ is available is what makes $B$ impermissible, and we can’t just reason $C$’s availability away. It seems likely that any theory according to which one treats $C$ (or any other outcome) as unavailable when it in fact is available, will deliver unappealing conclusions.\footnote{There is a version of this kind of concern even for theories that are not menu-dependent (though it does have less significance there). Say I face a decision between giving you ten, twenty, or a hundred dollars, and say I decide that I have already made up my mind that I will not give you the hundred dollars (but haven’t done anything to rule the option out: I have the money in my hand, I can still reach over and pass it to you). I don’t think that I should then take myself to face a different decision—just between giving you ten or twenty dollars. Although in this case the recommendation will be the same—I should give you twenty dollars—I should not take myself to have done something permissible in giving you just twenty dollars, when I genuinely could have given you a hundred.}

This discussion leads me back to the question of which decision a person actually faces, or when any given outcome is available. The rough idea I have relied on so far is that an outcome is available to a person at a time just in case it is possible for that person to bring about that outcome at that time.\footnote{That is, do the bringing about at that time; not that the outcome need occur at that time (outcomes being complete possible worlds and not temporal parts thereof). See Section 1.2.} This rough idea can be made more precise in a number of ways, by appeal to any of a number of versions of possibility, which fall on a spectrum from objective to subjective. For each of these versions of possibility and of availability, there is a corresponding version of choice-worthiness. The conclusions SU delivers when applied to decisions picked out by a highly objective version of availability should be interpreted in terms of a highly objective version of choice-worthiness, and so on for the other versions. Which version of availability we should use will depend on which version of choice-worthiness we are interested in.\footnote{This is a familiar distinction. Consequentialists have long debated whether the morally significant consequences (outcomes, or perhaps prospects) are the objectively possible ones or the epistemically possible ones. Should we assess an act on the basis of its actual consequences, or on the basis of its foreseen or foreseeable consequences? If I flick my light switch with the unforeseen consequence that my neighbour’s house explodes, did I do something permissible or something highly impermissible? Some people associate the term ‘rightness’ with more subjective moral analysis, and ‘goodness’ with more objective moral analysis, so that for example one can do what is most right without doing what is best. I think that using this terminology would be confusing here, though it may help the reader familiar with that distinction to understand what I’m talking about. See for example Thomson 1989.}

I happen to be most interested in a fairly subjective version of choice-worthiness, and thus think that the most useful version of availability is fairly subjective too. Here is a candidate necessary and sufficient condition for avail-
ability that approximates my approach: an outcome is available to me just in case I believe (or perhaps, ought rationally to believe) that it is within my power to bring that outcome about. Something like this condition seems to me to produce a relatively practically significant version of SU. After all, we can’t very well apply the theory to the decision that we objectively face, having no way of knowing what that is. Similarly, it would seem somehow inappropriate to attribute blame on the basis of a more objective version of the theory. However, someone might disagree with me about which version of the theory is most interesting or useful. Quite plausibly, there are various interesting and useful versions, along the scale of subjectivity and objectivity. The important point here is that, whichever version one chooses, if an outcome is available to some person at some time in the relevant sense, she cannot make it unavailable (in that same sense) merely by making up her mind not to choose it.

Let me close this section by summarising my responses to concerns about the menu-dependence of choice-worthiness according to SU. The concerns as I understand them all take this form: if the relative choice-worthiness (and potentially even permissibility) of a pair of outcomes depends on which other outcomes are available, might changes in availability affect choice-worthiness (and permissibility) in unappealing ways? In many cases, ‘changes in availability’ are understood in terms of series of decisions: first one decision is faced, and then another. I have argued that there is no place for series of decisions within the framework of moral theories I have adopted—there can be no series of decisions under certainty between complete possible worlds—and that this kind of objection therefore misses its target. I have indicated that a similar kind of concern might have some force when it comes to extending SU to deal with decisions under uncertainty. However, it is not obvious exactly what features an extension of SU might have, or whether these concerns would constitute an objection to such an extension; and figuring all of this out would take me too far afield. In some cases, ‘changes in availability’ are understood to occur within the context of a single decision. I have suggested that a useful version of SU will employ a version of availability that is objective enough to rule out such problematic changes in availability; and that simply ‘treating’ an outcome as unavailable constitutes a violation of SU, which I do not endorse.
CHAPTER 5. DEFENDING SHORTFALL UTILITARIANISM

5.2 More on overall betterness

Throughout this dissertation I have distinguished between choice-worthiness and overall betterness. I introduced in Section 1.2 the distinction between these two notions: how good an outcome is (or would be), in a morally relevant respect, or overall; and how morally good it is (or would be) to bring about that outcome when faced with a particular decision. I attached the ‘choice-worthiness’ terminology to the latter notion. In Section 2.2.2 I made some assumptions about goodness and betterness: I endorsed what I call the ‘standard’ structure of betterness relations. It is partly in light of these distinctions and decisions that I have further claimed that there is no overall betterness relation among outcomes according to some theories, including SU. Choice-worthiness according to these theories has a structure that cannot be accommodated within a single standardly structured betterness relation.

I noted in Section 4.2.4 that some readers may have qualms about this move away from overall betterness. I’ll now address those qualms. I’ll split them into two kinds. First, I’ll address some fairly direct possible objections to the claim that there is no overall betterness relation among outcomes (including Arrhenius’s ‘normative impossibility results’). I’ll argue that these objections fail. Second, I’ll address some possible alternative ways of representing the axiological structure of moral theories. Without taking this representational issue too seriously, I’ll reject some alternatives but endorse the option of representing moral theories in terms of an overall betterness relation among outcomes of a ‘finer grain’.

5.2.1 Objections to giving up overall betterness

The first objection to the claim that there is no overall betterness relation among outcomes is simply to insist that there must be such a relation. We are used to making judgments of overall betterness, in moral and other contexts. The picture of weighing goodness in the relevant respects against one another, and coming to an assessment of overall betterness, or betterness all-things-considered, is a very familiar and appealing one. Here is Huemer endorsing the assumption that comparisons of overall betterness (he says ‘overall value’) can be made.

The Repugnant Conclusion, like other theses in population ethics,
5.2. MORE ON OVERALL BETTERNESS

asks us to compare possible worlds in terms of their overall value. I assume that such comparisons are possible, and that we may rely on our ethical intuitions in making such comparisons—whether directly or through reasoning based on abstract principles. These assumptions are nontrivial. . . . When I ask myself whether it would be better for there to be one billion barely-worth-living lives or one million wonderful lives, it seems to me that I clearly understand the question, and thus that it is not incoherent or meaningless. I am not asking which would be better for some particular person or group, nor which would be better for some particular purpose, nor, in general, which would be better in some particular way. I am asking which would be better in the generic, agent-neutral sense—'from the point of view of the universe', in Sidgwick's phrase. . . . These assumptions are commonly taken for granted in population axiology. . . (Huemer 2008, 900–901).19

This reinforces and partly explains what we also saw in Section 2.2.3: that the assumption is often made in the population axiology literature (to varying degrees explicitly) that there is an overall betterness relation among outcomes. I rejected this assumption, and in fact I see no reason to make it (apart from habit and oversight of the alternatives). In this dissertation I have discussed numerous plausible moral theories according to which there is no overall betterness relation among outcomes—Harm Utilitarianism (HU), Non-existence Harm Utilitarianism (NHU), and SU the main examples. These theories posit the existence of various relations of betterness in morally relevant respects, and spell out how these relations should determine choice-worthiness in any given decision. Choice-worthiness itself can be represented in terms of betterness relations: one for each decision. The intuitive conclusions in variable popu-

---

19Temkin makes some similar claims (see Temkin 1987 and Temkin 2011). However, I am not quite sure how to interpret Temkin's claims about overall (or all-things-considered) betterness. My current focus is on the claim that there is no overall betterness relation among outcomes, where by 'overall betterness relation' I mean a single, fixed, standardly structured betterness relation. Using this terminology, and in particular given my (the standard) definition of betterness relations, Temkin would I think agree with this claim. However, he employs a broader definition of betterness relations, allowing that some (including the 'overall betterness relation') might be intransitive. Using his terminology, there is an overall betterness relation, but it is not fixed, or not standardly structured, or in fact made up of numerous relations, depending on your reading of Temkin. It is the existence of this broader thing (what I call choice-worthiness), that he adamantly defends. I take Huemer and others to assume the existence of what I call an overall betterness relation. More on alternative representations shortly.
lation decisions reached by SU, and the explanations in terms of betterness in respects underlying those conclusions, provide support for the idea that there need not be an overall betterness relation among outcomes. I have also illustrated by way of non-moral analogy (with beaches or bands, in Section 2.2.3) that the lack of an overall betterness relation might occur in other contexts. I see no reason whatsoever to persist in the assumption that there is an overall betterness relation in the face of these alternatives—betterness in respects determining choice-worthiness in decisions—and their success in capturing our intuitive evaluative judgments and decision-making processes.

There is a good explanation of why these assumptions are commonly taken for granted, in population axiology and beyond. First, in many contexts, including the context of fixed population decisions, respects of betterness seem to interact in such a way that they can be aggregated and represented in terms of a single betterness relation. It makes sense, in such contexts, to call this ‘overall betterness’. Second, perhaps in part due to this fact, the idea of combining respects of betterness to produce an overall betterness relation is a compelling, almost natural, one. We are used to proceeding in this way, and deviating from this picture complicates things quickly. However, in the face of good reasons for deviating from the picture, and for thinking that the various respects of betterness do not interact so as to produce a single betterness relation, we must be willing to break the convenient habit of thinking in terms of overall betterness.

Apart from thinking that there just must be an overall betterness relation among outcomes, one might worry that the absence of such a relation will have problematic implications. It can indeed cause problems to introduce some incomparability into an otherwise complete overall betterness relation, and derive choice-worthiness in a familiar way from that incomplete overall betterness relation. Let me illustrate the kind of problem I have in mind. Then I’ll explain why doing away with overall betterness altogether, in the way I propose, does not face the same kind of problem.

Imagine an overall betterness relation among outcomes such that, for any

---

20 Note that one need not think that questions or claims about overall betterness of outcomes are ‘incoherent or meaningless’. I agree with Huemer that they are not. I also understand such questions and claims; I can answer many of the questions, and judge the claims true or false. The answer to the question ‘Is outcome A overall better than outcome B?’ is ‘No’. The claim that A is overall better than B is false; as is the claim that A is overall worse than, or equally as good as, B.

21 We saw that this is the case for HU and SU, for example, when I showed that each agrees with CFPU and thus can be represented in terms of a single betterness relation in the context of fixed population decisions only. See Sections 2.3 and 4.2.
pair of outcomes, $A$ and $B$, $A$ is at least as good overall as $B$ if and only if the total wellbeing of the population of $A$ in $A$ is at least as great as the total wellbeing of the population of $B$ in $B$, and the population of $A$ is the same size as the population of $B$. So, among outcomes with populations of the same size one is always either overall better than, worse than, or equally as good as another; but no such relations hold between outcomes with populations of different sizes.\footnote{This is like one part of Bader’s theory (2015). His moral theory doesn’t necessarily face the problems I’m about to describe because he countenances deontological constraints.}

Say we thought this relation reflected overall goodness, which one ought to maximise. Now we would have to decide what one ought to do when faced with a decision between incomparable outcomes; say, this decision.

\[
\begin{array}{ccc}
\text{D_{44}}: & \text{Anne} & \text{Bob} & \text{Carl} \\
A & 8 & 10 & * \\
B & 10 & 8 & 10 \\
\end{array}
\]

We have supposed that $A$ and $B$ are overall incomparable. The standard response to such incomparability in a decision, in the framework of overall goodness maximisation, is to say that the options are equally choice-worthy in $D_{44}$, and both permissible, as neither is worse overall than another available option.\footnote{An alternative would be to say that there is no fact of the matter as to which is more choice-worthy. Still, one would either have to say that they are both permissible, or claim that this is a case of genuine moral dilemma (every available option is impermissible).} In this case, the conclusion is palatable. (Indeed, it is the conclusion that I would endorse and that SU delivers). However, in other cases this same response delivers entirely unpalatable results. Consider another decision.

\[
\begin{array}{ccc}
\text{D_{45}}: & \text{Anne} & \text{Bob} & \text{Carl} \\
A & 1 & 1 & * \\
C & 10 & 8 & 10 \\
\end{array}
\]

We have supposed also that $A$ and $C$ are overall incomparable. Neither is overall worse than any other available outcome. Therefore, according to this approach we should conclude that $A$ and $C$ are both permissible in $D_{45}$. This conclusion is highly unintuitive.\footnote{This is like the unpalatable ‘greediness’ of the incomparability account of neutrality discussed in Broome 2004, Broome 2009, Rabinowicz 2009, and Cusbert and Kath 2016.} We could flesh out some cases in detail, which $D_{45}$ might represent, to highlight the problem. $D_{45}$ might be a decision about whether to brutally murder Anne and Bob, who would otherwise have a happy child, Carl.\footnote{Toby Ord suggested this to me.}
Chapter 5. Defending Shortfall Utilitarianism

The idea that outcomes with different sized populations are overall incomparable has some initial appeal. However, introducing this incomparability into the overall betterness relation and continuing to make decisions on the basis of that relation is highly problematic. This kind of problem is well-known, and provides some reason to treat overall incomparability with caution. However, these problems do not arise when overall goodness is bypassed altogether. The lesson we should learn is that making decisions wholly on the basis of a significantly incomplete overall betterness relation can be problematic. In the context of a theory that bypasses overall betterness, making decisions based on relations of betterness in respects, widespread or even universal overall incomparability does not have this kind of unpalatable implication. According to SU (and HU and NHU, among others) choice-worthiness is not closely connected to overall betterness, but rather directly determined (in one way or another) by betterness in various respects. So for example, returning to decision \(D_{45}\), although we can’t say that \(C\) is better overall than \(A\), we can say that \(C\) is more choice-worthy in \(D_{45}\) than \(A\) (because \(C\) is better in a morally relevant respect than \(A\) and \(A\) is not better in any morally relevant respect than \(C\)), and that \(C\) is permissible in \(D_{45}\) while \(A\) is not. As I said in Section 1.2, each of the relations of choice-worthiness in a particular decision can also be represented by a standardly structured betterness relation: we can read ‘\(C\) is more choice-worthy in \(D_{45}\) than \(A\)’ as ‘It is better to choose \(C\) than to choose \(A\) when faced with \(D_{45}\)’.

