Incommensurability in Population Ethics

Jacob M. Nebel
BPhil Thesis (2015)

Abstract: Values are incommensurable when they cannot be measured on a single cardinal scale. Many philosophers suggest that incommensurability can help us solve the problems of population ethics. I agree. But some philosophers claim that populations bear incommensurable values merely because they contain different numbers of people, perhaps within some range. I argue that mere differences in how many people exist, even within some range, do not suffice for incommensurability. I argue that the intuitive neutrality of creating happy people is better captured by a version of average utilitarianism. But this view is problematic. So I suggest a version of total utilitarianism that avoids the repugnant conclusion by appealing to incommensurable dimensions of wellbeing.

Contents

Introduction

This thesis is about a family of views in population ethics. I am interested in views that appeal to *incommensurability*. Values are incommensurable when they cannot be measured on a single cardinal scale. My question is whether incommensurability can help us solve the problems of population ethics—most importantly, the problem of avoiding

The Repugnant Conclusion: For any population of excellent lives, there is some much larger population of people whose existence would be better, even though their lives would be barely worth living.[1]

Why might we be unable to measure the values of populations on a single cardinal scale? I consider three views.

Chapter 1 is about a view inspired by Parfit and developed by Blackorby, Bossert, and Donaldson (1996), Qizilbash (2007), and Rabinowicz (2009). This view says that adding people whose wellbeing is within some limited range makes things neither better nor worse, nor equally good. It results in what Chang (2002) calls *parity*. I discuss two objections to this view, raised by Broome (2004). I argue that one of Broome's objections succeeds, but that the other fails in a way that spells trouble for Broome's own view.

Chapter 2 is about a radical proposal that avoids Broome's objections. Bader (manuscript) argues that different-sized populations are *incomparable*. I present

[1] See (Parfit 1984, 388).

two objections to Bader's theory and argue that they are best avoided by average utilitarianism.[2] Philosophers tend not to take average utilitarianism seriously. But I argue that we can revise average utilitarianism in a natural way to capture the intuitive neutrality of creating happy people, by appealing to some incomparability. I find the resulting theory problematic, but still worth taking seriously.

In Chapter 3, I discuss a different kind of incommensurability: lexical superiority. One good is lexically superior to another if some amount of the one is better than any amount of the other. I discuss Parfit's (2004) lexical view, which he calls *perfectionism*, on which the loss of the best things in life cannot be outweighed by any improvements in wellbeing. I suggest a different lexical view, called *lexical total utilitarianism*, on which no population of people whose lives are all barely worth living could contain a greater sum of wellbeing than any population of excellent lives. I then argue that lexical total utilitarians should appeal to parity, in a way that avoids Broome's objections.

I draw two conclusions. The first conclusion is negative: incommensurability in population ethics is not plausibly explained by mere differences in population size. The second conclusion is positive: we can avoid the repugnant conclusion by appealing to incommensurable dimensions of wellbeing.

[2] I use "utilitarian" to refer only to the axiology, not to the conjunction of the axiology and consequentialism.

1 Parity and Mere Addition

The views that I discuss are best understood as different responses to Parfit's mere addition paradox. Here is one version of the paradox:

$$A = (100, \Omega, \Omega, \ldots, \Omega),$$

$$A+ = (100, 1, 1, \ldots, 1),$$

$$Z = (2, 2, 2, \ldots, 2)$$

These lists represent distributions of wellbeing. The numbers represent how good people's lives are. Level 100 for any person, for example, is a hundred times better than level 1 for any other person. All positive numbers represent lives that are worth living. I use Ω to indicate that a person has no life, and so no quality of life, in a distribution. The ellipses indicate that many other people exist at the relevant level.

For this example, imagine that each component in the list represents the welfare of ten billion people. In A, there are ten billion people with excellent lives. In $A+$, these same people's lives are just as good, and many more people exist with lives that are barely worth living. How many more people? Enough so that $A+$'s average wellbeing is less than 2. In Z, wellbeing is uniformly distributed so that every life is twice as good as the mediocre lives in $A+$.

The paradox consists in these claims:

(1) $A+$ is at least as good as A.

(2) Z is better than $A+$.

(3) Z is not better than A.

These claims are inconsistent. If Z is better than $A+$, which is at least as good as A, then Z must be better than A.[3]

Each claim seems highly plausible. (1) is supported by

The Mere Addition Principle: "For any population [A], if one adds any number of people with positive welfare to create a new population [$A+$], without affecting the [A]-people's welfare, then [$A+$] is at least as good as [A]" (Carlson 1998, 285).

And (2) is supported by

Non-Anti-Egalitarianism: "A population with perfect equality is better than a population with the same number of people, inequality, and lower average (and thus lower total) welfare" (Arrhenius 2000, 253).

If we accept the mere addition principle and non-anti-egalitarianism, then we must deny (3).[4] But if the lives in A are excellent, and the lives in Z are barely worth living, then this seems repugnant.

[3] I assume throughout that *at least as good as* is transitive and reflexive, and that *better than* is the asymmetric part of this relation. These assumptions are sufficient for the inference to be good (Sen 1970a, 10).

[4] See also Ng (1989) for both principles.

Parfit (1984, 431) suggests a response to the mere addition principle. He suggests that $A+$ is not worse than A, but that *not worse than* doesn't imply *at least as good as*. This chapter is about a theory inspired by Parfit's suggestion, called *critical-band utilitarianism*. The chapter comes in four sections. Section 1.1 introduces some concepts needed for making sense of Parfit's suggestion, along with related concepts that I use throughout the thesis. Section 1.2 introduces critical-band utilitarianism. Section 1.3 discusses Broome's objection that critical-band utilitarianism is *ad hoc*: I agree. Section 1.4 discusses Broome's objection that critical-band utilitarianism is incompatible with vagueness: I disagree, and I think Broome's own view is *ad hoc* too.

1.1 Concepts

Parfit's suggestion that $A+$ is neither worse than nor at least as good as A violates

The Completeness of *At Least as Good As*: For any distributions A and B, either A is at least as good as B, or B is at least as good as A.

Or, equivalently, A is either better than, worse than, or just as good as B. I use "incompleteness" to refer to violations of the completeness of *at least as good as*—i.e., cases in which neither of two things is at least as good as the other. In this section, I define two kinds of incompleteness—incomparability and imprecise equality—along with incommensurability.

Why might $A+$ be neither worse than nor at least as good as A?

On one view, *A* and *A+* are *incomparable*. My height, for example, is incomparable with your intelligence: I am neither taller, shorter, nor as tall as you are intelligent. I discuss the view that mere addition results in incomparability in the next chapter. But this is not Parfit's suggestion. Comparability is, I assume, transitive.[5] Parfit thinks that *Z* is worse than *A*. But if *A+* is comparable with (because worse than) *Z*, and *Z* is comparable with (because worse than) *A*, then *A+* must be comparable with *A*.

Parfit's suggestion is that *A* and *A+* are *imprecisely equally good*. For example, Einstein and Bach might have been imprecisely equally great geniuses: neither was a greater genius than the other, nor were they equally great with respect to genius. If they were equally great with respect to genius, than a slightly improved version of Bach (e.g., who finishes *The Art of the Fugue*) would've been a greater genius than Einstein. But they weren't incomparable with respect to genius, because Einstein was a greater genius than some musical geniuses, and Bach was a greater genius than some scientific geniuses. They were comparable with respect to genius, but only imprecisely so.

Instead of "imprecise equality," Chang (2002) uses "parity": she says that some values are on a par. I use these expressions interchangeably.

Incomparability and parity both involve failures of completeness. How do they relate to incommensurability?