A final possible objection to giving up overall betterness takes a slightly different form. Instead of objecting to the absence of such a relation itself, or to problems that might arise as a result of that absence, the objection is that problems that seem to concern overall betterness might arise in a slightly different form for choice-worthiness, too. I said in Section 2.2.3 that the rejection of the overall betterness requirement means that many of the impossibility results in population axiology fail to apply to general utilitarian theories that don’t rely on an overall betterness relation, like SU. Those impossibility results show that there is no satisfactory overall betterness relation among outcomes. I accept this, and so I have not engaged with those impossibility results. One might wonder, however, whether there are versions of those impossibility results that concern choice-worthiness rather than overall betterness.

\(^{26}\)Of course I haven’t shown that introducing any overall incomparability is problematic; just that there is at least one way of doing so, which has some intuitive appeal and is occasionally endorsed in the population ethics literature, which is problematic.
5.2. MORE ON OVERALL BETTENESS

Arrhenius addresses this question in the draft of his forthcoming book, answering in the affirmative (2016). Towards the end of that draft, Arrhenius extends his ‘axiological impossibility results’ (i.e. those concerning overall betterness) to what he calls ‘normative impossibility’. Rather than considering what is possible within an overall betterness relation, Arrhenius considers what is possible within a ‘population morality’. A population morality ‘at least assigns the normative status wrong to some actions in some possible choice situations’ (2016, 376). He provides some normative versions of his ‘axiological adequacy conditions’, claims that they are just as compelling as the overall betterness versions, and proves that they are inconsistent. He aims thus to extend his conclusions: not only is there no adequate overall betterness relation among outcomes (in his terms, ‘population axiology’); there is also no adequate population morality.

SU is a population morality: it assigns the normative status wrong—I have been using the term ‘impermissible’—to some actions in some possible choice situations. In any decision, all but the outcome(s) most choice-worthy in that decision are impermissible. Arrhenius’s normative impossibility theorems thus apply to SU. I accept that those theorems are valid. It must therefore be the case that SU does not meet some of his normative adequacy conditions. Does this constitute an objection to SU? I will address this possible objection as follows. First, I’ll show that, and why, SU does not meet two of Arrhenius’s normative adequacy conditions in particular (one from each of his normative impossibility theorems). I’ll argue that SU provides a compelling justification for the falsehood of each of those conditions, and that it is therefore no objection to SU that it does not meet them. I’ll draw the lesson that translations of overall betterness adequacy conditions into normative conditions, though they may be appealing on their face, should be treated with caution.

One of Arrhenius’s conditions that is false according to SU is his ‘Normative Dominance Addition Condition’. His rough characterisation of the condition

---

27Note that Arrhenius associates what I call the overall betterness requirement with both axiology and consequentialism. Thus, I don’t accept the claim that his axiological results show that certain things are impossible in a consequentialist framework. See Section 2.2 for my broader definitions of consequentialism and axiology. To repeat my claim there: nothing in this dissertation rests on this being more than a terminological disagreement.

28Arrhenius: ‘What the above discussion shows, I think, is that as long as welfare is only taken into account in a consequentialist manner, the axiological impossibility theorems challenge the existence of an acceptable population morality’ (2016, 365–366). The discussion is about the shift from ‘axiology’ to ‘population morality’.

29I made this general assumption in Section 1.2.
is: ‘If it is wrong in a certain situation to add lives with positive welfare and increase the welfare of the rest of the population, then it is also wrong in the same situation to add no lives, other things being equal’ (2016, 378). Here is the condition more carefully stated, in my terms.

**The Normative Dominance Addition Condition**: For any decision, $D_1$, any pair of outcomes available in $D_1$, $A$ and $B$, and any pair of wellbeings, $x$ and $y$, $y > 0$: if every member of the population of $A$ has a wellbeing below $x$ in $A$; and the people who exist in $B$ consist of the same number of people as in the population of $A$ with wellbeings at or above $x$, in addition to some number of people at wellbeing $y$; and choosing $B$ in $D_1$ is impermissible; then choosing $A$ in $D_1$ is impermissible.$^{30}$

This condition is false according to SU. Here is a decision to illustrate.

$$
\begin{array}{c|cc|c}

<table>
<thead>
<tr>
<th></th>
<th>p_1</th>
<th>p_2</th>
</tr>
</thead>
</table>

D_{46}: & A & 15 & * \\
| B & 16 |
| C & 13 & 13 |

\end{array}
$$

The outcomes $A$ and $B$ in $D_{46}$ fit the description of the pair of outcomes given in the Normative Dominance Addition Condition. The population of $B$ consists of the same number of people as the population of $A$ but with higher wellbeing (in this case also in fact the very same person), in addition to an extra happy person. Choosing $B$ in $D_{46}$ is impermissible according to SU, but choosing $A$ is not impermissible. In fact, choosing $A$ is required, as it is the only permissible available outcome. This is because the shortfall of $A$ in $D_{46}$ is 1, the shortfall of $B$ in $D_{46}$ is 2, and the shortfall of $C$ in $D_{46}$ is 2.$^{31}$

That much is clear: SU does not meet the Normative Dominance Addition Condition. I want to further argue that SU is correct in this regard, and that it is not desirable that a moral theory meet the Normative Dominance Addition Condition. I make this argument on the basis of the intuitive correctness of the verdicts delivered by SU in $D_{46}$, and the justification of these verdicts provided by SU. Outcome $B$ above is just like $A$, except that the $A$-person is better off and

---

$^{30}$For Arrhenius’s formal statement see Arrhenius 2016, 379. For the overall goodness version see page 310 of that work.

$^{31}$I will discuss this decision again in the context of the benign addition argument in the next section.
an additional happy person exists. What, one might think, could make B less choice-worthy in any decision than A? The answer is: the existence of another available outcome that is better than B in some morally relevant respect, but not better than A in any morally relevant respect. It is true that A is worse in a morally relevant respect than B, and that in a decision just between those two, B would be more choice-worthy than A. However, B is worse in a morally relevant respect than another available outcome in D_{46} by a greater amount. Things tell against A by less than they tell against B in D_{46}. SU concludes that one ought to choose A when faced with D_{46}, in line with my intuitions and for good reasons.\footnote{Arrhenius considers the Normative Dominance Addition Condition the most eligible for rejection of the normative adequacy conditions (likewise for the overall betterness version), for, I think, not dissimilar reasons. He provides a normative impossibility theorem that does not appeal to it (and numerous overall betterness theorems that don’t appeal to the analogue) (2016, 387–388).}

Another of Arrhenius’s conditions that is false according to SU is his ‘Normative Non-Sadism Condition’. His rough characterisation of the condition is: ‘If it is wrong in a certain situation to add any number of lives with positive welfare, then it is also wrong in the same situation to add any number of lives with negative welfare, other things being equal’ (2016, 382). Here is the condition more carefully stated, in my terms.

The Normative Non-Sadism Condition: For any decision, D_{1}, any pair of outcomes available in D_{1}, A and B, and any pair of wellbeings, x and y, x > 0 and y < 0: if the A-and-B people have the same wellbeings in A and B, the A-not-B people have wellbeing x in A, there is at least one B-not-A person and the B-not-A people have wellbeing y in B, and it is impermissible to choose A in D_{1}, then it is impermissible to choose B in D_{1}.\footnote{For Arrhenius’s formal statement see Arrhenius 2016, 382. For the overall goodness version see pages 95–96 of that work.}

This condition is also false according to SU. Here is a decision to illustrate.

<table>
<thead>
<tr>
<th>D_{47}</th>
<th>p_1</th>
<th>p_2</th>
<th>p_3</th>
<th>p_4</th>
<th>p_5</th>
<th>p_6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>50</td>
<td>5</td>
<td>5</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>B</td>
<td>50</td>
<td>*</td>
<td>*</td>
<td>-5</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>C</td>
<td>25</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>20</td>
</tr>
</tbody>
</table>

The outcomes A and B in D_{47} fit the description of the pair of outcomes given in the Normative Non-Sadism Condition. The A-and-B person, p_1, has the same
wellbeing in A and in B. The A-not-B people, p2 and p3, have positive wellbeing in A. The B-not-A person, p4, has negative wellbeing in B. However, according to SU it is impermissible to choose A in D47 and permissible to choose B in D47. This is because the shortfall of A in D47 is 30, while the shortfall of B in D47 is 25 and the shortfall of C in D47 is 25.

Once again, I think that SU not only conflicts with the Normative Non-Sadism Condition, but also provides a compelling justification for the falsity of that condition. That justification, spelled out in terms of D47, is as follows. It is true that B is worse in a morally relevant respect than A, that A is not worse in any morally relevant respect than B, and that this means that in a decision just between A and B, A is permissible (in fact required) and B is impermissible. One might naturally wonder what could invert this permissibility structure, to make B permissible and A impermissible. (The Normative Non-Sadism Condition just asserts that nothing can.) The answer is once again: the existence of another available alternative related to A and B in certain ways. The crucial fact in D47 is that the amount by which A is worse than C in a morally relevant respect is greater than the amounts by which either of the other available outcomes is worse than another.

I find that the Normative Non-Sadism Condition has some intuitive appeal (though I don’t agree with Arrhenius that it is as compelling as the overall betterness version). However, I am convinced that SU reaches the correct conclusion in D47, and thus shows the condition to be in fact undesirable. Even setting aside the theoretically-laden justification provided by SU, I reach the same conclusion by thinking through the options in D47 at a more intuitive level. Here is how. Considering first B, it strikes me that certainly there is something unappealing about B. It counts against B that p4 exists in B wretched, and it counts further against B that there is another available outcome in which, instead of p4 existing wretched, someone else exists happy (with wellbeing 20, at best). Equally clear is that there is something unappealing about C. It counts against C that p1 exists with wellbeing 25 rather than with wellbeing 50, as she does in each other available outcome. So far, things seem to count against B and C to a similar degree. Finally, I consider A. Certainly, in the context of D47, there is also something unappealing about A. My intuitions about non-identity decisions lead me fairly directly to this conclusion. In A two people exist with low-ish happy wellbeing, while in C two different people exist with
higher happy wellbeing instead. That C is also available in D_{47} means that the
low-ish wellbeings of p_2 and p_3 in A count against choosing A. The weight
of this consideration, it seems to me, is greater than the weight against C, or
against B. So, in D_{47} one ought to choose B or C and not A.

In addition to arguing that these two normative conditions are false, I want
to suggest a diagnosis for their intuitive appeal. I attribute this appeal, the
reader may be unsurprised to learn, to the prevalence and naturalness of rea-
soning in terms of overall betterness. Arrhenius developed his normative
adequacy conditions by translating adequacy conditions for an overall better-
ness relation. However, the influence of overall betterness can be less direct
than this. Reasoning in terms of overall betterness might include things like
decision-making by elimination in pairwise comparisons: it is tempting to think
that if X is better in some respect than Y and Y is better in no respect than X, then
X is more choice-worthy in any decision than Y (and can thus be eliminated).
This condition holds if the relevant respects of betterness can be combined to
produce an overall betterness relation. If not, however, that condition does
not hold. The two conditions I have focussed on here were initially intuitively
compelling to me, as they seem to be to many others. The means by which I
was convinced of their falsehood (set out in the justifications provided above)
leads me to think that that intuitive appeal was due to some tacit appeal to
overall betterness, perhaps via elimination by pairwise comparison.

Say I am right that these two normative conditions are false, and that their
intuitive appeal can be explained in terms of the prevalence—misleading, in
these cases—of reasoning in terms of overall betterness. Where does this leave
us? What kind of response does this offer to Arrhenius’s results, and the other
similar results that may follow? It does not show that we are free to embrace the
rest of Arrhenius’s normative adequacy conditions, nor that there are not other
impossibility theorems that do not appeal to these two conditions. I think that
my arguments here fall into the second category of the three kinds of possible
response suggested by Arrhenius himself: they ‘uncover a reason for discarding
one or more of the adequacy conditions as evidence against the theories that
violate them’ 2016, 391–394.\textsuperscript{34} I have provided reasons for rejecting two of the
conditions in particular, and also a general ‘debunking’ reason for exercising

\textsuperscript{34}The other options he gives are to ‘become moral sceptics and accept that our considered moral
beliefs cannot be epistemically justified’, or to ‘try to find a way to explain away the relevance of
the theorem for moral justification’ (391).
caution regarding other intuitively appealing conditions.

5.2.2 A finer grain of outcomes

At various points throughout this dissertation, I have touched on other ways of representing theories like SU so as to retain an overall betterness relation of some kind. I don’t think that there is any very good reason to try to retain an overall betterness relation among outcomes (I’ve just tried to explain why), and I have presented moral theories in the way that I think does least violence to the concepts and terminology involved. Some may disagree with me, and find the idea that there is an overall betterness relation among outcomes so compelling that it should be maintained at the expense of other compelling ideas. They may think that it would be better to accept that the overall betterness relation among outcomes is very incomplete and not particularly closely connected to choice-worthiness. (I mentioned this possibility in footnote 22 on page 117.) Or they may prefer to accept that the overall betterness relation among outcomes is intransitive, context-dependent, or radically structured in some other way. (I mentioned the possibility that Temkin should be read in this way in the previous section.) I find each of these options less perspicuous, and more prone to create confusion, than my approach of simply giving up overall betterness in favour of betterness in various respects and choice-worthiness in decisions. However, I accept that some people may prefer one of these other ways of representing SU and similar moral theories. While I prefer my own way of doing things, this seems to me to be largely a presentational issue and as such of secondary importance.

There is one other way of representing SU and similar theories in terms of an overall betterness relation, which I have so far not addressed and which I find quite appealing. The idea is that although SU does not deliver an overall betterness relation among outcomes as I have so far characterised them, it does deliver an overall betterness relation among outcomes distinguished on a finer grain. Though we can’t say that $A$ is better overall than $B$, we can say that $A$-rather-than-$B$ (or perhaps $A$-in-$D_1$) is better overall than $B$-rather-than-$A$ (or $B$-in-$D_1$).\footnote{This is a possibility that has been touched on in various places, and spelled out as an alternative to Temkin’s apparent ‘intransitivity’ of betterness in Voorhoeve 2013. Thanks to Stefan Riedener for a very useful discussion of this possibility.} I’ll set it out my preferred version of this idea here, and defend it
from some possible objections.

Here is the proposed alternative representation of SU. The relations of betterness in respects among outcomes—A, B, C, and so on—remain the same, as do the definitions of intersectional and complementary worseness, total worseness, and shortfalls. In addition, we posit a relation of overall betterness among outcome-in-decisions. The stock of outcome-in-decisions are as follows. For any decision, $D_1$, and for any available outcome in $D_1$, $A$, there is the outcome-in-decision $A$-in-$D_1$. These are the only outcome-in-decisions. The overall betterness relation orders outcome-in-decisions on the basis of shortfalls in this way: for any pair of outcome-in-decisions, $A$-in-$D_1$ and $B$-in-$D_2$, $A$-in-$D_1$ is at least as good overall as $B$-in-$D_2$, just in case the shortfall of $A$ in $D_1$ is at most as high as the shortfall of $B$ in $D_2$.

This replicates the choice-worthiness orderings of SU in every decision. Wherever SU as I have presented it says that $A$ is at least as choice-worthy in $D_1$ as $B$, it says according to this other presentation that $A$-in-$D_1$ is at least as good overall as $B$-in-$D_1$. By focussing on a finer grain of outcome, we are able to represent the choice-worthiness output of SU within a single betterness relation. This fine-graining approach also makes it particularly evident that SU provides a plausible way of making inter-decision moral comparisons. We often make this kind of comparison, saying things like ‘What she did was wrong, but not as wrong as what he did’ (where ‘she’ and ‘he’ faced different decisions). It is an appealing aspect of the current proposal that it brings this kind of comparison to the fore (though we can endorse those comparisons without endorsing the fine-graining).

Consider for example the following decisions, faced by two prospective parents (or scientists), Paul and Quinn. Paul faces $D_{48}$, between creating Anne very happy, creating Anne barely happy, and not creating Anne at all. Quinn faces $D_{49}$, between creating Bob very happy, creating Bob moderately happy, and not creating Bob at all.

<table>
<thead>
<tr>
<th>$D_{48}$:</th>
<th>Anne</th>
<th>$D_{49}$:</th>
<th>Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>9</td>
<td>D</td>
<td>9</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>E</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>*</td>
<td>F</td>
<td>*</td>
</tr>
</tbody>
</table>

On the current proposal we can say that according to SU $A$-in-$D_{48}$ is equally as

\[^{36}\text{So there is no } B\text{-in-}D_2 \text{ if } D_2 = \{A, C, D\}, \text{ for example.}\]
good overall as C-in-D_{48}, which are both better overall than B-in-D_{48}. (This is just a different way of representing the claim that A and C are equally choice-worthy in D_{48}, and both more choice-worthy in D_{48} than B.) Likewise, we can say that D-in-D_{49} is equally as good overall as F-in-D_{49}, which are both better overall than E-in-D_{49}. These claims capture the idea that Paul ought to choose A or C, and ought not to choose B, and Quinn ought to choose D or F and ought not to choose E. We can also, though, say that A-in-D_{48} is equally as good overall as F-in-D_{49} (the shortfall of each outcome in the relevant decision is 0), and that B-in-D_{48} is worse overall than E-in-D_{49} (the shortfall of B in D_{48} is higher than the shortfall of E in D_{49}). This seems plausible to me. I have the intuition that if Paul does wrong and chooses B, he does more wrong than Quinn does if she chooses E. Quinn makes it that case that Bob lives a life only a little worse than he might have lived, whereas Paul makes it the case that Anne lives a life much worse than she might have lived.

The menu-dependence of choice-worthiness is no problem for this kind of fine-grained overall betterness. Consider D_{14} and D_{17} again.

<table>
<thead>
<tr>
<th></th>
<th>p_1</th>
<th>p_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>15</td>
<td>*</td>
</tr>
<tr>
<td>B</td>
<td>15</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>p_1</th>
<th>p_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>15</td>
<td>*</td>
</tr>
<tr>
<td>B</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>C</td>
<td>15</td>
<td>17</td>
</tr>
</tbody>
</table>

I previously explained that SU cannot be represented in terms of an overall betterness relation among outcomes because this would require it to be the case both that B is equally as good as A (in virtue of the choice-worthiness output for D_{14}), and that B is overall worse than A (in virtue of the choice-worthiness output for D_{17}). This violates the standard structure of betterness relations. This problem does not arise for the overall betterness of fine-grained outcomes. We can represent SU as saying that B-in-D_{14} is equally as good overall as A-in-D_{14}, and that B-in-D_{17} is worse overall than A-in-D_{17}. This is perfectly consistent with the standard structure of betterness relations.

I find this way of representing moral theories (including SU), in terms of an overall betterness relation among fine-grained outcomes, somewhat appealing. The main objection I have encountered against it is a sense that outcomes-in-decisions are not the appropriate object of moral concern. I initially found this objection to have some force, but have come to think that it may be misguided. Let me explain the objection as I understand it, and then offer my response.
At least for consequentialists, outcomes—possible worlds—seem to be the appropriate objects of moral concern. (This motivated the way I framed my project in Section 2.2.) We think that what matters for the purposes of the moral assessment of people’s possible actions is what the world will be like if those actions are performed. At first glance, it seems as though the proposed fine-graining of outcomes constitutes a move away from this focus on possible worlds to some other, mysteriously complex kind of entity. Rather than evaluating available acts by reference to their consequences, in the intuitively compelling sense of how they make the world ‘go’, the suggestion seems to be to evaluate available acts by reference to these curious ‘outcome-in-decisions’. It’s not obvious what kind of thing outcome-in-decisions are—some kind of possible-world/decision hybrid?—but they don’t seem to be a natural interpretation of ‘consequences’. So the fine-graining proposal seems to be a shift away from consequentialism; away from the idea that only outcomes matter, to the idea that both outcomes and decision context matter. The shift seems somewhat mysterious, poorly motivated, or ad-hoc.

I suggest that this is not the best way to understand the fine-graining proposal. Rather, we should understand the proposal as identifying previously overlooked, morally relevant modal features of possible worlds. The objects of moral concern according to this proposal are, as we always thought they were, possible worlds. The re-individuation of those possible worlds, to produce a finer grain of outcomes, reflects a recognition of the modal properties of those worlds. On this understanding, an outcome-in-decision is not a mysterious entity at all. It is a possible world: with a population of people with particular wellbeings, and modally connected to other possible worlds in particular ways. Take $B$-in-$D_{17}$, for example. This is a possible world in which only $p_1$ and $p_2$ exist, with wellbeings 15 and 13 respectively. It is also a possible world in which some person at some time faces $D_{17}$. This means that it is a world in which it was possible (in the relevant sense, at the relevant time) for that person to choose not to create $p_2$, or to create her with wellbeing 17 instead of 13. The proposal is that this kind of feature of possible worlds—which other possible worlds are available to agents in them—is morally relevant.

Imagine I face $D_{17}$ and choose $B$. Then, the actual world is a world in which I face $D_{17}$ and choose $B$. It is a world in which $p_2$ exists with wellbeing 13, and it is a world in which it is at some time possible that $p_2$ will not exist, and possible
that \( p_2 \) will exist with wellbeing 17. It is a world in which someone (me) chooses to create \( p_2 \) with wellbeing 13 rather than not creating \( p_2 \), or creating \( p_2 \) better off. These are all features of the actual world in which I face \( D_{17} \) and choose \( B \). On the other hand, imagine I face \( D_{14} \) and choose \( B \). Then, the actual world is a world in which I face \( D_{14} \) and choose \( B \). This is another world in which \( p_2 \) exists with wellbeing 13, and in which it is at some time possible that \( p_2 \) will not exist; but it is not a world in which it is at any time possible that \( p_2 \) will exist with wellbeing 17. It is a world in which I choose to create \( p_2 \) with wellbeing 13 when the only other possibility is that I don’t create \( p_2 \) at all. \( B \)-in-\( D_{17} \) and \( B \)-in-\( D_{14} \) are two possible worlds that share some features (populations and wellbeings) but differ in their modal features. This might justify the proposed focus on outcomes of a finer grain by moral theories. \( B \)-in-\( D_{17} \) is overall worse than \( B \)-in-\( D_{14} \) because \( B \)-in-\( D_{17} \) is a possible world in which it is (at some time) possible for \( p_2 \) to have higher wellbeing than she in fact does, whereas this is not possible in \( B \)-in-\( D_{14} \).

It seems fairly plausible to me that this kind of modal feature of a world might be morally significant. On this interpretation, the fine-graining proposal does not constitute a move away from consequentialism, or from the focus on possible worlds. I also don’t think it constitutes a move away from welfarism. One might at first think this: according to the current proposal, wellbeings are not the only things that matter for the moral evaluation of possible worlds (outcomes); modal features of those worlds also matter. Strictly I suppose this is true. However, the relevant modal features all themselves concern wellbeings. One might say that the focus shifts from only actual wellbeings, to include also possible wellbeings. This seems to me to maintain the spirit of welfarism.\(^{37}\)

To summarise this section: my preferred way of representing and understanding SU (and other, similar moral theories) is as a theory according to which there is no single overall betterness relation among outcomes, and betterness in various respects determine the choice-worthiness of outcomes in decisions directly. This is consistent with thinking of each choice-worthiness relation as a relation of overall betterness in a decision—what it would be morally better

\(^{37}\)This might raise some doubts about my claim that this is simply a representational matter. The difference between concern with only actual wellbeings and concern also with possible wellbeings may seem more substantive. I don’t think that it really is. Rather, I think that the fine-graining approach once again brings to the fore something that was there all along. Our concern always included merely possible wellbeings (which is why we rejected actualist or necessitarian versions of utilitarianism; see page 123).
5.3 Some conclusions

In this section I consider the conclusions of SU in four kinds of variable population decision. First are decisions concerning human extinction. According to SU, one is not obliged to prevent or delay the time of human extinction, for its own sake. This is a conclusion that some people find counterintuitive, though I am not among them. I find this conclusion perfectly intuitive, and consider it a straightforward extension of my intuition of neutrality. The second kind of decision is somewhat complex, and the conclusions of SU here are to my mind somewhat troubling. I defend those conclusions, and also consider some possible changes to SU in light of this kind of decision. I have a similar response to the third kind of decision, which constitutes a version of the benign addition argument (which I mentioned in Section 4.2). Finally, I consider the conclusions of SU in some decisions concerning the creation of wretched people. I think that even if some of these conclusions are initially unsettling, they are well justified.

5.3.1 Extinction

Some people have the intuition that the extinction of people would be a bad thing in itself, and something that ought to be avoided (at least other things being equal). SU does not vindicate this intuition. Consider the following case. Everyone who now exists is sterile. If I do nothing, no person who doesn’t already exist will ever come to exist and humans will, within a lifetime, go extinct. The alternative is to release a gas that will restore fertility (at least to some degree). The wellbeings of the inevitable people (which in this case is just those who already exist) will be the same whether we release the gas or not. Any ‘benefits’ they receive from the restoration of fertility are cancelled out by ‘harmful’ side-effects of the released gas. The wellbeings of the contingent
people (the merely possible future people) will be positive; say, the same as the inevitable people’s. Ought I to release the gas or not? Imagine this is the decision.

<table>
<thead>
<tr>
<th>$D_{50}$</th>
<th>$n$ people</th>
<th>$m$ people</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>$30,30,\ldots,30$</td>
<td>$<em>,</em>,\ldots,*$</td>
</tr>
<tr>
<td>$B$</td>
<td>$30,30,\ldots,30$</td>
<td>$30,30,\ldots,30$</td>
</tr>
</tbody>
</table>

Some people have the intuition that releasing the gas is not only morally permissible in this case, but morally required; that $B$ is more choice-worthy in $D_{50}$ than $A$.

According to SU, releasing the gas is permissible, but not required: $A$ and $B$ are equally choice-worthy in $D_{50}$. Furthermore, any difference in the wellbeings of the inevitable people in a case like this is decisive, according to SU. In a slightly different case in which releasing the gas would increase the wellbeings of the inevitable people, its release would be obligatory. In a slightly different case in which releasing the gas would decrease the wellbeings of the inevitable people, its release would be impermissible. My intuitions agree with SU on all of these points, and so I don’t feel any pressure to adjust SU in light of this decision. I’m not sure how many people will have the intuition that I ought to release the fertility-restoring gas. I’m sure some will. I don’t, and I have encountered in conversation many people who also don’t. The only response I can offer is to provide some further motivation for my intuitions and some possible partial (‘debunking’) explanations of the conflicting intuition.