Incommensurability is a relation between values, but "values" can refer either to

[5]This assumption is not usually made. But it makes it easier for me to distinguish incomparability from parity, discussed below. Chang (2002) distinguishes them by appealing to *positive* versus *negative* value relations.

objects' *degrees* of value (i.e., how good they are) or to their *dimensions* of value (i.e., the ways or respects in which they are valuable). The relation between degrees and dimensions is one of determinate to determinable. *Height*, for example, is a dimension along which people and other things vary. Your particular height is a determinate of this determinable. Similarly, *3 grams* is a determinate of the determinable *mass*. I use "values" only to refer to the determinates, not to their determinables, unless otherwise specified. When some determinable property *F* has some ordinal structure—i.e., when some things are *F*er than others—I call *F* a *dimension*.

I use Chang's (2013) definition of incommensurability: "Two values, such as pleasure and fairness, are incommensurable if there is no cardinal scale of value according to which both can be measured" (2595). Chang's examples of pleasure and fairness suggests that she understands incommensurability as a relation between dimensions, rather than degrees, of value. That is how I, too, shall understand incommensurability.

I define a *scale* as a set of scalar values. A value is scalar iff it can be represented by a single real number (e.g., 3 grams). A cardinal scale is one that preserves meaningful differences between values: it allows us to say that *A* is *F*er than *B by more* than *C* is *F*er than *D*. Two dimensions *F* and *G* are incommensurable iff the difference between some quantity of *F* and some quantity of *G* is not a scalar value. For example, the difference between my height and your intelligence is not a scalar value: it is undefined, so they are incomparable. And the difference between Einstein's intelligence and Bach's intelligence is, plausibly, not a scalar value: the values are too imprecisely comparable to be represented by real numbers. But the difference

between my height and the width of my toe is a scalar value: it is some real number of centimeters. Height and width are commensurable because they can be measured on the same cardinal scale. Height and intelligence are incommensurable because they cannot be measured on the same ordinal scale, let alone a cardinal one. Different kinds of intelligence are, plausibly, incommensurable because they cannot be measured on a single cardinal scale.

What is the relation between incommensurability and incompleteness?

First, incompleteness (due to either incomparability or parity) requires incommensurability. If A is not better than, worse than, or just as good as B, then the evaluative difference between them cannot be represented by a single real number. For then it would be either positive (better), negative (worse), or zero (equally good). But if we cannot represent the evaluative difference between A and B with a real number, then we cannot measure their values on a single cardinal scale.

Second, incommensurability is more general than incompleteness. In Chapter 3, I discuss lexical superiority. Like Ross (2002, 150) and Laird (1936, 255), I understand lexical superiority as a relation between incommensurable dimensions of value. If the *A*s are lexically superior to the *B*s, then we cannot represent the evaluative difference between them with a single real number. They cannot fit on a single cardinal scale. (I explain this in more detail in Chapter 3.) But lexical superiority need not violate completeness. Whereas parity involves imprecise comparability, lexical superiority involves *emphatic* comparability (Chang 2014, 116).

I have defined incompleteness, incomparability, parity, and incommensurability. Now back to population ethics.

1.2 Critical-Band Utilitarianism

Blackorby, Bossert, and Donaldson (2005) try to develop Parfit's suggestion that $A+$ is not worse than, nor at least as good as, A. They suggest an interval of non-negative *critical levels*, such that the addition of lives within the interval makes the world neither better nor worse, all things considered, nor precisely equally as good. According to Qizilbash (2007) and Rabinowicz (2009), adding such lives results in parity. More specifically, according to

Critical-Band Utilitarianism: One distribution is at least as good as another iff its critical-level-adjusted sum of wellbeing is at least as great for every level in the interval. That is, for any distributions A and B whose welfare levels are $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$ respectively, A is at least as good as B iff

$$\forall c \in C, \sum_{i=1}^{n}[a_i - c] \geq \sum_{i=1}^{m}[b_i - c],$$

where C is the set of critical levels.

Suppose the critical band extends from 0 to 5, so that the extra lives in the mere addition paradox are within the interval. Critical-band utilitarianism therefore denies the mere addition principle stated on page 8: $A+$ not, on this view, at least as good as A. This is because, for some critical levels (namely, those above 1), $A+$ is not at least as good as A. Critical-band utilitarianism can avoid the repugnant conclusion because, according to any critical level at least as great as the Z-people's wellbeing, Z is not better than A.[6]

[6]Critical-band utilitarianism also avoids what Parfit (manuscript) calls the *callous conclusion*,

This theory's incompleteness is best understood as parity, rather than incomparability, because the interval is bounded. If we interpreted it as incomparability, then comparability would become intransitive. For example,

$$B = (100, \Omega)$$

$$C = (100, 100)$$

$$D = (100, 1)$$

D is comparable with (because worse than) C, which is comparable with (because better than) B. But neither of B and D is at least as good as the other, according to critical-band utilitarianism. So the incompleteness cannot be incomparability if comparability is transitive. It is best understood as parity.[7]

Critical-band utilitarianism, however, implies

The Weak Repugnant Conclusion: For any number of excellent lives, there is some larger number of mediocre lives whose existence would not be worse.

This is because, according to *some* critical level within the interval (namely, any level below the Z-people), A is not better than Z. This could be avoided if some positive welfare levels, including the Z-people's, were below the interval's lower bound.

on which adding a life towards the bottom of the band isn't worse than adding a better life. Critical-band utilitarianism avoids this because, for any critical level, the critical-level-adjusted sum of well-being would be greater if we add the better life rather than the worse one. But adding either life would be neither better nor worse than adding no life at all.

[7] Rabinowicz and Qizilbash provide other reasons to interpret the incompleteness as parity.

Suppose, for example, that the interval's lower bound is 3. On this view, A is better than Z because the Z-people fall below the critical band, so they contribute only negative value. But the view would then imply

The Sadistic Conclusion: When adding people without affecting the original people's welfare, it can be better to add people with negative welfare rather than positive welfare. (Arrhenius 2000)

For example, if the interval's lower bound is 3, then $A+$ becomes worse than

$$E = (100, -1, \Omega, \ldots, \Omega).$$

Critical-band utilitarianism can avoid the sadistic conclusion if the interval extends down to 0. But it nonetheless implies

The Weak Sadistic Conclusion: When adding people without affecting the original people's welfare, it is not always worse to add people with negative welfare rather than positive welfare.

This is because, whenever many people would have positive welfare below some critical level in the interval, it might not be worse, according to that critical level, to add fewer people with negative welfare. So, for some critical level, the sum of critical-level-adjusted wellbeing may be greater in the population of lives that are worth not living.

The weak repugnant and sadistic conclusions are two problems for critical-band utilitarianism. Broome raises further objections to the theory, which raise more general questions about incommensurability in population ethics.

1.3 Arbitrariness

Broome's (2004) first objection to critical-band utilitarianism is that the theory is *ad hoc*. Why should we expect mere addition to yield parity? The most plausible cases of parity involve tradeoffs between incommensurable dimensions of value. Broome considers Sartre's student, who had to choose between fighting in the Second World War and staying home to care for his mother. Many find it plausible that neither of the student's options is at least as good as the other, because there are multiple ideals at stake: he has reasons of honor and patriotism to fight in the war, and reasons of love and care to stay home. He has to weigh these conflicting ideals, but their weights seem imprecise. Even if there were a precise amount by which fighting in the war is better *with respect to honor* and a precise amount by which staying home is better *with respect to love*, there could be no precise amount by which honor is more or less important than love. This makes it plausible that the options are on a par. The parity seems best explained by the imprecise weights of the incommensurable dimensions of value. But, Broome writes,

> our case is not at all like Sartre's. We are not dealing with differing values. One option has a different number of people from the other. Whatever the value of people might be, each option realizes that value; one simply realizes a greater quantity of it than the other. (168)

By "the value of people," Broome means the value of lives worth living, or of wellbeing. Broome's point is that critical-band utilitarianism does not identify a conflict between dimensions of value, and that without such a conflict, we have no reason to expect parity.