I have deliberately presented a slightly different case from one often used in discussions of extinction. In the case often discussed, the continued fertility of existing people is the ‘default’ (what will happen if I do nothing), and the decision faced is whether to release a gas that will harmlessly sterilise everyone.\(^{38}\) I suspect that this alternative setup might skew our intuitions towards anti-extinction. For one thing, we are asked to imagine that immediate universal sterility will not negatively affect people’s wellbeings. This is extremely difficult to imagine. I assume that it is uncontroversial that immediate universal sterility would, in fact, make billions of people worse off in many ways. Living through the decades lead up to the extinction of humans would be very unpleasant.\(^{39}\) We might try to imagine significant compensatory ‘benefits’, to

---

\(^{38}\text{McMahan in conversation, and Ng 1983 (165).}\)

\(^{39}\text{I can’t overstate this. Most people would find the process emotionally traumatic. It would also}\)
outweigh these negative affects, but quite frankly I think that is very difficult to do. What kind of thing could outweigh living through the decades leading up to the extinction of humans?

In the case I have presented, we don’t have to imagine a harmless universal sterilisation. We only have to compare the unpleasantness of living through the extinction of humans with some other, equally unpleasant prospect. That a gas might have significant negative side effects is not at all difficult to imagine. My case also differs from this other case, in that universal sterilisation is the default and we have to perform some act (release the second gas) to restore fertility. It is very plausible, I think, that people intuitively shy away from the prospect of performing an act that will affect everyone on earth (even if we try to stipulate that it will not affect their wellbeings). It is hard to imagine that this would not carry unforeseen risks, and intuitive considerations of autonomy might also tell against this. Though we are in theory supposed to be excluding these considerations from our intuitive assessment of the case (thing are ‘otherwise equal’), this is a case in which I doubt that we can do so effectively.

My response to this decision is largely the same as my response to a decision I discussed in the context of the neutrality intuition in Section 3.2—the decision of whether or not to create some happy people on a distant planet. I have a strong intuition of neutrality: that creating a happy person is not more choice-worthy than not creating her, other things being equal. Nor is creating two happy people more choice-worthy (in a decision between the two) than not creating them; nor three people, nor four, nor a trillion. The kind of act involved does not affect my intuition. I have it whether I could ‘create’ these people by bearing a child, by pressing a button, or by releasing a gas. Nor does what else happens in the world affect my intuition. I have it whether the rest of the population is large or small; whether they are spread out through time or not.

A reminder of two points that I made regarding the neutrality intuition and the distant planet decision, which apply also to the extinction decision. First, a negative attitude to the extinction of humanity should not necessarily be interpreted as a moral intuition, to the effect that one ought to prevent extinction if possible. To the extent that I have such a negative attitude, I think that it is a non-moral one; it concerns something that I value in some be physically awful. The last generation would have no-one to care for them in old age and illness. Many would likely starve to death, I suppose. In all likelihood the best way ‘through’ would be early suicide.
sense, but which I do not think is directly morally relevant (like jam and beauty and various other things). The second point is that even a moral intuition should not necessarily be interpreted as a utilitarian intuition; as something we should want a general utilitarian theory to capture. I am especially confident that any negative attitude I have towards the extinction of humans is not a utilitarian intuition. My utilitarian intuition is that creating happy people does not constitute making people better off, in the relevant sense. Some people think that there is more to morality than utilitarian concerns. I suspect that there isn’t, but accept that this is a reasonable thing to disagree about.\footnote{My project, remember, is to find the best utilitarian component of morality—see Section 2.1.} And if one were convinced that the extinction of people has moral disvalue in itself, I think that this would provide a reason to think that utilitarian concerns are not all there is to morality.

Finally, here is one thought that relates to intuitions about extinction decisions (and possibly to other neutrality-related decisions too). It is hardly surprising that people would have anti-extinction (and pro-procreation) intuitions. Each of us is the product of billions of years of evolution, which can be described as a process that favours life-forms that have more descendants. Our desire to have descendants, and to favour the continued survival of our species, is very plausibly a product of this process. This kind of debunking argument doesn’t prove that our anti-extinction intuitions conflict with morality. But it does give us reason to suspect them, I think, if they conflict with the rest of our moral theory.

### 5.3.2 McDermott Cases

Here is an interesting decision suggested to me by Michael McDermott (in correspondence). I’ve included in the table the shortfall of each available outcome in the decision.\footnote{There are very similar cases, where instead of the empty population we have some population of inevitable people, whose wellbeings are the same on each of the five outcomes. According to SU (I agree), these cases are to be treated in the same way as the present case. If you think that there is a morally important difference between these cases (for example you think that the existence of humanity is morally significant) I would ask you to consider one of these related cases, rather than the case I have given. The issues I want to address here do not have anything to do with the emptiness of one outcome.}
Here is how we assess this decision according to SU. The shortfall of each of $A$, $C$, and $D$ in $D_{S1}$ is 0 (each of these outcomes is intersectionally and complementarily worse than each other available outcome in the decision by 0). The shortfall of $B$ in $D_{S1}$ is 5 ($B$ is complementarily worse than $C$ by 5). And the shortfall of $E$ in $D_{S1}$ is 4 ($E$ is intersectionally worse than $B$ by 4, as well as intersectionally worse than $D$ by 4). So $A$, $C$, and $D$ are all equally choice-worthy in $D_{S1}$, $E$ is less choice-worthy in $D_{S1}$ than both, and $B$ is least choice-worthy in $D_{S1}$.

This choice-worthiness ordering strikes some people as unintuitive. In an email, McDermott describes his view. Note that I have replaced McDermott’s use of ‘better’ and ‘worse’ here with ‘more choice-worthy’ and ‘less choice-worthy’. I think that this represents his views.

I can sort of see the intuition that $C$ is [more choice-worthy] than $B$. But it seems unintuitive to say that $B$ is [less choice-worthy] than $A$ ($p_1$ has no complaints). And it seems unintuitive to say that $D$ is [more choice-worthy] than $B$. And it seems especially unintuitive to say that $E$ is [more choice-worthy] than $B$.

Let me address each of these claims produced by SU in turn.

That $C$ is more choice-worthy in $D_{S1}$ than $B$: McDermott says that he can ‘sort of see’ the intuition that this is the case. McDermott favours a theory much like NHU, and has fairly person-affecting intuitions. For those of us who find non-identity decisions to be compelling counter-cases to NHU, this claim will not seem strange at all (see Sections 3.3 and 4.3.3). In a decision just between $B$ and $C$ I have a strong intuition that $C$ is more choice-worthy than $B$, so the fact that this is so in a larger decision is not concerning (which is not to say that this should be the case in every decision in which $B$ and $C$ are available). The reason $C$ is more choice-worthy in $D_{S1}$ than $B$ is that $B$ is worse than $C$ in some morally relevant respect. According to SU, it is worse complementarily,
That B is less choice-worthy in D \text{51} than A: B is less choice-worthy in D \text{51} than A despite the fact that B is not worse than A in any morally relevant respect; and in virtue of the fact that B is worse in some morally relevant respect than another available outcome, while A is not worse in any morally relevant respect than any available outcome. Again, I think that this conclusion will not seem strange to anyone who shares my intuitions about non-identity cases. In the decision consisting of just A, B and C, my intuition is that A and C are permissible and B is not; B is less choice-worthy in that decision than A and (as I just discussed) C.

That D is more choice-worthy in D \text{51} than B: Here is where I start to agree with McDermott, at least as far as intuitions go. D is just B with an extra happy person. The conclusion that D is more choice-worthy in D \text{51} than B smacks of non-neutrality about happy people. Why should I, instead of creating just p_1 at 5, create p_1 at 5 and p_2 at 10? However, when I think through the process by which SU reaches this conclusion, I find that process fairly compelling as a justification. Here it is. B is worse than C in a morally relevant respect. On the other hand, D is not worse than C, nor worse than any available outcome, in any morally relevant respect. Set aside E and A for the moment. The explanation, I think, can be given with reference to the comparisons between B, C, and D. B compares disfavourably to C. To choose B rather than C would be to choose to create someone less happy instead of someone (else) more happy. To choose D rather than C would be to choose to create someone less happy in addition to someone more happy. The comparison between B and D seems neutral in both directions: to choose B instead of D would be to create one happy person instead of that same person equally happy and another happy person (and vice versa). Likewise for the comparison between C and D. The difference in how B and D compare to C is decisive.

That E is more choice-worthy in D \text{51} than B: I agree that this is the most unsettling of SU’s conclusions in this decision. Here is the explanation of this conclusion according to SU. There is a reason not to choose E in this decision. E is worse than B. It is intersectionally worse (worse with respect to (p_1, p_1)) by 4. There is also, as we have seen, a reason not to choose B. B is worse than C. It is complementarily worse (worse with respect to (p_1, p_1)) by 5. These are the only two morally relevant worseness relations among the available outcomes, and
so they alone determine choice-worthiness in D49. The reason not to choose B is stronger than the reason not to choose E.

This explanation makes sense to me, and yet it does not fully dispel my unease about the conclusion (as the explanation in the case of the previous conclusion did). I think that two things explain why the conclusions of SU are unsettling here. First, not only is B not worse than E in any morally relevant respect, but E is worse than B in such a respect (according to SU it is intersectionally worse with respect to p1). To choose E instead of B would be to choose to create one person less happy and a second happy person, rather than just creating the one person more happy. This means that, in a decision just between B and E, B is more choice-worthy than E according to SU: this is the reversal of the conclusion reached in the larger decision we are considering. I argued above that it is not clear that this kind of menu-dependence of choice-worthiness is problematic. However, it is understandable that it may sometimes be unintuitive (though in other cases, particularly relating to neutrality, we have seen that it is highly intuitive).

Second, what tells against E in this decision is intersectional worseness (and the special case of intersectional worseness that coincides with worseness for a particular person), while what tells against B is complementary worseness. People are very practiced at recognising and responding to something like intersectional worseness (and especially the special case of worseness for a particular person), and less practiced at recognising or responding to complementary worseness. It is not entirely surprising if intuitions side with intersectional worseness rather than complementary worseness where the two conflict, even for people without particularly strong person-affecting intuitions. One needs to genuinely embrace the moral significance of complementary worseness in order to accept that the complementary worseness of B (than C) outweighs the complementary worseness of E (than B), to make it the case that B is less choice-worthy in D51 than E.

These conclusions would not be so troubling if they only affected minor placings in the choice-worthiness ordering. SU says that in D51 both B and E are impermissible; both are significantly worse than the other available options. The troubling element is only that B is a slightly less choice-worthy impermissible option than E. However, as McDermott points out, there is a variation of this decision in which similar factors do make the difference between permissibility
and impermissibility.

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>Shortfall</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>*</td>
<td>*</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>$B$</td>
<td>5</td>
<td>*</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>$C$</td>
<td>*</td>
<td>10</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>$D$</td>
<td>5</td>
<td>10</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>$E$</td>
<td>1</td>
<td>10</td>
<td>9</td>
<td>4</td>
</tr>
</tbody>
</table>

$D_{52}$ is like $D_{51}$, except that there is an extra inevitable person. The inevitable person is worse off in those three outcomes which were permissible in the previous case ($A$, $C$, and $D$), better off in $B$ and $C$. This means that according to SU, in $D_{52}$ $E$ is permissible (along with $A$, $C$, and $D$) and $B$ is impermissible. (We could alter the case in a similar way to make the other three options impermissible too, and only $E$ permissible.) Now, it is permissible to bring about $E$ and impermissible to bring about $B$. The considerations telling against each of these options are as before—only the remaining three options have been worsened and made less choice-worthy.

It does seem unintuitive to me that it is permissible to choose $E$ and impermissible to choose $B$ in this decision. However, I am prepared to accept this conclusion—partly, I think, on the basis of the explanation provided by SU, and partly because it seems less troubling than implications of each of the alternatives to SU I have encountered. This decision is complex, and in general I am inclined to trust intuitions about complex cases less than intuitions about simpler cases. So I am not inclined to avoid this kind of conclusion at the cost of any of those intuitions set out in Chapter 3. If this is an implication of any asymmetric theory that avoids repugnance and the non-identity problem, then I would accept it as a moral discovery. However, I have no reason to think that this is the case. SU is one such theory. It is possible that a thorough exploration of the space of general utilitarian theories will reveal a better one. Until that time, I think the best approach is to accept that in this decision $E$ is permissible and $B$ is not.

### 5.3.3 Benign addition

In Section 4.3.5 I discussed the mere addition argument, an argument that is supposed to lead to a repugnant conclusion (or to show that avoiding repugnant
conclusions is difficult). I showed that according to SU the argument is either
invalid (if it is made in terms of choice-worthiness) or unsound (if made in
terms of overall betterness). I also mentioned, and postponed discussion of,
the similar benign addition argument. I return to the benign addition argument
now. Focussing on this argument reveals something interesting about SU: that
in some complex decisions, an outcome with a larger less happy population
is more choice-worthy than an outcome with a smaller happier population.
I’ll examine why this is the case according to SU, and what the limits to this
phenomenon are; and I’ll argue that this limited amount of ‘repugnance’ is
acceptable.

Here are three decisions.

<table>
<thead>
<tr>
<th>Decision</th>
<th>P1</th>
<th>P2</th>
<th>Decision</th>
<th>P1</th>
<th>P2</th>
<th>Decision</th>
<th>P1</th>
<th>P2</th>
</tr>
</thead>
<tbody>
<tr>
<td>D_{53}</td>
<td>A</td>
<td>15</td>
<td>+</td>
<td>D_{54}</td>
<td>A'</td>
<td>16</td>
<td>10</td>
<td>D_{55}</td>
</tr>
<tr>
<td>A'</td>
<td>16</td>
<td>10</td>
<td>B</td>
<td>14.5</td>
<td>14.5</td>
<td>B</td>
<td>14.5</td>
<td>14.5</td>
</tr>
</tbody>
</table>

And here is how the argument goes. The first decision is the ‘benign addition’
decision. One of the available outcomes contains a person with a particular
wellbeing. The other available outcome contains that same person with a higher
wellbeing as well as an additional happy person. According to SU, A' is more
choice-worthy in D_{53} than A. We’re neutral about creating a happy person, so
the fact that p1’s wellbeing is higher in A' than in A is decisive. The second
decision is a fixed population ‘redistribution’ decision. According to SU (and
as dictated by CFPU), B is more choice-worthy in D_{54} than A'. The average
wellbeing in A' is 13, which is lower than the average wellbeing in B. The third
decision is a (mildly) repugnant one. SU says that A is more choice-worthy in
D_{55} than B. B is intersectionally worse than A, and contains an extra happy
person (about the latter we are neutral).

These are very similar conclusions to the ones delivered by SU in the mere
addition argument. We saw there that they are inconsistent in the framework of
overall betterness, but not in the framework of choice-worthiness determined
by betterness in respects. We have no problem explaining axiologically why
A' is not less choice-worthy in D_{53} than A, B is more choice-worthy in D_{54} than
A', and A is more choice-worthy in D_{55} than B.

In mere addition cases SU also avoids any degree of repugnance in the
relevant three-way decisions. (See page 130.) In benign addition cases things
are not so simple. Consider the three-way decision between all of the outcomes
CHAPTER 5. DEFENDING SHORTFALL UTILITARIANISM

This is one of a limited range of cases in which SU says that an outcome with a larger, less happy population is more choice-worthy than an outcome with a smaller, happier population. Here is how SU reaches that conclusion. The shortfall of $A$ in $D_{56}$ is 1 ($A$ is intersectionally worse than $A^+$ by 1). The shortfall of $A^+$ in $D_{56}$ is 3 ($A^+$ is intersectionally worse than $B$ by 3). The shortfall of $B$ in $D_{56}$ is 0.5 ($B$ is intersectionally worse than $A$ by 0.5). Thus, $B$ is more choice-worthy in $D_{56}$ than $A$, which in turn is more choice-worthy in $D_{56}$ than $A^+$.

$A$ is more choice-worthy than $B$ in a decision just between the two, and making $A^+$ available reverses this order. It does this because $A^+$ is better than $A$ in some morally relevant respect (and thus increases the shortfall of $A$), but not better than $B$ in any morally relevant respect (in fact it is worse than $B$ in a morally relevant respect). One might find this kind of menu dependence unintuitive in itself. I addressed potential concerns about menu-dependence in Section 5.1. Here I want to try to set aside the issue of menu-dependence and address the conclusions of SU in this three-way decision directly. (It may, of course, not be possible to effectively quarantine our intuitions about particular decisions from the influence of qualms about menu-dependence. We can try, however.)

This kind of conclusion occurs only in a fairly special kind of decision, and not in all three-way mere addition decisions. We encountered in the previous section a decision in which it does not occur. Recall $D_{46}$.

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>Shortfall</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{56}$: $A$</td>
<td>15</td>
<td>*</td>
<td>1</td>
</tr>
<tr>
<td>$A^+$</td>
<td>16</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>$B$</td>
<td>14.5</td>
<td>14.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

$A$ is more choice-worthy in $D_{46}$ than $C$ because the shortfall of $A$ in $D_{46}$ is 1, whereas the shortfall of $C$ in $D_{46}$ is 2. (The shortfall of $B$ is also 2, which makes it equally as choice-worthy in $D_{46}$ as $C$.)
5.3. SOME CONCLUSIONS

So, what are the conditions under which SU produces this kind of conclusion, in these benign addition cases? Consider this larger decision (again with the shortfall of each outcome included in the table).

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
<th>Shortfall</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>15</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>1</td>
</tr>
<tr>
<td>$A^+$</td>
<td>16</td>
<td>10</td>
<td>*</td>
<td>*</td>
<td>3</td>
</tr>
<tr>
<td>$B$</td>
<td>14.5</td>
<td>14.5</td>
<td>*</td>
<td>*</td>
<td>0.5</td>
</tr>
<tr>
<td>$B^+$</td>
<td>14.5</td>
<td>14.5</td>
<td>7</td>
<td>*</td>
<td>6</td>
</tr>
<tr>
<td>$C$</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>*</td>
<td>1</td>
</tr>
<tr>
<td>$C^+$</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>$D$</td>
<td>13.5</td>
<td>13.5</td>
<td>13.5</td>
<td>13.5</td>
<td>2</td>
</tr>
</tbody>
</table>

In this case the slide towards larger, less happy populations stops when we get to wellbeing 14. This amounts to a ‘reduction’ of 1 unit of wellbeing for the inevitable person, $p_1$ (from her maximum available wellbeing). Outcome $C$, with a large population of people at wellbeing 14, is worse than $A$ by 1. It is also worse than $B$ by 1. Its shortfall in $D_{57}$ is the same as $A$’s, which means that $A$ and $C$ are equally choice-worthy in $D_{57}$. $D$, with an even larger population of people at wellbeing 13.5, has a higher shortfall and so is less choice-worthy in $D_{57}$ than $A$. In general, I think that we can see that the allowable (and the required) decrease in wellbeing will depend on just how benign the ‘initial’ addition is. Here, the decrease is capped at one unit of wellbeing, which is the amount by which $p_1$’s wellbeing is increased in $A^+$ over $A$. (Of course, for all I’ve said one unit of wellbeing might be highly significant, so I don’t claim that this is unimportant. In any case, we can change the numbers in this case to make the differences involved arbitrarily large.)

I am prepared to accept these implications of SU as somewhat surprising moral discoveries. In this dissertation, because of the nature of my core utilitarian intuitions, I have focussed largely on two-way variable population decisions. My intuitions about more complex decisions are much weaker. I also have less faith in them, as I suspect they may be particularly susceptible to influence by tacit thought patterns that assume an overall betterness relation (for example, pairwise elimination of alternatives). I think that we should expect to be surprised when we begin to apply respect-based decision-making procedures to these more complex decisions. However, I also think it not unlikely that further investigation of such decisions might lead us to revise the
way we use respects of betterness to make decisions (i.e. revise the shortfall minimising approach), or even what the respects of betterness are. For this reason, my acceptance of the implications of SU in decisions between more than two alternatives is somewhat provisional.

5.3.4 Wretchedness: too significant or not significant enough?

Lastly, I want to consider two kinds of decision concerning the creation of wretched people. One is a kind of decision in which wretchedness appears highly morally significant, according to SU. The other is a kind of decision in which the moral significance of wretchedness according to SU appears relatively small.\(^{42}\) I am aware that some people find the conclusions of SU in one or the other of these kinds of decision (or even both) unintuitive. I am not among them, and will defend SU’s conclusions in both kinds of decision.

According to SU, one ought not create a group of people, however large, if any of them will be wretched (other things being equal, in a decision between the two). Here is a decision to illustrate.

\[
D_{58}: \begin{array}{ccc}
  & x \text{ people} & y \text{ people} \\
 A & 50, 50, \ldots, 50 & \ast, \ast, \ldots, \ast \\
 B & 50, 50, \ldots, 50 & -20, 50, 50, \ldots, 50
\end{array}
\]

This is a decision between creating a group of people (\(y\) of them), and not creating that group of people. (I include the \(x\) inevitable people just to show that things are otherwise equal, and to avoid this being a decision about whether people should exist at all.) If created, almost all of the people will be happy. One of them, however, will be wretched. According to SU \(B\) is complementarily worse than \(A\) by 20, in virtue of the existence in \(B\) of this one wretched person. This is the decisive fact in \(D_{58}\): \(A\) is more choice-worthy in \(D_{58}\) than \(B\). One ought not create the \(y\) possible people if faced with this decision.

Some people find this conclusion unintuitive.\(^{43}\) I don’t. I don’t find it unintuitive, but nor do I find particularly unintuitive the conflicting claim that it is permissible to create a large number of happy people and one wretched person (that \(B\) is permissible in \(D_{58}\), for example). I don’t have strong direct intuitions about this kind of decision. For this reason I found it difficult, in

---

\(^{42}\)This contrast indicates that we cannot simply assess the significance of wretchedness according to a particular theory, across the board. This should become clearer as we go on.

\(^{43}\)Sikora (1978), for one.
developing SU, to decide how it should deal with wretchedness. On the one hand, I think that some people will find this conclusion unintuitive, and that carries some weight. Even if I don’t have the intuition myself, if many other people do perhaps there is some reason to try to respect the intuition. On the other hand, in the absence of a strong intuition about this kind of case I am inclined to trust to compelling general principles, or consistency with the strong intuitions that I do have. And here these seem to favour the conclusion given by SU. Speaking loosely, if it is worse that someone exists wretched than that they don’t exist at all, and it is neutral that someone exists happy instead of not existing at all, then it should be on balance worse that someone exists wretched and someone exists happy than that neither of them exist at all. The badness of a wretched life should not be ‘soaked up’ by merely neutral factors.\footnote{Thanks to Melinda Roberts for useful discussion on this, and putting this latter view forcefully.}

So, SU seems to me to be right on this issue. For the sake of those who disagree with me, I shall mention a possible alteration to the definition of complementary worseness. As it stands, SU says that we should treat wretched wellbeings and happy wellbeings differently in a complementary comparison. Roughly, we are interested in the sum of the wretched wellbeings and the average of the happy wellbeings (really the average of the \textit{adjusted} wellbeings—see the definitions in Section 4.1.3). Instead, and to make wretchedness less ‘weighty’ in decisions like this, we might simply compare the average of all wellbeings, wretched and happy, in each of the relevant complementary groups. So when comparing $A$ and $B$ in $D_{58}$ above, we would focus on the average wellbeing among all of the $B$-not-$A$ people in $B$ (something less than 50 but still positive) and conclude, contrary to SU, that $B$ is complementarily worse than $A$ by 0. The two outcomes would then be equally choice-worthy in $D_{58}$. I considered this option, but in the end decided that I prefer SU as it stands. I find it difficult to explain why it might be that the badness of wretched lives could be ‘soaked up’ or neutralised in this way by the existence of happy lives, which after all are themselves merely neutral.

The other kind of decision I want to address is a kind of non-identity decision, between creating one group of all happy people, or a different group of people, some happy and some wretched. Here is a decision of the kind I have in mind.\footnote{Thanks to Hilary Greaves for pointing out this kind of decision. It is a non-identity version of the ‘sadistic conclusion’ (see Arrhenius 2000). In this decision there are no inevitable people. This}
In outcome A five people exist with wellbeing 10 and five people exist with wretched lives, at wellbeing −2. In B five people exist with wellbeing 10 and twenty people exist with wellbeing 2. Some people have the intuition that B is more choice-worthy in D59 than A. They may reason as follows. A and B are very similar in one respect: in each of them, five people exist with wellbeing 10. Granted they are a different five people in one outcome than in the other; still, it is tempting to match up and disregard these parts of each outcome, and focus on what seems to distinguish A and B. In A five wretched people exist (in addition to five people at wellbeing 10), whereas in B twenty happy people exist (in addition to five people at wellbeing 10). These considerations seem to tell against A. The existence of wretched people is a bad thing, whereas the existence of happy people is neutral. A, we might be tempted to conclude, is less choice-worthy in D59 than B.

SU disagrees. According to SU, A is more choice-worthy in D59 than B. Here is how SU comes to this conclusion. Each outcome is intersectionally worse than the other by 0, of course. So choice-worthiness in D59 comes down to complementary worseness. We have to compare how well the A-not-B people fare in A with how well the B-not-A people fare in B. First, we compare the amounts of wretchedness among these two groups. There is 10 units of wretchedness more in A than in B. This ‘wretched element’ of complementary worseness counts against A. But we also have to compare the levels of wellbeing of the happy people in the two outcomes. We replace negative wellbeings with 0, and then compare the adjusted average wellbeings in each outcome. The adjusted average wellbeing of the A-not-B people in A is 5, and the adjusted average wellbeing of the B-not-A people in B is 3.6. The difference between the two is 1.4, and there are ten people in the smaller of the two groups. This ‘happiness element’ of complementary worseness counts against B by 1.4 × 10 = 14. Things count against A by 10, and things count against B by 14: SU concludes that B is complementarily worse than A by 4 and A is complementarily worse than B by 0. This makes the shortfall of A in D59 0, and

<table>
<thead>
<tr>
<th>D59:</th>
<th>p1, . . . , p5</th>
<th>p6, . . . , p10</th>
<th>p11, . . . , p15</th>
<th>p16, . . . , p35</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10, 10, . . . , 10</td>
<td>* , * , . . . , *</td>
<td>−2, −2, . . . , −2</td>
<td>* , * , . . . , *</td>
</tr>
</tbody>
</table>

can be difficult to imagine. Surely at least the decision-maker is inevitable? Feel free to add some inevitable people to this decision, whose wellbeings are not affected.
5.3. SOME CONCLUSIONS

the shortfall of \( B \) in \( D_{59} \). \( A \) is more choice-worthy in \( D_{59} \) than \( B \).

My initial intuition is that \( B \) is more choice-worthy in \( D_{59} \) than \( A \); that one ought to create the group of all happy people rather than the group containing wretched people. However, on reflection I accept the conclusion of SU, that \( A \) is more choice-worthy in \( D_{59} \) than \( B \). I think that there is a compelling explanation of why this conclusion seems unintuitive but is in fact correct. The mistake in the intuitive way of reasoning I presented above is in ‘matching up’ the five people at wellbeing 10 in \( A \) with the five people at wellbeing 10 in \( B \). It is tempting to think of each of the people with wellbeing 10 in \( A \): ‘in \( B \), instead of that person existing at wellbeing 10 some different person exists at wellbeing 10’. I don’t think that this is right. As I said in Section 4.1.3, there doesn’t seem to be any good way to ‘pair up’ the people in two complementary sets in order to be able to make this kind of pairwise comparison. In \( D_{59} \), for example, why should we think that \( p_1 \) exists in \( A \) instead of \( p_6 \) existing in \( B \)? Why not think that \( p_1 \) exists in \( A \) instead of \( p_{16} \) existing in \( B \) (with wellbeing 2 instead of 10)? Now, any pairing is equivalent to any other if the number of people in each group is the same. Where the number of people differs, the possible pairings can lead to very different results. This is the first step in the defence of the conclusion that \( A \) is more choice-worthy in \( D_{59} \) than \( B \): to overcome the inclination to match up and set aside the two groups of people with wellbeing 10.