In Section 1.3.1, I defend Broome's diagnosis of parity against some objections due to Parfit. In Section 1.3.2, I respond to Rabinowicz's attempt to make critical-band utilitarianism less arbitrary.

1.3.1 The Tradeoff Constraint

Broome's diagnosis is that parity is explained by conflicts between incommensurable dimensions of value. This diagnosis can be generalized beyond imprecise equality in the goodness of outcomes. Consider which of two people is healthier, more creative, more intelligent, nicer, lazier, or more beautiful. These relations, which Chang (2002) calls *covering considerations*, depend on multiple respects or criteria, which I have been calling *dimensions*. Things and people can be good, creative, healthy, intelligent, nice, lazy, or beautiful in some respects and not in others. These covering considerations are *multidimensional*. When there are conflicts between these dimensions, we may get failures of completeness, because the dimensions cannot be precisely weighed against each other. Something like this view has been suggested by many philosophers,[8] economists,[9] and linguists.[10] Chang (2002, 679), too, characterizes her cases of parity as conflicts between imprecisely

[8] See Laird (1936, 255f.), Rashdall (1902, 155–6), Schoenfield (2012, 37), and Rabinowicz (2012, 134, 158).

[9] See Sen (1997), Eliaz and Ok (2006), Ok (2002), von Neumann and Morgenstern (2007, 29).

[10] See Sassoon (2013a, 385), Klein (1980, 22), and Kamp and Sassoon (2015).

weighted aspects of a covering consideration (although she does not say that these are the only possible cases).

This suggests the following desideratum:

The Tradeoff Constraint: If A and B are on a par, then there is some way in which A is better than B and some way in which B is better than A.

Critical-band utilitarianism violates this constraint because adding good lives seems not to make things in any way worse. Critical-band utilitarians might object that each critical level within the band is a dimension of population goodness. On this view, $A+$ is better than A in some respects and worse in others, because it is better according to some critical levels and worse according to others. So $A+$ and A may be on a par. But this claim is *ad hoc*. Being better according to some critical level doesn't seem like a *way* of being better any more than being a better president according to some pundit is a way of being a better president.

Some might try to justify the appeal to parity by identifying conflicting ideals in cases of mere addition. Some pluralists believe that mere addition makes things in some way better—e.g., with respect to total wellbeing—and in other ways worse—e.g., with respect to inequality.[11] Such conflicts might generate parity. If we accept that these are ways in which populations can be better or worse, then we can reconcile the parity of mere addition with the tradeoff constraint.[12] But this view is very different from critical-band utilitarianism. Critical-band utilitarianism appeals only to the sum of (critical-level-adjusted) wellbeing. Inequality and other im-

[11] See Temkin (2012) and Carter (1999).
[12] See Goodrich (2014).

personal ideals are simply irrelevant to critical-band utilitarianism: critical-band utilitarianism predicts cases of parity even when there's no inequality and the universal level of wellbeing is increased. Appealing to such ideals would be a different theory with a different structure. Such a theory might be plausible, but I am not confident that the plurality of ideals needed to give plausible results in population ethics can be integrated in a systematic way. So I do not pursue that option here.

The tradeoff constraint doesn't apply to incomparability. For example, I am neither taller, shorter, nor as tall as you are intelligent, but that's not because I am in some ways taller and in other ways shorter than you are intelligent: degrees of height and of intelligence are incomparable. Moreover, I am neither better nor worse nor just as well off as $\sqrt{2}$. But this isn't because $\sqrt{2}$ is well off in a different way than I am. It's because $\sqrt{2}$ doesn't have a degree of wellbeing. These cases do not involve tradeoffs between imprecisely weighted dimensions of value. In the case of $\sqrt{2}$, we are comparing items along a dimension that one item doesn't instantiate: it has no determinate of that determinable. When comparing my height and your intelligence, however, we have values, but they are determinates of determinables that share no ordinal structure or scale. Comparing them is like asking, "Which is greater: crimson or 4 grams?" Incomparability is not explained by tradeoffs between dimensions of value.

Parfit raises two objections to the tradeoff constraint.[13] First, he suggests that tradeoffs are not required for imprecise *comparability*, and so they may not be required for imprecise equality. I agree that tradeoffs may not be necessary for imprecise comparability, but imprecise comparability is compatible with completeness. Sup-

[13] In correspondence.

pose, for example, that wellbeing is just a function of how much money one has, measured in some privileged currency. More specifically, suppose that wellbeing is a strictly increasing function of money, and only money. And suppose further that money has diminishing marginal contributions to wellbeing. If the factor by which money's marginal contribution to wellbeing diminishes is imprecise, then it may be imprecise how much better one amount of money is than another. There we have imprecise comparability.

Crucially, however, although the imprecise weight of marginal differences in money with respect to wellbeing can generate imprecise comparability, it cannot generate imprecise equality. This is because, for any two amounts of money in some currency, one is either greater than the other or they are precisely equal. If they are precisely equal, then they are precisely equally good. If one is greater than the other, then one may be imprecisely better than the other. But there is no room for imprecise equality. That is how imprecise weights can generate imprecise comparability along a single dimension without generating imprecise equality. This doesn't entail the tradeoff constraint. But it shows that imprecise comparability is not sufficient for imprecise equality. We also need multidimensionality.

Second, Parfit objects that *A* and *B* might be on a par because they are good in different ways, without each being better in some way and worse in another. On Parfit's view, the imprecise equality of Einstein and Bach with respect to genius is not grounded in the fact that Einstein was a greater genius in some ways and a lesser genius in others. Rather, it is grounded in the fact that they are geniuses of such different kinds. Whereas the tradeoff constraint focuses on the different ways in which *A* is better than *B* and *B* is better than *A*, Parfit's suggestion focuses on the

different ways in which *A* is good and *B* is good.

Consider an analogy. We can apply multidimensional adjectives to many different kinds of entities. For example, we can apply *healthy* to to *human*, *girl*, *cat*, *plant*, *savings account*, and *policy*. The kind of entity to which an adjective is applied determines the relevant dimensions. The dimensions that are relevant to the health of a savings account are very different from those for a cat. This may be true, in general, for distinct determinates of a determinable. For example, there is some overlap in the criteria for the health of animals, but the criteria for donkey health are different from the criteria for human health. Asking whether Einstein is a greater genius than Bach may be relevantly like asking whether I am healthier than a donkey. If my donkey Morton is diseased and I am functioning normally, I am plausibly healthier than Morton. But, after curing Morton, it may be true that neither of us is at least as healthy as the other—not because I am in some ways healthier, and in other ways sicker, than Morton, but rather because human health is very different from donkey health.

On the surface, the questions are these:

(1) Was Einstein a greater genius than Bach?

(2) Am I healthier than Morton?

But perhaps we really mean,

(3) Was Einstein a greater scientific genius than Bach was a musical genius?

(4) Am I healthier for a human than Morton is for a donkey?

We might call comparisons like (3) and (4) *interdimensional*, to emphasize the parallel with what linguists call *interadjective* comparisons. Intuitively, they involve imprecise comparability, rather than incomparability, because Einstein was a greater genius than some musical geniuses and I am healthier than some donkeys.

Interdimensional imprecision may also arise in more familiar contexts. Compare the *frisson* one sometimes experiences in response to music with the pleasure of knowing that one has solved a challenging and important puzzle. Experiences of these different kinds may in some cases be imprecisely equally pleasurable, not because each is more pleasurable than the other along some dimension, but rather because they are pleasures of different kinds. Or consider aesthetic comparisons: some landscapes are more beautiful than some paintings, which are more beautiful than some sculptures. The dimensions of beauty here are arguably distinct. This may generate imprecision when comparing the beauty of different kinds of things. And consider interpersonal comparisons of wellbeing: some outcomes are better for some people than they are for others. But even if we could make precise comparisons with respect to a single person's wellbeing, some doubt that different people's wellbeing levels are precisely comparable (Sen 1970b).[14]

I am not claiming that there *is* interdimensional imprecision. I am merely suggesting this as a way of developing Parfit's alternative to tradeoff-based imprecision. But I now wish to claim that the imprecision predicted by critical-band utilitarianism seems not to be justified on interdimensional grounds either: what are the different dimensions along which the relevant distributions are good?

[14] van Rooij (2011) points out the analogy between interadjective comparisons and interpersonal comparisons of wellbeing

I already dismissed two answers to this question as *ad hoc* or inconsistent with critical-band utilitarianism: different critical levels and different impersonal ideals (e.g., total wellbeing and inequality). But two more answers naturally suggest themselves. One answer is that the distributions are good *for different people*. The problem with this suggestion is that it predicts imprecision in the following case:

$$A = (4, \Omega)$$

$$B = (\Omega, 4)$$

Intuitively, these are equally good. We are now considering the view that they are instead on a par.

This view, however, faces the nonidentity problem (Parfit 1984, 359). Suppose we slightly improve the welfare of the second person in B:

$$A = (4, \Omega)$$

$$B' = (\Omega, 5)$$

Intuitively, B' is better and is what we ought to choose. But this makes it unlikely that A and B were on a par. A distinctive feature of parity is insensitivity to small improvements: a slightly improved Bach may be a greater genius than Bach, but not a greater genius than Einstein. We have slightly improved B, but the resulting B' is better than A. If anything better than B, by however little, is better than A, then A and B are not on a par. They are equally good (Broome 2004, 21).

A different kind of interdimensional imprecision has to do with different *numbers* of people. Parfit has suggested that distributions containing different numbers of people are only imprecisely comparable.[15] He suggests that adding people creates a *margin of imprecision*. Unless an extra person's life would be sufficiently good, the value of her life is "swallowed up" (Broome 2004, 170) by the margin of imprecision, which results in imprecise equality. This might justify something like critical-band utilitarianism.

It is not clear, however, *why* populations of different sizes should be imprecisely comparable. It doesn't seem like interdimensional imprecision, because being good for 100 people and being good for 1000 people don't seem like different *ways* of being good. Moreover, if different-sized populations are less precisely comparable by some margin, then this shouldn't depend on how good people's lives are: the goodness of a life can overcome or be swallowed up by the imprecision, but it shouldn't affect the presence or magnitude of imprecision. So there ought to be some *bad* lives whose badness is swallowed up by the margin of imprecision. For example:

$$C = (-1, \Omega)$$

$$D = (-1, -1)$$

My objection is that, if every additional life creates a margin of imprecision that must be overcome, then D is not worse than C. But D is worse than C. So not

[15] In previous versions of Parfit (manuscript). He has since dropped this claim, partly in response to these arguments.

every additional life creates a margin of imprecision that must be overcome. Parfit might respond that the second life in *D* does create a margin of imprecision, but the badness of her life overcomes this margin, and that this is true for any life worth not living. But if the disvalue of a bad life, however small, is enough to overcome the margin of imprecision, then why isn't the value of a good life always enough to do so? Alternatively, Parfit might respond that bad lives do not create a margin of imprecision. But this makes it even more arbitrary to claim that populations containing different numbers of people are less precisely comparable.

There is a more extreme version of this view. Bader's view says that populations containing different numbers of people are not even imprecisely comparable: they are *in*comparable. I discuss this view in Chapter 2.

I conclude that although there may be interdimensional, non-tradeoff-based imprecision, this source of imprecision is not helpful to critical-band utilitarianism.

1.3.2 Rabinowicz's Responses

Rabinowicz (2009, 405) tries to meet Broome's arbitrariness objection by drawing on his (2008) definition of parity. On this definition, two outcomes are on a par iff it is permissible to prefer either outcome to the other. Rabinowicz then claims that the parity of mere addition is explained by the permissibility of preferring either the larger population or the smaller population to the other. But *why* it is permissible to prefer either world to the other? One plausible reason why it might be permissible to prefer either of two outcomes to the other is that one has some reason to have each preference, and neither set of reasons outweighs the other. That

is how Rabinowicz himself characterizes parity at times (2012, 134, 158). But there appears to be no conflict in the balance of reasons in cases of mere addition: what reason do we have to prefer the smaller distribution to the larger one? Rabinowicz claims that the permissible preference orderings "correspond to different choices of subintervals within the neutral range" (2009, 405). But this is a merely formal point: Rabinowicz does not explain what *makes* these different choices permissible. So Broome's request for an explanation has not been met.

Rabinowicz (2009, 406) suggests another way of meeting the arbitrariness objection. He suggests that the critical band can be interpreted as a *personal* neutral range—i.e., a range such that a person's existence within the range is neither better nor worse for her than nonexistence. Life above this range is better for the person than nonexistence and, therefore, worth living; life below it is worse for the person than nonexistence and, therefore, worth not living. Adding lives worth living always makes things better, and adding lives worth not living always makes things worse. Adding lives that are neither worth living nor worth not living—i.e., in the personal neutral range—makes things neither better nor worse, nor equally good. This view is formally just like critical-band utilitarianism, but it reinterprets the band's lower bound as the point below which life is worse than nonexistence and the upper bound as the point above which life is better than nonexistence.

This interpretation of the critical band could explain why the addition of lives within the band results in parity: the lives themselves are neither better than, worse than, nor as good as nonexistence. But Broome's request for an explanation could arise there too. Why think that some lives are not better than, worse than, or just as good as nonexistence? The most natural reason to think this is because lives are

incomparable in personal value with nonexistence: nonexistence has no personal value, in the same way that $\sqrt{2}$ has no height or color. That is not *ad hoc*. However, this would make comparability intransitive. Nonexistence is comparable with (because better than) lives below the band, which are comparable with (because worse than) lives within the band. So the transitivity of comparability requires nonexistence to be comparable with lives within the band, contrary to the suggestion we are considering.

Rabinowicz's view is instead, I think, that many lives are *on a par* with nonexistence. This might seem arbitrary. But there may be a good justification for it. On Rabinowicz's view, a life is on a par with nonexistence iff it is permissible to prefer either the life or nonexistence for the person's sake. This most plausibly obtains when lives contain both good things and bad things. There may be no single quantity of happiness, for example, that precisely compensates for any quantity of suffering. This is incoherent on some views about quantities of happiness and quantities of suffering, on which such quantities are defined only as values on the same cardinal scale of wellbeing (Griffin 1979). But if we understand such quantities independently of evaluative comparisons, we might wonder how much happiness is needed to compensate for some quantity of suffering. According to psychologists Fredrickson and Losada (2005), the answer is 2.9013. That is their "critical positivity ratio" that separates "flourishing" from "languishing" lives. But that's absurd. If the question even makes sense, the answer is not some real number. Rabinowicz might say that there is a range of permissible tradeoff ratios between goods and bads, such that some lives are better than nonexistence according to some tradeoff ratios but not according to others. This might explain why such lives are on a par with nonexis-

tence.

Rabinowicz claims that his view avoids the repugnant conclusion. But it doesn't. A life that is *barely worth living* is one that, on this view, is barely better for the person than nonexistence. Rabinowicz's view entails that, for any number of excellent lives, there is some larger number of people whose existence would be better, even though their lives are *barely* better for them than their nonexistence. Rabinowicz might be understanding the repugnant conclusion differently—e.g., as the claim that, for any number of excellent lives, there is some much larger number of people whose existence would be better, even though their lives would barely be better than lives that are worth *not* living (406).[16] Rabinowicz avoids this conclusion because he thinks that lives worth living cannot be *barely* better than lives worth not living: such lives are separated by the critical band, which is wide enough that no lives above it are just barely better than lives below it.