The second step is to think through the comparison without making this mistake. The following strikes me as a persuasive way of thinking through the decision, and is a kind of intuitive version of the decision-making process of SU. First: it is bad that people exist wretched in \( A \) instead of not existing at all in \( B \). Noted. However, this does not capture all elements of different between \( A \) and \( B \). It is true that the people who exist wretched in \( A \) don’t exist at all in \( B \). It also true that some different people exist in \( B \) instead of the ones who exist in \( A \). We can’t say which individual replaces which other individual, but we can say that on average the \( B \)-not-\( A \) people fare worse in \( B \) than the \( A \)-not-\( B \) people fare in \( B \). It happens that in this decision, the worseness of wretched lives than none is outweighed by the worseness of lower wellbeings than higher. This won’t always be the case. Here is one decision in which one ought, according to SU, create a group of all happy people rather than one containing a wretched person.
Here, the wretchedness of \( p_5 \) and \( p_6 \) in \( A \) outweighs the lower average in \( B \). \( B \) is more choice-worthy in \( D_{60} \) than \( A \).

### 5.3.5 Conclusion

In this section I have considered the conclusions of SU in some particular variable population decisions. I have defended each of the conclusions reached by SU. In some cases (for example concerning extinction) I considered this to be clearly the best response. In others (for example concerning benign addition) I was less sure, and acknowledged the possibility that some adjustment to SU may improve it. I can summarise my position like this: none of the conclusions that I have considered constitutes a reason to return to any of the familiar general utilitarian theories (TU, AU, HU, or NHU), each of which has more counterintuitive conclusions; however, some of the conclusions that I consider may constitute a reason to explore possible alterations to SU, and other general utilitarian theories.

### 5.4 SU and risk

In this final section of this chapter I shall propose a way of extending SU to deal with risky population decisions. So far I’ve only addressed decisions in which the outcome of each available act is certain. (I made a simplifying assumption to this effect in Section 1.2.) But a moral theory that only makes this kind of decision is not very useful. In practice, the outcomes of available acts are never certain. Our actual decisions are risky. A useful moral theory must therefore make decisions under risk—must assess prospects and not just outcomes. I’ll begin with a brief summary of the standard decision theoretic method for extending CFPU to risky decisions. I’ll then suggest a way of extending SU to risky decisions.

As I’ll be discussing making risky decisions, it will be useful to be able to present these in table form as well. Whereas I have presented decisions under certainty as decisions between outcomes, to present risky decisions I
shall return to the framework of decisions between acts. Each act is associated with a prospect, and each prospect consists of a number of outcome/probability pairs, one for each possible state of the world. Here is a representation of a risky decision.

\[
\begin{array}{c|c|c}
 & 30\% & 70\% \\
\hline
a & p_1(8), p_2(10) & p_1(12), p_2(\ast) \\
b & p_1(\ast), p_2(15) & p_1(\ast), p_2(15) \\
\end{array}
\]

This is the decision between acts \(a\) and \(b\). If \(a\) is chosen (i.e. performed), there is a 30% chance that both \(p_1\) and \(p_2\) will exist with wellbeings 8 and 10 respectively, and a 70% chance that only \(p_1\) will exist with wellbeing 12. Act \(b\), in fact, is not a risky option. If \(b\) is chosen only \(p_2\) will exist, with wellbeing 15. These tables would be quite impractical where large populations are involved, but as I shall focus on simple decisions involving just one or two people they should suffice for my purposes.

### 5.4.1 CFPU, and SU, and risky decisions

Here is a well-known way of extending CFPU to deal with risky decisions. Rather than judging an act on the basis of the outcome it will lead to, we judge it on the basis of its ‘expected outcome’, which is a kind of summary of the possible outcomes of that act and the probabilities with which they will result from that act. Consider this decision.

\[
\begin{array}{c|c|c}
 & 30\% & 70\% \\
\hline
a & p_1(8), p_2(10) & p_1(12), p_2(0) \\
b & p_1(0), p_2(15) & p_1(0), p_2(15) \\
\end{array}
\]

This is a lot like \(D_{61}\) above, but with wellbeing 0 replacing the \(\ast\)s in that decision. This makes it a risky fixed population decision rather than a risky variable population decision. Choosing act \(a\) will result in one of two possible outcomes:

---

46I shall assume that all risk is a matter of epistemic indeterminacy about which state the world is in. A minor adjustment to the framework would allow for physical indeterminacy too (i.e. that the state of the world might not settle which outcome will result from a given act).

47Decision theory has been used to extend various moral theories, but particularly versions of utilitarianism, to risky acts. Hansson observes that in general ‘issues such as risk-taking and risk imposition have been left out of ethics since they are believed to belong to decision theory, and consequently the ethical aspects of these issues have not been treated in either discipline’ (2010, 585). Colyvan, Cox and Steele observe something similar, and shown how utilitarian, deontological, and virtue-based moral theories might be extended (2010).
with 30% probability it will lead to an outcome with a total wellbeing of 18; with 70% probability it will lead to an outcome with a total wellbeing of 12. We can calculate the expected total wellbeing associated with act $a$ from these features of its prospect, as follows: $(0.3 \times 18) + (0.7 \times 12) = 13.8$. We know with certainty that $a$ will not lead to an outcome with a total wellbeing of 13.8. (In this sense the expected outcome is not expected at all!) But if we’re interested in total wellbeings, this is a useful summary of the prospect of $a$. The expected total wellbeing associated with act $b$ is $(0.3 \times 15) + (0.7 \times 15) = 15.48$

According to the standard extension of CFPU, one ought to maximise expected total wellbeing. And, more generally, one available act is more choice-worthy in some decision than another, just in case the expected total wellbeing of the one is higher than the expected total wellbeing of the other. So in $D_{62}$ one ought to choose $b$ rather than $a$; the former is more choice-worthy in $D_{62}$ than the latter. This is not the only possible way of extending CFPU to make risky decisions. One might want to adapt the method to incorporate some kind or risk aversion, for example. However, this does seem like one plausible way. My intention in this section is to suggest one possible similar extension for SU.

Let’s now return to the risky variable population decision, $D_{61}$.

<table>
<thead>
<tr>
<th>$D_{61}$:</th>
<th>30%</th>
<th>70%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>$p_1(8), p_2(10)$</td>
<td>$p_1(12), p_2(*)$</td>
</tr>
<tr>
<td>$b$</td>
<td>$p_1(*), p_2(15)$</td>
<td>$p_1(*), p_2(15)$</td>
</tr>
</tbody>
</table>

My proposal is that we extend SU to risky decisions by minimising expected (rather than actual) shortfalls. This method is roughly analogous to the method of extending CFPU by maximising expected total wellbeing. The first step is to treat each possible state of nature in turn as if it were the certain, known state of nature, and calculate the shortfall of the outcome of each act in the decision according to SU. So in $D_{61}$ we would first assume that the state of nature that is 30% probable is the actual state of nature, and calculate the shortfall of the outcome of each available act in that hypothetical decision under certainty.

---

48We can also make this calculation in a different way. Instead of calculating total wellbeings first, and then assessing our expectation in terms of totals, we could assess the expectation for each individual’s wellbeing first, and then find their total. In $D_{62}$, the expected wellbeing of act $a$ for $p_1$ is $(0.3 \times 8) + (0.7 \times 12) = 10.8$, and for $p_2$ is $(0.3 \times 10) + (0.7 \times 0) = 3$. The sum of these expected wellbeings is 13.8. The expected wellbeing of act $b$ for $p_1$ is $(0.3 \times 0) + (0.7 \times 0) = 0$, and for $p_2$ is $(0.3 \times 15) + (0.7 \times 15) = 15$. The sum of these expected wellbeings is 15. So, there are these two different ways of assessing the prospect of risky fixed population decision that amount to the same thing: the expected total wellbeing of an act is the same as the total expected wellbeing of that act.
Then we would do the same assuming that the other possible state of nature is actual. So, we would calculate shortfalls according to SU for these two decisions under certainty, as indicated.

\[
\begin{array}{c|cc|c}
D_{63} & p_1 & p_2 & \text{Shortfall} \\
\hline
A & 8 & 10 & 5 \\
B & * & 15 & 0 \\
\end{array}
\quad
\begin{array}{c|cc|c}
D_{64} & p_1 & p_2 & \text{Shortfall} \\
\hline
A & 12 & * & 3 \\
B & * & 15 & 0 \\
\end{array}
\]

We now have the shortfall for each possible state of nature, and of course we know how probable each state of nature is. So, we can calculate the expected shortfall of each act. The expected shortfall of act \(a\) is \((0.3 \times 5) + (0.7 \times 3) = 3.6\). The expected shortfall of act \(b\) is \((0.3 \times 0) + (0.7 \times 0) = 0\). If one ought to minimise expected shortfalls, act \(b\) is more choice-worthy in \(D_{61}\) than act \(a\), and one ought to choose \(b\) and not \(a\) when faced with that decision.

5.4.2 Using risky decisions

My main goal in this section was to suggest one way of extending SU to deal with risky decisions. I want to make one final and distinct point about risky decisions. I want to suggest that risky variable population decisions might be useful in providing further support for SU, and further criticism of the contending general utilitarian theories (or vice versa, of course). When we consider what the implications might be of any plausible extensions of the various general utilitarian theories, we might find that this tells in favour of one or another theory. We could add this to our stock of intuition-based tests for those theories.

I haven’t defended or even fully spelled out a way of extending any of the general utilitarian theories we have encountered. However, based on the suggestions I’ve made in this section and some compelling ideas of how, for

---

49 There is a second proposal, analogous to the alternative suggested in the previous footnote. One might minimise shortfalls in expected wellbeings conditional on existence. In \(D_{63}\) the expected wellbeing of \(p_2\) if \(a\) is performed and \(p_2\) exists is \(1 \times 10 = 10\). If \(a\) is performed \(p_1\) will exist with certainty, so his expected wellbeing conditional on existence is just the same as his expected wellbeing: \((0.3 \times 8) + (0.7 \times 12) = 10.8\). Likewise for the expected wellbeing conditional on existence of \(p_2\) if \(b\) is performed: \((0.3 \times 15) + (0.7 \times 15) = 15\). As for the expected wellbeing conditional on existence of \(p_1\) if \(b\) is performed: \(p_1\) will certainly not exist if \(b\) is performed, and I think we should carry on treating this as *. Having made all of these calculations of expected wellbeing conditional on existence, we can then construct a decision under certainty using them, and calculate shortfalls for that decision. These two methods of extending SU agree in the case of \(D_{61}\). I don’t know whether they will agree in all cases.
example, TU might be extended, I’ll illustrate this idea by reference to one risky variable population decision.

Imagine a woman has just conceived. The possible future child does not yet exist. As things stand, there is a 20% chance of miscarriage and an 80% chance that the child will come to exist and live a very good life. Available to the woman is a miscarriage prevention drug that would reduce the risk of miscarriage to 5% but make it the case that the child, if he does exist, will live a worse life (either significantly or slightly worse). Also available to the woman is an abortion drug, which would with certainty terminate the pregnancy before the possible person comes into existence. Here is a representation of this decision.

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th>15%</th>
<th>80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Act a</td>
<td>*</td>
<td>*</td>
<td>100</td>
</tr>
<tr>
<td>Act b</td>
<td>60</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>Act c</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

Ought the woman to take the miscarriage prevention drug? What is the moral status of taking the abortion drug?

The most plausible way to extend TU is, I think, the same way we extended CFPU, with 0s replacing *s where applicable. The expected total wellbeing for act a is \((0.5 \times 0) + (0.15 \times 0) + (0.8 \times 100) = 80\). The expected total wellbeing for act b is \((0.5 \times 0) + (0.15 \times 60) + (0.8 \times 100) = 85\). The expected total wellbeing for c is \((0.5 \times 0) + (0.15 \times 0) + (0.8 \times 0) = 0\). If one ought to maximise expected total wellbeing, then b is more choice-worthy in \(D_{65}\) than a, which is more choice-worthy in \(D_{65}\) than c. Our pregnant woman ought to take the disability-inducing miscarriage prevention drug. The morally-worst thing she could do would be to abort the pregnancy.

The extension of SU that I suggested above disagrees. The expected shortfall of acts a and c in this decision are both 0, and the expected shortfall of b is \(0.8 \times 5 = 4\). The pregnant woman ought not take the miscarriage prevention drug. She may either continue with the pregnancy, or abort before the child comes into existence. This seems to me to be a more intuitive conclusion. I would be concerned if people started taking (or doctors started prescribing) \(50\) Again, if you can’t imagine this adjust the decision to a moment pre-conception.

\(51\) By stipulation this is not a non-identity decision situation. Jeff McMahan and Luara Ferracioli pointed out to me in a seminar that, if the woman could ‘try again’ until she had a successful, drug-free pregnancy, this might be the most choice-worthy option according to the extension of TU being considered. This option is not available in \(D_{65}\).
this kind of miscarriage prevention drug. It is quite appalling to me, to think that one might reduce a possible person’s wellbeing in order to increase the chance that that person exists at all. Others’ intuitions may differ. (I would not be surprised if those who don’t share my intuition about the moral neutrality of creating a happy person disagreed.) My main point here, though, is that risky variable population decisions may be an untapped source of differentiation between general utilitarian theories.  