This claim's plausibility, however, depends on why certain lives are merely on a par with nonexistence. Consider a psychologically simple creature—e.g., a lizard—whose life contains only good things. It might contain, for example, a stream of mild pleasure for two minutes. This life seems to me barely worth living, and indeed barely better than a life that contains a stream of mild discomfort for two minutes. That some number of positive lizard lives could be better than some number of excellent lives strikes me as repugnant. To avoid this repugnant conclusion, Rabinowicz would have to claim that the happy lizards' lives are not worth living: they are on a par with nonexistence. But if we explain the parity of some lives with nonexistence by appealing to permissible tradeoffs between good things and

[16] But see his standard statement of the repugnant conclusion on p. 398.

bad things, we cannot say that the lives in question are on a par with nonexistence: they contain no bad things. Rabinowicz might say that such lives are in some ways bad—e.g., because they are boring. On this view, it may be permissible to prefer nonexistence to a lizard life of good feelings for this creature's own sake, because it is so boring. But why? Nonexistence isn't any more exciting. Alternatively, Rabinowicz could reject the tradeoff constraint, in which case Broome's arbitrariness objection rears it head again. We would need some explanation of why it is permissible to prefer nonexistence to a mildly happy life for the being's own sake. Without such an explanation, Rabinowicz's interpretation makes little progress.

Moreover, in addition to implying the standard repugnant conclusion, Rabinowicz's view implies other counterintuitive conclusions. We saw on pages 14–15 that critical-band utilitarianism implies weak versions of the repugnant and sadistic conclusions. Rabinowicz's view simply reinterprets those, but in ways that are even less plausible. Its analogue of the weak repugnant conclusion is that, for any number of excellent lives, there is some much larger number of people whose existence would not be worse, even though none of their lives would be worth living. (It implies this because lives within the band are not, for Rabinowicz, worth living.) Its analogue of the weak sadistic conclusion is that, when adding people without affecting the original people's welfare, it is not always worse to add people whose lives are much worse than nonexistence rather than people whose lives are not worse than nonexistence. (It implies this because lives within the band are not worse than nonexistence.) Rabinowicz's personal neutral range may help to avoid the arbitrariness objection and an idiosyncratic version of the repugnant conclusion, but it implies the standard repugnant conclusion and other implausible conclusions.

I conclude that Broome's arbitrariness objection succeeds. I now turn to Broome's vagueness objection, which I think fails.

1.4 Vagueness

Broome argues that parity is incompatible with vagueness. Completeness asserts a disjunction (stated on page 9), and parity implies that the disjuncts are false: *A* is neither better than, worse than, nor just as good as *B*. Broome, by contrast, accepts the disjunction that *A* is better than, worse than, or as good as *B*. But he thinks it's often vague which disjunct holds.

Broome claims that *better* is obviously vague. And critical-band utilitarianism should predict some vagueness: the upper bound of the critical band, for example, is vague. It would be hard to believe that adding any life of some quality results in parity, but that adding any slightly better life makes things better. Say that a life is *borderline* if it is vague whether the life is within the critical band or above it. It's vague whether adding a borderline life makes things better or merely imprecisely as good. But, according to critical-band utilitarianism, it's determinate that adding a borderline life doesn't make things worse. Broome claims that this combination—that adding the life would determinately not make things worse, but that it's vague whether it makes things better—is impossible.

This claim follows from his

Collapsing Principle: "For any predicate *F* and any things *A* and *B*, if we can deny that *B* is *F*er than *A*, but we cannot deny that *A* is *F*er than *B*, then *A* is *F*er

than *B*" (2004, 174).

Let's apply the collapsing principle to critical-band utilitarianism. Let *A* be a population with the borderline life, and let *B* be that same population without this life. The critical-band utilitarian denies that *B* is better than *A*. And he refuses to deny that *A* is better than *B*, because it's vague whether the life is above the critical band. But, in violation of the collapsing principle, he refuses to assert that *A* is better than *B*. Because critical-band utilitarianism violates the collapsing principle, Broome concludes that it is false, and indeed that *any* appeal to incompleteness is incompatible with the vagueness of *better*.

Broome's alternative theory is formally just like critical-band utilitarianism, except he interprets the critical band as a zone of vagueness. There is, on Broome's view, a single critical level, but it's vague which level it is. One distribution is at least as good another iff it's at least as good according to every admissible precisification of the critical level. Suppose that all levels from 0 to 5 are admissible precisifications of the critical level. This means, in the mere addition paradox, that $A+$ is either better than, worse than, or just as good as *A*, but we cannot say which.

In Section 1.4.1, I argue that the collapsing principle is false. In Section 1.4.2, I challenge Broome's assumption that *better* is *obviously* vague. I argue that it is vague only if it is multidimensional, and that this is bad news for Broome's own theory.

1.4.1 Counterexamples

In this section, I present new counterexamples to the collapsing principle. My goal here is to show that the principle is threatened from many different angles, so I aim for breadth rather than depth.

Consider first absolute adjectives, such as *flat*, *certain*, or *empty*. Suppose we can assert, in some context, that some table is flat. We can deny, within that context, that the pavement is flatter than the table: if x is flat (within a context), then y isn't flatter than x (within that context). But it might be vague whether the pavement is flat. And so we may be unable to assert that the table is flatter than the pavement. But we can deny that the pavement is flatter than the table, because the table is flat.

Consider next phenomenal predicates, such as *painful*. I cannot tell whether some sensation is painful. The sensation becomes slightly more intense: *if* the first was painful, then the second was more so. But I can't tell whether the second was more painful, because it might not have been painful at all. (I find this plausible for some buzzings and musical tones.) The second sensation definitely wasn't *less* painful. We can deny that the first felt more painful than the second, but the second might not be more painful than the second: it might not be painful at all.

Further problems arise with degree modifiers, such as *slightly* or *rather*. Let F be the predicate *slightly sweet*. Take two cups of coffee, call them A and B, and add some sugar to A. Add just enough so that it's borderline whether A is slightly sweeter than B.[17] You can deny that B is slightly sweeter than A—it has less sugar—but you can

[17] It might be on the borderline between slightly sweeter and not sweeter at all, or on the borderline between slightly sweeter and much sweeter.

neither assert nor deny that *A* is slightly sweeter than *B*. Or consider more precise measure phrases, such as *four centimeters tall*. Suppose that I'm shorter than you, but that the difference between our heights is vague: it hovers somewhere around four centimeters. We can assert that I'm not four centimeters taller than you, and we can't deny that you're four centimeters taller than me, but you might not be four centimeters taller than me: you might be taller by slightly less than that.

Other complex comparative constructions pose additional trouble. Consider conjunctive comparisons, such as *bigger and better*. Consider two populations, *A* and *B*, where *B* is bigger than *A*. On Broome's population axiology, it may be vague whether *B* is better than *A*; let that be so. Broome would deny that *A* is bigger and better than *B*, because *B* is bigger than *A*. And Broome could not deny that *B* is bigger and better than *A*, because it is vague whether *B* is better than *A*. But Broome could not assert, by his own lights, that *B* is bigger and better than *A*. Broome could simply exclude such comparisons from his collapsing principle. But some linguists argue that *better* is implicitly conjunctive over multiple dimensions (Sassoon 2013b). If these linguists are right, then excluding conjunctive comparisons from the collapsing principle would thereby exclude *better*.

More interesting examples are due to vague identities. Suppose, with Lewis (1988), that it's vague whether *Princeton = Princeton Borough*: our use does not settle whether "Princeton" refers to the Borough or to the Borough plus the Township or one of many larger areas. But we know that Princeton is at least as large as Princeton Borough. So we can deny that Princeton Borough is larger than Princeton, although we cannot deny that Princeton is larger than Princeton Borough. The collapsing principle implies that Princeton is therefore larger than Princeton Bor-

ough. But then Lewis's identity statement would be determinately false, because nothing is larger than itself.

Vague identities pose additional problems. Suppose that someone undergoes one of Parfit's (1984) spectrum operations: we tweak her body and mind so that we cannot say whether the resulting person is identical to the person we started with. Then it's vague whether the operation increases the number of people who ever exist. There is someone before the operation and someone after the operation, but it's vague whether they are the same someone. So it's vague whether there are two people where there would've otherwise been one. Let A be the world's timeless population—i.e., all those who ever live—had the spectrum operation not occurred, and let B be the population given that the operation occurs. If the operation preserved identity, then A and B are equally big: the same number of people ever exist. But if the operation didn't preserve identity, then B is bigger than A. We can deny that A is bigger than B, but since we cannot say whether the operation preserved identity, we cannot deny that B is bigger than A. Nonetheless, B might not be bigger than A. This violates the collapsing principle.

Broome might avoid these problems by restricting the collapsing principle to comparisons between determinately distinct items. But similar problems remain, having to do with vague spatiotemporal boundaries of objects. It may be vague whether I am taller than my identical twin because it's vague whether some hair curl counts as part of my head. If the curl counts as part of my head, then I am taller, but if not, we're equally tall. So we can deny that my twin is taller than me, but can neither assert nor deny that I am taller than him. The vagueness here has nothing to do with whether I am my twin: I'm determinately not.

1.4.2 Multidimensional Vagueness

I'm sure that for each of my counterexamples, there is someone who would reject it. But it seems to me that at least some of the counterexamples are genuine. The collapsing principle looks hard to defend. We could try to revise the principle to avoid the counterexamples. But they don't seem to have much in common. Why should the collapsing principle hold for *better* but not for absolute predicates, phenomenal predicates, degree-modified predicates, conjunctive predicates, or comparisons involving vague identities or boundaries?

Constantinescu (2012) suggests a unified revision.[18] He distinguishes between *genuine* and *derivative* comparative vagueness. The vagueness in the counterexamples is, in some sense, not genuinely comparative. They have to do with how little curvature is required to say that something is flat, our inability to tell whether some sensation is painful, the vagueness of degree modifiers, the semantics of conjunctive comparisons, and the vagueness of singular terms, personal identity, and spatiotemporal boundaries. Those phenomena are somehow different from, and have less to do with comparison than, the vagueness of *better*. The trouble is to say what *exactly* is different about *better*: why is *better* vague? Constantinescu suggests that the vagueness of some comparison is genuinely comparative iff "it is not determinate how the contributory values of [two things] should be ranked. In an important sense, then, this type of vagueness belongs entirely in the realm of comparisons" (12). And, says Constantinescu, the collapsing principle holds only for *this* type of vagueness. We can formulate Constantinescu's view as follows:

[18] In response to Carlson (2004). Broome (2009) finds Carlson's counterexample unconvincing. But Broome's response does not apply to my counterexamples.

Multidimensional Collapsing Principle: For any predicate *F* and any things *A* and *B*, if we can deny that *B* is *F*er than *A*, but, because the relative weights of *F*'s dimensions are vague, we cannot deny that *A* is *F*er than *B*, then *A* is *F*er than *B*.

If Constantinescu is right, then Broome's argument succeeds only if the vagueness of *better* is genuinely comparative in this sense—that is, multidimensional vagueness. This is important because it limits the scope of Broome's objection: we needn't worry about it unless our incomplete axiology is multidimensional.[19]

I do not endorse the multidimensional collapsing principle. But the only counterexamples I can think of are cases of parity, which would beg the question in this context. And I think the principle has something going for it: why should conflicts between imprecisely weighted dimensions generate both parity *and* vagueness? This is a question, not an argument. I hesitate to accept or reject the principle. I am, though, inclined to agree that the only genuinely comparative vagueness is multidimensional vagueness.

Broome (2007) sees no need to provide an explanation for why *better* is vague, on the grounds that almost every predicate in natural language is vague. But some linguists and philosophers of language hold that adjectives only exhibit vagueness in the positive form—e.g., *good*, *tall*, and *bald*.[20] And others hold that unidimensional predicates are never vague in the comparative form (Egre and Klinedinst 2010). The best explanation of why *better* is vague seems to me that it is multidi-

[19] Alternatively, we might worry that no appeal to parity can meet both the multidimensional collapsing principle and the tradeoff constraint. I discuss this worry on page 112.

[20] See Cooper (1995), Kennedy (2007), and Kennedy (2011).

mensional. When it is vague which of two things is better, that is because each is better than the other in some ways but not in others. The weights of these dimensions may be imprecise, and this generates vagueness. This is the same story that Broome tells about parity, but applied to vagueness.

Broome (1999, 123), however, argues that unidimensional comparisons can be vague. It can be vague, for example, which of two color patches is *redder*. Redness, Broome thinks, is unidimensional. But it's not: redness depends on hue, brightness, and saturation (Hyde 2012, 17). It's often vague which of two patches is redder because they differ along these dimensions. Broome might say that it could be vague which of two patches is redder even if they differ only with respect to hue. But these patches could, in principle, be placed on a color wheel, and the one closer to 0° hue would be redder.[21]

Broome (2007, 555f.) offers another example. He asks us to compare a range of sauvignon blancs with a moderately good chardonnay. Suppose that we cannot distinguish between each wine in the range, but that our best sauvignon blancs are better, and our worst sauvignon blancs are worse, than than our chardonnay. Broome infers that because there is no sharp borderline between which wines are better and which are not better than the chardonnay, *better* must be vague.

There is a source of noise in Broome's example, which makes it hard to identify exactly what is vague and why. When Broome stipulates that we cannot distinguish between the skillfully blended wines, this makes it hard to tell whether the result-

[21] Broome might respond that we cannot detect which patch is closer to 0° hue. But Raffman (2011) shows that we can detect such differences, although we often don't know when we detect them.

ing vagueness is due to the vagueness of *better*. We need to distinguish between the vagueness of *better* and the vagueness of *tastes* and other observational or phenomenal predicates.[22] To control for issues of perceptual discrimination, we should imagine that we can distinguish between how the wines taste, and that we have sharp preferences between sauvignon blancs. If it is still vague which sauvignon blancs are better than the chardonnay, then that vagueness seems multidimensional. Wine quality depends on many different factors, and that is why Broome's example requires wines of different grapes, which are good in different ways. I cannot see why else we would have sharp preferences over wines of one grape but not over wines of different grapes.

I conclude that Broome's examples of comparative vagueness are cases of multidimensional vagueness. They are, therefore, consistent with the hypothesis that if *better* is vague, this is because of multidimensionality.

This is a problem for Broome's axiology, because the vagueness in Broome's axiology is not multidimensional. It has no more to do with tradeoffs between imprecisely weighted dimensions of value than critical-band utilitarianism. Broome would say that he predicts vagueness when and because adding lives is better according to some precisifications of the critical level, but not according to others. But this is no less *ad hoc* than the critical-band utilitarian's appeal to different critical levels as distinct dimensions of value, which we considered on page 18. It looks nothing like the multidimensional vagueness that we find in other adjectives just

[22] For example, we may be unable to say which of two massive numbers *seems smaller*, because they are represented in complicated ways. But we can't infer from this that *smaller number* is vague: it is a paradigm of precision. The vagueness, if there is any, is in how the numbers *seem*, not in how small they are.

like *better*.[23] Broome's arbitrariness objection applies to his own view.