Conclusion

In this chapter I have considered some possible objections to SU. First, in Section 5.1, I addressed the menu-dependence of choice-worthiness according to SU. I argued that some objections to this kind of menu-dependence concern series of decisions, and thus miss their target: SU is a theory developed in a framework in which there can be no series of decisions. Second, in Section 5.2, I responded to concerns about the claim that there is no overall betterness relation among outcomes. I defended this possibility and the way I have represented theories like SU, but endorsed one way of salvaging overall betterness: by appeal to a finer grain of outcomes. In Section 5.3 I considered the conclusions of SU in some particular variable population decisions. I defended (at least provisionally) the conclusions of SU regarding extinction, McDermott cases, benign addition, and wretchedness. Finally, in Section 5.4 I proposed a method for extending SU to deal with risky decisions, and suggested that risky variable population decisions might provide further grounds for preferring one general utilitarian theory or another.

---

52 My focus here is on the acceptability of conclusions about these decisions directly, or in themselves. In Section 5.1 I also raised the related issue of possible structural problems with extensions of SU to risky decisions (see pages 141–143).
CHAPTER 5. DEFENDING SHORTFALL UTILITARIANISM
Chapter 6

Conclusion

Sometimes—probably much more often than is immediately obvious—we face decisions that affect who comes to exist. How ought we to make such decisions? In this dissertation I have proposed an answer to this question, in the form of a moral theory: Shortfall Utilitarianism. In this concluding chapter, let me briefly review my reasons for proposing this theory, what I have been able to say in favour of it, and the persisting questions I have.

I began in Chapter 1 by adopting a framework for representing decisions and moral theories. I defined a decision as a set of acts available to an agent at a time, where each act would, if performed, lead with certainty to a particular outcome. I defined an outcome as a complete possible world. This picture of decisions is widely adopted in the population ethics literature (particularly in ‘population axiology’). My reasons for accepting it, which I take to be the usual reasons, were as follows. The assumption of certainty serves to isolate for moral examination some aspects decisions—in particular, who exists and how well they fare—from the complications associated with risk. This assumption was intended to be temporary. It must eventually be given up in order to develop a practically useful moral theory. Complete possible worlds are a natural object of attention in this field, in which we attempt to grapple with consequences of actions into the vast temporal and spatial distance.

A moral theory, I asserted, takes decisions as input and returns a choice-worthiness ordering for each; that is, an ordering of the available acts in the decision in terms of how morally good it would be to perform that act when
faced with that decision. (At least, the theories that I would consider do this; I acknowledged that others may provide less detailed choice-guiding output.) My depiction of moral theories was intended to be quite permissive. I focussed on choice-worthiness, and carefully distinguished this from notions of intrinsic goodness of acts or outcomes, because this seemed to me the best way to represent the theories forthcoming in the dissertation.

I spelled out what I meant by ‘decisions that affect who will come to exist’ by way of the distinction between fixed population decisions and variable population decisions. A fixed population decision is one that does not affect who ever exists: exactly the same people will ever exist, whichever of the available acts is chosen. A variable population decision is any non-fixed population decision: at least one possible person will exist if one available act is chosen and will not exist if some other available act is chosen.

In Chapter 2 I narrowed my focus to the question of how we ought to make decisions as utilitarians. I introduced Classical Fixed Population Utilitarianism (CFPU), a partial utilitarian theory and, I suggested, the historical and conceptual core of utilitarianism. According to CFPU, choice-worthiness of acts in fixed population decisions tracks the total wellbeing of those acts’ outcomes. There are many (perhaps surprisingly many) ways to extend CFPU to variable population decisions; many general moral theories that agree with CFPU in fixed population decisions and also provide choice-worthiness orderings for variable population decisions. I appealed to three characteristics of CFPU central to the utilitarian spirit to further narrow my focus: consequentialism, axiology, and welfarism. I defined a general utilitarian theory as a consequentialist, axiological, welfarist extension of CFPU, and adopted the goal of finding a plausible general utilitarian theory.

My reasons for focussing on general utilitarian theories were as follows. First, I focused on utilitarianism because I think it highly likely that utilitarianism is at least part of the moral story. I am inclined to think that, at least other things being equal, one ought to make people as well off as possible; and that CFPU captures what it means to make people as well off as possible in the context of fixed population decisions. It seemed reasonable, therefore, to set aside any other morally relevant ‘things’ (deontic constraints or rights or equality, for example) and focus on finding the best version of utilitarianism; the best general spelling out of ‘make people as well off as possible’. This goal
would be of particular interest, of course, to someone who shares my belief that there is at least a utilitarian element to morality. However, I suggested that the project may also be of broader interest than this. For one thing, it might help critics of utilitarianism to understand both the variety, and the limits, of utilitarian theories. For another, it might be that the issues addressed in the context of utilitarianism also occur, or have analogues, outside that context (in which case, the solutions I propose here may be useful there too).

The details of my definition of a general utilitarian theory were also important; particularly the repudiation of the popular overall betterness requirement. I avoided this requirement, while embracing both consequentialism and axiology, in order to make space for theories in which multiple respects of betterness among outcomes, irreducible to an overall betterness relation, determine choice-worthiness. I made use of an analogy with deciding which beach to go to, in order to demonstrate the possibility of this kind of theory.

At the end of Chapter 2 I introduced three general utilitarian theories: Total Utilitarianism (TU), Average Utilitarianism (AU), and Harm Utilitarianism (HU). According to TU, choice-worthiness in any decision tracks total well-being in the available outcomes. According to AU, choice-worthiness in any decision tracks average wellbeing in the available outcomes. According to HU, choice-worthiness in any decision tracks total harm incurred in the available outcomes. TU and AU are well known as versions of utilitarianism, whereas the status of HU is more controversial—in virtue, I think, of its failure to conform to the overall betterness requirement.

In Chapter 3 I described four of my strongest utilitarian intuitions about variable population decisions. Ideally, I thought, a general utilitarian theory should agree with all four of these intuitions. My first intuition was that there is a normative wretchedness threshold on the wellbeing scale: a wellbeing level such that, other things being equal, one ought not create a person who will live a life at a wellbeing below that level. My second intuition was the other half of ‘the Asymmetry’: that, other things being equal, it is not the case that one ought to create a person who will live at or above the wretchedness threshold. This is the intuition of neutrality. My third intuition was that one ought to create a person with higher wellbeing, rather than creating a person with lower wellbeing, even if they are two different possible people. This is the intuition that the non-identity problem really is a problem. My fourth and final intuition
was that, other things being equal, one ought to create a smaller population of people with higher wellbeing, rather than a larger population with lower wellbeing. This is a strong anti-repugnance intuition.

As I spelled out these intuitions, I tested TU, AU, HU, and some variations of those theories, for whether they agree with the intuitions. None of these familiar general utilitarian theories agrees with all four intuitions. TU fails on the neutrality intuition and the anti-repugnance intuition. AU fails on the wretchedness and neutrality intuitions (both halves of the Asymmetry). HU fails on the wretchedness intuition; and HU and its cousin Non-existence Harm Utilitarianism (NHU) both fail on the non-identity intuition and the anti-repugnance intuition. This brought me roughly up to the current state of play: all of the extant general utilitarian theories seem thoroughly deficient in one way or another.

Enter Shortfall Utilitarianism (SU). In Chapter 4 I set out SU in full. According to SU one outcome can be worse than another in two morally relevant ways; and if one available outcome in a decision is worse than a second available outcome in that decision in either of these ways, that tells against choosing the first available outcome in that decision. When faced with any decision, one ought to choose the available outcome that is worse than another available outcome in either of these ways by the least amount.

According to SU, the two morally relevant kinds of worseness are intersectional and complementary worseness. Intersectional worseness is a matter of comparing a pair of outcomes in terms of the wellbeings of the people who exist in both. The idea here is that, if the people who exist in both outcomes fare worse in one than in the other (as a group), that tells against choosing the one rather than the other. Complementary worseness is a matter of comparing a pair of outcomes in terms of the wellbeings of the people who exist in one outcome (either one), but not the other. If the people who exist only in the one outcome fare worse there, than the people who exist only in the other outcome fare there, that tells against choosing the one outcome rather than the other.

Having set out the theory, I showed that SU is a general utilitarian theory that agrees with my four key intuitions. SU is a welfarist, axiological, consequentialist extension of CFPU: it is a theory according to which the choice-worthiness of the available acts in a decision depends only on the relations of betterness among the outcomes of those acts. However, according to SU there is no overall
betterness relation among outcomes (at least, not a standardly structured one). Axiology and consequentialism have both sometimes been taken to require an overall betterness relation among outcomes. No theory based on an overall betterness relation agrees with all four of the intuitions I focussed on in Chapter 3. In showing that at least one theory based on multiple betterness relations does agree with all of these intuitions, I think I have provided some reason for people who share those intuitions, and who are drawn to consequentialist and axiological theories, to shift their attention away from simple overall betterness and towards theories like SU. I offered SU as a plausible utilitarian moral theory. I further suggested that the class of welfarist, axiological, consequentialist moral theories that do not provide an overall betterness relation among outcomes is under-explored.

In Chapter 5 I addressed some possible concerns about SU, structured around four themes. First, I argued that the kind of concern often raised about menu-dependent choice-worthiness, concerning as it does series of decisions, is mis-directed. There remains the question of whether similar concerns will arise for a theory that extends SU to deal with risky decisions. A conclusive investigation of this question was beyond my scope. Second, I addressed some questions about overall betterness. I argued that there is no good reason to try to salvage an overall betterness relation among outcomes, and rejected some possible ways of doing so. However, I endorsed one such way: by means of ‘fine-graining’ outcomes. I also addressed Arrhenius’s normative impossibility theorems. I argued that at least two of his choice-worthiness ‘adequacy conditions’ are false, and attributed their intuitive appeal to the habit of thinking in overall betterness terms. Third, I considered the conclusions of SU in some particular variable population decisions, which may strike some people as un-intuitive. I defended those conclusions, and occasionally also suggested some possible alterations to SU. Finally, I proposed a way of extending SU to deal with risky decisions, based on expected shortfalls.

I hope to have persuaded my reader of two claims. Claim 1: SU is an attractive general utilitarian theory for anyone who shares my utilitarian intuitions. SU reflects the Asymmetry between wretched and happy lives, delivers intuitive verdicts in non-identity cases, and avoids repugnant conclusions. It achieves these results on the basis of compelling morally relevant betterness relations among outcomes, which are determined by the wellbeings of people
in those outcomes. Claim 2: The class of general utilitarian theories includes welfarist, axiological, consequentialist theories that are incompatible with the overall betterness requirement. In light of this second claim, anyone attracted to the former characteristics but dissatisfied with the possibilities afforded under the overall betterness restriction should be encouraged to investigate the options afforded by rejecting that restriction.

Let me conclude by returning to the question of the significance of my project. In Section 1.2.4 I said that it was important to address the question of how we ought to make variable population decisions for three reasons: because those decisions are common, because they are morally perplexing and divisive, and because the way we address them will have significant consequences for our assessment of many pressing practical decisions. I can now perhaps make the final point more clearly.

I mentioned there the possibility that proper appreciation of the significance of variable population decisions should lead us to accept that we have strong moral obligations to pursue space exploration programs, perhaps even outweighing our obligations to alleviate ongoing poverty and suffering. It should now be clear why this is a possible implication of accepting TU (or a plausible risk-sensitive extension thereof). Colonising distant planets may lead to the existence of vastly many happy people. According to TU securing the existence of those happy people may, if they are numerous enough, be worth reducing the wellbeings of people who already exist or who will inevitably exist in the future, and even bringing into existence many wretched people. Of course, what the implications of TU will be for any realistic decision is a complex matter. However, philosophers who have felt compelled to accept a moral theory like TU, thinking that there is no better option, have endorsed space exploration as a moral endeavour on a par with improving lives, preventing war, and the like.\(^1\)

I do not think that we are yet at the point, in the field of population ethics

---

\(^1\)I should point out that other general utilitarian theories threaten similarly disturbing practical implications. Were we to accept NHU, we would have to conclude that, for the purposes of making many decisions, people who will exist in the more distant future are less morally significant than people who will exist sooner. This is because many decisions affect which people will exist in the future, and this effect becomes more significant over time. A decision that changes the identities of a small proportion of the population in the next decade will change the identities of a larger proportion of the population in the next century, and so on. Were we to accept NHU, we would, I think, have to accept that the sacrifices required by people presently alive to reduce the impact of climate change on wellbeings in three hundred years is minimal.
or even in the broadly utilitarian branch of that field, at which we should be willing to accept such disturbing conclusions. The prospects for finding better options seem to me to be good. I think that SU is one better option. SU reconciles four of my strongest intuitions about variable population decisions. SU also supports my intuitions about space exploration: according to SU we are not obliged to bring into existence more possible future people; and certainly not at the expense of inevitable people’s wellbeings. I also think that there are likely to be other appealing options in the class of general utilitarian theories.
# Appendix A

## List of Decisions

### D1

<table>
<thead>
<tr>
<th></th>
<th>Anne</th>
<th>Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a \Rightarrow A$</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>$b \Rightarrow B$</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

### D2

<table>
<thead>
<tr>
<th></th>
<th>Anne</th>
<th>Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a \Rightarrow A$</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>$c \Rightarrow C$</td>
<td>9</td>
<td>*</td>
</tr>
</tbody>
</table>

### D3

<table>
<thead>
<tr>
<th></th>
<th>Anne</th>
<th>Bob</th>
<th>Carl</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a \Rightarrow A$</td>
<td>10</td>
<td>8</td>
<td>*</td>
</tr>
<tr>
<td>$d \Rightarrow D$</td>
<td>9</td>
<td>*</td>
<td>9</td>
</tr>
</tbody>
</table>