1.5 Conclusion

I began this chapter with a version of the mere addition paradox. Parfit suggests that adding lives worth living doesn't make things worse, nor does it make things at least as good. This violates completeness. I introduced critical-band utilitarianism, which tries to develop Parfit's suggestion, and interpreted the critical band as a zone of parity. I then agreed with Broome that the appeal to parity is arbitrary: mere addition seems not to involve a conflict between incommensurable dimensions of value. Critical-band utilitarianism, therefore, violates the tradeoff constraint stated on page 18. There might be exceptions to the tradeoff constraint—e.g., interdimensional imprecision—but they do not seem promising for population axiology. I found Rabinowicz's attempts to justify critical-band utilitarianism unconvincing.

I then considered Broome's vagueness objection. I argued that the collapsing principle is subject to a wide range of counterexamples. But it might survive in a more limited form, if the multidimensional collapsing principle is true. Constantinescu's restriction of the collapsing principle to multidimensional vagueness is motivated by the thought that comparatives are usually vague when and because the weights of their dimensions are vague. I have no argument against the multidimensional collapsing principle, so I hesitate to reject it. It may be hard to appeal to parity in a way that meets both the tradeoff constraint and the multidimensional collapsing principle; I suggest a view that might do this in Section 3.2. But I argued that

[23] See Sassoon (2013a) and Grinsell (2013).

Broome's vague axiology is also *ad hoc*: we should expect vague betterness, like parity, when there are tradeoffs between dimensions of value.[24]

In Chapter 2, I discuss a view that avoids Broome's objections. It is what critical-band utilitarianism would look like if the critical band was unbounded above and below. On Bader's view, mere addition results in incomparability, rather than parity.

[24] You might wonder why I haven't addressed Broome's *greediness* objection, which he identifies as his most serious problem for critical-band utilitarianism. The reason is that I think it brings up the same issue as the arbitrariness objection, but in a less pure form. The greediness objection trades on an ambiguity in what Broome calls "a neutral thing" (170): if mere addition is good in some ways and bad in others, it might be neutral all things considered, but the ways in which it is good or bad may "swallow up" (198) other values. So if the tradeoff constraint can be met, then so can the greediness objection. This is clearest in Rabinowicz's response to the greediness objection: he accepts that adding people "has a value that counts against other values" (203). This avoids the greediness objection. But what is that value? That's the arbitrariness objection. The greediness objection also raises important practical difficulties, to which I have no solution.

2 Incomparability and Average Utilitarianism

Bader rejects the mere addition principle because he is, like Narveson (1973, 80), "in favour of making people happy, but neutral about making happy people." This chapter is about Bader's attempt to explain the neutrality of creating happy people in terms of incomparability. In Section 2.1, I summarize Bader's view. The view has many moving parts, some of which I skip and most of which I oversimplify. After summarizing the view, I present two objections to it. I then claim that average utilitarianism can avoid these objections in a way that captures some of the virtues of Bader's view. In Section 2.2, I argue that average utilitarianism can elegantly capture the neutrality of creating happy people. In Section 2.3, I argue that average utilitarianism is impartial without being impersonal, and that it gives intuitively correct verdicts in infinite cases. In Section 2.4, I suggest a way of dealing with the most counterintuitive implications of average utilitarianism.

2.1 Bader's View

The core of Bader's population axiology is

The Equinumerosity Constraint: If two distributions contain different numbers of people, neither is better than the other, nor are they equally good: they are incomparable.[25]

[25] Bader instead uses "non-comparable" to emphasize a formal, rather than substantive, failure of the betterness relation.

Bader argues that the equinumerosity constraint follows from the fact that lives are incomparable with nonexistence for the people living them, along with the reducibility of betterness to betterness for people. Although Bader prohibits comparisons between different-sized populations, he—unlike Heyd (1988) and Morton (1994)—allows comparisons between disjoint populations of the same size. He accepts

Strong Impartiality: For any distributions A and B containing the same number of people, if A and B are permutations of each other, then they are equally good.[26]

Bader's view, therefore, avoids the nonidentity problem. It is better to bring a better life rather than a worse life into existence, even if they would be lived by different people. Moreover, Bader allows for other comparisons between populations of the same size. He appeals to a method of

Balancing Gains and Losses: A is at least as good as B iff there is a bijection $f : A \to B$ such that $\sum_{a \in A} [a - f(a)] \geq 0$.

A bijection from A to B is just a function that pairs every member of A with just one member of B, and pairs every member B with just one member of A. It is a one-to-one correspondence between A and B.[27] Balancing says that which of two

[26] Two distinct distributions are permutations of each other iff all and only same wellbeing levels exist in both, but they are rearranged among people.

[27] The identity-preserving bijection, in which each person is paired with herself, is a special case. For Bader, it doesn't matter which bijection we use. In infinite cases, he requires distributions to be better according to *every* bijection.

distributions is better depends not on the sum of people's wellbeing in each distribution, but rather on the sum of the *differences* between people's wellbeing across the distributions. It therefore avoids "conflating all persons into one" (Rawls 1999, 24) or caring about "utilities that are not someone's" (Bennett 1978, 63–64). In finite, fixed-population cases, this method coincides with total and average utilitarianism. But it entails the equinumerosity constraint. For if A and B contain different numbers of people, then there can be no bijection between them, so neither is at least as good as the other.

The equinumerosity constraint implies that the repugnant conclusion is false, and that mere addition does not make things better. But it also implies the weak repugnant conclusion stated on page 14: Z is not worse than A. However, Bader introduces further machinery that softens this conclusion. He allows for a kind of dominance reasoning for preferring and choosing one distribution rather than another. If A is smaller than B, then we can compare A with each A-sized subset of B. If every A-sized subset of B is at least as good as A according to balancing, and some such subset is better, then B *dominates* A. Similarly, if A is at least as good as every A-sized subset of B according to balancing, and A is better than some such subset, then A dominates B.

Bader then uses this dominance reasoning to help us choose between distributions. When choosing between various distributions, we first separate the smallest distribution(s) from all the distributions that are bigger than it. (The smallest distribution might contain no one.) We have no reason to choose a minimal distribution (i.e., one of the smallest ones) rather than a bigger one, or vice versa. But, if we choose a bigger one, then we should not choose one that is *dominated* by another.

We might call this a *conditional choiceworthiness* relation: a non-minimal distribution that is dominated by another other non-minimal distribution is less choiceworthy, conditional on choosing a non-minimal distribution. But non-minimal distributions are neither more nor less choiceworthy than minimal ones. And between all the minimal ones, the better ones are more choiceworthy. We ought not to choose an option that is less choiceworthy than some available alternative.

Suppose that we must choose between an empty distribution, a quality distribution (A), and a quantity distribution (Z). Bader says that we have no more or less reason to choose A or Z than we have to choose the empty one. But, conditional on choosing either of the non-empty distributions, we shouldn't choose Z. Suppose next that we must choose between A and Z, with no empty option. In this case, Bader's choiceworthiness relation is silent, because it does not compare minimal with non-minimal distributions. But, if we have several intermediate options—e.g., B, C, etc.—it becomes wrong to chose Z, although permissible to choose either A or B.

Finally, Bader introduces deontic constraints against adding lives that are worth not living. According to the procreative asymmetry, the fact that someone's life would be worth living is not (in itself) a moral reason to create her, but the fact that someone's life would be worth not living is (in itself) a moral reason not to create her (McMahan 1981). Bader has a compelling explanation of the second half of the asymmetry. Roughly, miserable lives are such that, if we add them, we would have reason to shorten them as much as possible so that, in the limit, they would not exist. But we have reason not to do things that, if we do them, we would have reason to undo them. So, although Bader's method of balancing is silent regarding the addition of miserable lives, and although such lives are not worse for people than

nonexistence, the choiceworthiness relation says not to add them. The constraint against adding such lives trumps the conditional choiceworthiness of choosing a dominant distribution.[28] However, it can be permissible to add miserable lives if the positive externalities (for either existing people or others who are added) are sufficiently great. Adding some lives of total misery $-H$ is permissible iff doing so increases the total amount of happiness by at least as much as H.[29]

Bader's theory avoids the repugnant conclusion, solves the nonidentity problem, captures the neutrality intuition, and explains the procreative asymmetry. It is also not threatened by Broome's arbitrariness or vagueness objections. I observed on page 19 that incomparability is not subject to the tradeoff constraint. What about vagueness? I suggested on page 34 that it can be vague how many people exist. This might have been a problem for Bader if Broome's original collapsing principle were true. But the most plausible collapsing principle is something like the multidimensional one stated on page 36, which applies only when the weights of conflicting dimensions are vague. Bader's view appeals to no such weights. If it is ever vague whether A is better than or incomparable with B, the vagueness will not be due to imprecisely weighted dimensions of value.