### D4

<table>
<thead>
<tr>
<th></th>
<th>Ocean</th>
<th>Sand</th>
<th>Pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bondi</td>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Bronte</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

### D5

<table>
<thead>
<tr>
<th></th>
<th>Ocean</th>
<th>Sand</th>
<th>Pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bondi</td>
<td>5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Tamarama</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

### D6

<table>
<thead>
<tr>
<th></th>
<th>Ocean</th>
<th>Sand</th>
<th>Pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tamarama</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Bronte</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

### D7

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>8</td>
<td>*</td>
<td>*</td>
<td>12</td>
</tr>
</tbody>
</table>

### D8

<table>
<thead>
<tr>
<th></th>
<th>Anne</th>
<th>Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>$B$</td>
<td>10</td>
<td>*</td>
</tr>
<tr>
<td>Decision</td>
<td>Alternatives</td>
<td>Payoffs</td>
</tr>
<tr>
<td>----------</td>
<td>--------------</td>
<td>---------</td>
</tr>
<tr>
<td>( D_9 ):</td>
<td>A, B, C</td>
<td>( p_1 ), ( p_2 )</td>
</tr>
<tr>
<td>( D_{10} ):</td>
<td>Anne, Bob</td>
<td>A, B</td>
</tr>
<tr>
<td>( D_{11} ):</td>
<td>Carl, Dave</td>
<td>A, B</td>
</tr>
<tr>
<td>( D_{12} ):</td>
<td>Anne, Bob</td>
<td>A, B, C</td>
</tr>
<tr>
<td>( D_{13} ):</td>
<td>( p_1 ), ( p_2 )</td>
<td>A, B</td>
</tr>
<tr>
<td>( D_{14} ):</td>
<td>( p_1 ), ( p_2 )</td>
<td>A, B, C</td>
</tr>
<tr>
<td>( D_{15} ):</td>
<td>( p_1 ), ( p_2 )</td>
<td>A, B, C</td>
</tr>
<tr>
<td>( D_{16} ):</td>
<td>( p_1 ), ( p_2 )</td>
<td>A, B, C</td>
</tr>
<tr>
<td>( D_{17} ):</td>
<td>( p_1 ), ( p_2 )</td>
<td>A, B, C</td>
</tr>
<tr>
<td>( D_{18} ):</td>
<td>( x ) billion Earthlings, ( y ) billion Martians</td>
<td>( p_1 ), ( p_2 ), ( p_3 )</td>
</tr>
<tr>
<td>( D_{21} )</td>
<td>( x ) billion Earthlings</td>
<td>( y ) billion Martians</td>
</tr>
<tr>
<td>-------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>( A )</td>
<td>80,80,\ldots,80</td>
<td>80,80,\ldots,80</td>
</tr>
<tr>
<td>( B )</td>
<td>80,80,\ldots,80</td>
<td><em>\ldots,</em></td>
</tr>
</tbody>
</table>

| \( A \) | 1 trillion people | \( x \) trillion people |
| \( B \) | 1 trillion people | \( x \) trillion people |

<table>
<thead>
<tr>
<th>( D_{23} )</th>
<th>( x ) trillion people</th>
<th>( (x + 1) ) trillion people</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>100,100,\ldots,100</td>
<td><em>\ldots,</em></td>
</tr>
<tr>
<td>( B )</td>
<td><em>\ldots,</em></td>
<td>1,1,\ldots,1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( x ) people</th>
<th>( p_1 )</th>
<th>( p_2 )</th>
<th>100 people</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D_{24} )</td>
<td>( A )</td>
<td>58,58,\ldots,58</td>
<td>50</td>
</tr>
<tr>
<td>( B )</td>
<td>57,57,\ldots,57</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( D_{25} )</th>
<th>( p_1 )</th>
<th>( p_2 )</th>
<th>101 people</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>50</td>
<td>50</td>
<td><em>\ldots,</em></td>
</tr>
<tr>
<td>( B )</td>
<td>1</td>
<td>1</td>
<td>1,1,\ldots,1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( D_{26} )</th>
<th>( p_1 )</th>
<th>( p_2 )</th>
<th>102 people</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>50</td>
<td>50</td>
<td><em>\ldots,</em></td>
</tr>
<tr>
<td>( B )</td>
<td>1</td>
<td>1</td>
<td>1,1,\ldots,1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( D_{27} )</th>
<th>( p_1 )</th>
<th>( p_2 )</th>
<th>102 people</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>50</td>
<td>50</td>
<td><em>\ldots,</em></td>
</tr>
<tr>
<td>( B )</td>
<td>1</td>
<td>1</td>
<td>1,1,\ldots,1</td>
</tr>
</tbody>
</table>

| \( D_{28} \) | 10 million people | 3 million people |
|\( A \) | 100,100,\ldots,100 | *\ldots,* |
| \( A^+ \) | 100,100,\ldots,100 | 70,70,\ldots,70 |

| \( D_{29} \) | 10 million people | 3 million people |
|\( A^+ \) | 100,100,\ldots,100 | 70,70,\ldots,70 |
| \( B \) | 95,95,\ldots,95 | 95,95,\ldots,95 |

| \( D_{30} \) | 10 million people | 3 million people |
|\( A \) | 100,100,\ldots,100 | *\ldots,* |
| \( B \) | 95,95,\ldots,95 | 95,95,\ldots,95 |
### APPENDIX A. LIST OF DECISIONS

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{31}$:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A$</td>
<td>6</td>
<td>9</td>
<td>9</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>10</td>
<td>8</td>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{32}$:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A$</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>$B$</td>
<td>5</td>
<td>10</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>$C$</td>
<td>10</td>
<td>5</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>$D$</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{33}$:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A$</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>5</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>$B$</td>
<td>5</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{34}$:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A$</td>
<td>4</td>
<td>2</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>6</td>
<td>*</td>
<td>−2</td>
<td>−1</td>
</tr>
<tr>
<td>$C$</td>
<td>*</td>
<td>1</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{35}$:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A$</td>
<td>8</td>
<td>2</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>6</td>
<td>*</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>$C$</td>
<td>*</td>
<td>4</td>
<td>4</td>
<td>−6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{36}$:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A$</td>
<td>8</td>
<td>8</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>$B$</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{37}$:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A$</td>
<td>8</td>
<td>8</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>*</td>
</tr>
<tr>
<td>$C$</td>
<td>8</td>
<td>8</td>
<td>*</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{38}$:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A$</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>$B$</td>
<td>10</td>
<td>*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{39}$:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A$</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>$B$</td>
<td>10</td>
<td>*</td>
</tr>
<tr>
<td>$C$</td>
<td>10</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{40}$:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A$</td>
<td>19</td>
<td>*</td>
<td>10</td>
</tr>
<tr>
<td>$B$</td>
<td>22</td>
<td>10</td>
<td>*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{41}$:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$B$</td>
<td>22</td>
<td>10</td>
<td>*</td>
</tr>
<tr>
<td>$C$</td>
<td>*</td>
<td>12</td>
<td>*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{42}$:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A$</td>
<td>19</td>
<td>*</td>
<td>10</td>
</tr>
<tr>
<td>$C$</td>
<td>*</td>
<td>12</td>
<td>*</td>
</tr>
</tbody>
</table>
### D₄₃:

<table>
<thead>
<tr>
<th></th>
<th>p₁</th>
<th>p₂</th>
<th>p₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>19</td>
<td>*</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>22</td>
<td>10</td>
<td>*</td>
</tr>
<tr>
<td>C</td>
<td>*</td>
<td>12</td>
<td>*</td>
</tr>
</tbody>
</table>

### D₄₄:

<table>
<thead>
<tr>
<th></th>
<th>Anne</th>
<th>Bob</th>
<th>Carl</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8</td>
<td>10</td>
<td>*</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>8</td>
<td>10</td>
</tr>
</tbody>
</table>

### D₄₅:

<table>
<thead>
<tr>
<th>Anne</th>
<th>Bob</th>
<th>Carl</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### D₄₆:

<table>
<thead>
<tr>
<th></th>
<th>p₁</th>
<th>p₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>15</td>
<td>*</td>
</tr>
<tr>
<td>A⁺</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>B</td>
<td>13</td>
<td>13</td>
</tr>
</tbody>
</table>

### D₄₇:

<table>
<thead>
<tr>
<th></th>
<th>p₁</th>
<th>p₂</th>
<th>p₃</th>
<th>p₄</th>
<th>p₅</th>
<th>p₆</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>50</td>
<td>5</td>
<td>5</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>B</td>
<td>50</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>C</td>
<td>25</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>20</td>
</tr>
</tbody>
</table>

### D₄₈:

<table>
<thead>
<tr>
<th>Anne</th>
<th>Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>9</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>*</td>
</tr>
</tbody>
</table>

### D₄₉:

<table>
<thead>
<tr>
<th>Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>F</td>
</tr>
</tbody>
</table>

### D₅₀:

<table>
<thead>
<tr>
<th>n people</th>
<th>m people</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 30, 30, ..., 30</td>
<td>*, *, ..., *</td>
</tr>
<tr>
<td>B 30, 30, ..., 30</td>
<td>30, 30, ..., 30</td>
</tr>
</tbody>
</table>

### D₅₁:

<table>
<thead>
<tr>
<th></th>
<th>p₁</th>
<th>p₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>*</td>
</tr>
<tr>
<td>C</td>
<td>*</td>
<td>10</td>
</tr>
<tr>
<td>D</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>

### D₅₂:

<table>
<thead>
<tr>
<th></th>
<th>p₁</th>
<th>p₂</th>
<th>p₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>*</td>
<td>*</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>*</td>
<td>9</td>
</tr>
<tr>
<td>C</td>
<td>*</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>D</td>
<td>5</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>10</td>
<td>9</td>
</tr>
</tbody>
</table>
### APPENDIX A. LIST OF DECISIONS

<table>
<thead>
<tr>
<th>$D_53$:</th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$D_54$:</th>
<th>$p_1$</th>
<th>$p_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>15</td>
<td>*</td>
<td>$A^+$</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>$A^+$</td>
<td>16</td>
<td>10</td>
<td>$B$</td>
<td>14.5</td>
<td>14.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$D_55$:</th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$D_56$:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>15</td>
<td>*</td>
<td>$A^+$</td>
</tr>
<tr>
<td>$B$</td>
<td>14.5</td>
<td>14.5</td>
<td>$B$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$D_57$:</th>
<th>$p_1$</th>
<th>$p_2$</th>
<th>$p_3$</th>
<th>$p_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>15</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$A^+$</td>
<td>16</td>
<td>10</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>14.5</td>
<td>14.5</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$B^+$</td>
<td>14.5</td>
<td>14.5</td>
<td>7</td>
<td>*</td>
</tr>
<tr>
<td>$C$</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>*</td>
</tr>
<tr>
<td>$C^+$</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>$D$</td>
<td>13.5</td>
<td>13.5</td>
<td>13.5</td>
<td>13.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$D_58$:</th>
<th>$x$ people</th>
<th>$y$ people</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>50, 50, ... , 50</td>
<td>* , * , ... , *</td>
</tr>
<tr>
<td>$B$</td>
<td>50, 50, ... , 50</td>
<td>-20, 50, 50, ... , 50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$D_59$:</th>
<th>$p_1, p_2$</th>
<th>$p_3, p_4$</th>
<th>$p_5, p_6$</th>
<th>$p_7, ... , p_10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>10, 10, ... , 10</td>
<td>* , * , ... , *</td>
<td>-2, -2, ... , -2</td>
<td>* , * , ... , *</td>
</tr>
<tr>
<td>$B$</td>
<td>* , * , ... , *</td>
<td>10, 10, ... , 10</td>
<td>* , * , ... , *</td>
<td>2, 2, ... , 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$D_60$:</th>
<th>$p_1, p_2$</th>
<th>$p_3, p_4$</th>
<th>$p_5, p_6$</th>
<th>$p_7, ... , p_10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>10, 10</td>
<td>* , *</td>
<td>-2, -2</td>
<td>* , *</td>
</tr>
<tr>
<td>$B$</td>
<td>* , *</td>
<td>10, 10</td>
<td>* , *</td>
<td>2, ... , 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$D_61$:</th>
<th>30%</th>
<th>70%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>$p_1(8), p_2(10)$</td>
<td>$p_1(12), p_2(*)$</td>
</tr>
<tr>
<td>$b$</td>
<td>$p_1(*), p_2(15)$</td>
<td>$p_1(*), p_2(15)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$D_62$:</th>
<th>30%</th>
<th>70%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>$p_1(8), p_2(10)$</td>
<td>$p_1(12), p_2(0)$</td>
</tr>
<tr>
<td>$b$</td>
<td>$p_1(0), p_2(15)$</td>
<td>$p_1(0), p_2(15)$</td>
</tr>
<tr>
<td></td>
<td>$p_1$</td>
<td>$p_2$</td>
</tr>
<tr>
<td>----</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>$A$</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>$B$</td>
<td>*</td>
<td>15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$p_1$</th>
<th>$p_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>12</td>
<td>*</td>
</tr>
<tr>
<td>$B$</td>
<td>*</td>
<td>15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>5%</th>
<th>15%</th>
<th>80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Act $a$</td>
<td>*</td>
<td>*</td>
<td>100</td>
</tr>
<tr>
<td>Act $b$</td>
<td>*</td>
<td>60</td>
<td>95</td>
</tr>
<tr>
<td>Act $c$</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
Bibliography


Gustafsson, J. E.: 2010, A money-pump for acyclic intransitive preferences, *Dialectica* 64(2), 251–257.


Hare, C.: 2007, Voices from another world: Must we respect the interests of people who do not, and will never, exist?, *Ethics* 117(3), 498–523.


Parfit, D.: 2015a, Can we avoid the repugnant conclusion? Unpublished manuscript.


BIBLIOGRAPHY