I have two objections to Bader's view.

First, Bader's view implies

The Weak Deontic Very Repugnant Conclusion: For any number of excellent

[28] This is not, strictly speaking, part of Bader's view. But I think it makes it more plausible in cases where one chooses to violate the deontic constraint against adding such lives and then must choose between distributions that violate this constraint, some of which dominate others.

[29] Bader's justification for this is more elegant than a brute appeal to sums of happiness. It has to do with shortenings of the positive-externality lives.

> lives and any number of horrible lives, there is some number of mediocre lives whose existence, along with the horrible lives, would be permissible to bring about, when the only alternative is to create only the excellent lives instead.[30]

For example, suppose that we must choose between two worlds. In *A*, there are ten billion people with excellent lives. In *Z*, these same people and 90 billion others suffer great agony for fifty years. They suffer so that *n* other people can exist with lives that are barely worth living. The weak deontic very repugnant conclusion implies that, for some *n*, it is permissible to choose *Z*.

Bader's view implies the weak deontic very repugnant conclusion because the choiceworthiness relation is silent when deciding whether to add people: *A* is the minimal distribution, and the question is whether to expand the population to *Z*. Although *A* dominates *Z*, Bader says that non-minimal distributions are neither more nor less choiceworthy than minimal ones. And the deontic constraint against adding miserable lives is not violated in choosing *Z*: the positive externalities for *n* people make it permissible to bring about the hundred billion. So Bader's view implies the weak deontic very repugnant conclusion.

Bader might appeal to additional deontic constraints against choosing *Z* in some circumstances. We might, for example, be using some people merely as means to our end of expanding the population. But suppose that the agent doesn't know whether the populations share any members, doesn't care whether there are more or fewer people, and, under the influence of Bader's theory, sees no reason to

[30] See Arrhenius (forthcoming) for the very repugnant conclusion. This is a "weak deontic" version because it claims that adding the worse lives is permissible.

choose either distribution. It seems that, if she plumps for Z, she doesn't thereby use anyone merely as a means to any end, although she does act wrongly.

Bader might simply bite the bullet of the weak deontic very repugnant conclusion. But Bader's view is subject to a particularly counterintuitive version of the *Egyptology objection*.[31] Suppose that I could have either zero, two, or three children. None of them would have further children, and my choice wouldn't affect anyone else's wellbeing in any other way. They would all have equally great lives at, say, 10. I know, from reading my Bader, that it doesn't matter morally whether I have children or not. But, if I have children, then I should have whichever number of children would bring about a dominant distribution. I know the welfare level of everyone who has ever existed or will ever exist—except for the ancient Egyptians. Suppose that the ancient Egyptians were either all at 9 or all at 11, and that everyone else who has ever lived is at 10. If they were at level 9, then the distribution where I have three children dominates the distribution where I have two children. They look like these:

$$A = (9, 9, \ldots, 9, 10, 10, \ldots, 10, \Omega, \Omega, \Omega)$$

$$B = (9, 9, \ldots, 9, 10, 10, \ldots, 10, 10, 10, \Omega)$$

$$C = (9, 9, \ldots, 9, 10, 10, \ldots, 10, 10, 10, 10)$$

The ancient Egyptians are the first people in each list. Everyone who comes after them is better off. I am the last person in A, and the next three slots are those of my

[31] See Parfit (1984, 120) and McMahan (1981).

children. Every B-sized subset of C is at least as good as B according to balancing, and some subsets (any including my three children) are better. So C dominates B. Therefore, conditional on my having children, I should have three children rather than two. But if the ancient Egyptians were better off, then the distributions look like these:

$$A' = (11, 11, \ldots, 11, 10, 10, \ldots, 10, \Omega, \Omega, \Omega)$$

$$B' = (11, 11, \ldots, 11, 10, 10, \ldots, 10, 10, 10, \Omega)$$

$$C' = (11, 11, \ldots, 11, 10, 10, \ldots, 10, 10, 10, 10)$$

Now B' dominates C'. B' is at least as good as every B'-sized subset of C', and better than some such subsets (any including my three children). Therefore, conditional on my having children, I should have two children rather than three.

Bader's choiceworthiness relation is supposed to tell me what I ought to choose. It doesn't say that having children is more choiceworthy than not having children, or vice versa: I can permissibly either have children or remain childless. But if the ancient Egyptians lived one way, then I shouldn't have two children. If they lived another way, then I shouldn't have three. The number of children that I can permissibly have depends on the welfare of the ancient Egyptians. That is absurd. Call this the *deontic* Egyptology objection.

On some population axiologies, the contributive value of my having two rather than three children—i.e., their impact on the goodness of the world—can depend

on the welfare of the ancient Egyptians. That is the *axiological Egyptology objection*, and it too seems absurd. But, as Carlson (1998) points out, it is less absurd than the deontic version. Our primary intuition is that the welfare of the ancient Egyptians cannot affect whether or how many children you *ought* to have. Bader's view, however, violates this deontic intuition. On Bader's view, it is neither better nor worse for me to have two children rather than three, no matter the welfare of the ancient Egyptians. But, depending on how the ancient Egyptians lived, I either ought not to have two or ought not to have three, even though neither choice would make things worse. I cannot see how that could possibly be true.

Bader might respond by appealing to what he calls *fine-grained* expansion conditions—e.g., expanding the population to size n, rather than merely expanding the population. He says that we can always permissibly choose there to be n people (assuming no violation of the asymmetry). Our conduct is then guided by the choiceworthiness relation whose condition is expanding to size n. But if we merely choose to expand the population, without a specific number in mind, our conduct is guided by the choiceworthiness relation whose condition is expanding *simpliciter*. In the present case, therefore, it is permissible for me to choose to have any number of children I'd like. But if I merely choose to have children, without a number in mind, I ought not to have either two or three children, depending on how the ancient Egyptians lived. This strikes me as no more plausible than the original verdict. It also makes Bader's response to the repugnant conclusion less satisfying: we make permissibly choose Z just by wanting or deciding that there be as many people as possible.

Bader could avoid my objection by restricting his evaluation to distributions of

welfare over the people affected by our choice. But the view is already quite complicated. I explain in Section 2.2 how, if Bader makes this kind of move, he might as well be an average utilitarian.

2.2 Deontic Neutrality

Bader can avoid the deontic Egyptology objection by accepting

Independence of the Unaffected: What one has most (welfare-based) reason to do in some choice situation depends only on the distributions over people whose welfare is affected by our choice.

Who are these people? If some people's wellbeing would remain the same whatever we do (e.g., the ancient Egyptians), then they are not affected by what we do, so they are excluded from the distributions of welfare whose value (or choiceworthiness) determines what we ought to do. If their wellbeing would be greater or less, then they are included in the relevant distributions. If they might not exist, depending on what we do, then we also include them: their happiness is affected, although not for good or ill.[32] I say that people are affected in a choice situation iff their welfare differs in at least some of the distributions available.

Suppose we accept independence of the unaffected. Bader's view can then avoid the deontic Egyptology objection. But so can average utilitarianism. According to average utilitarianism, A is better than B iff the A-people are, on average, better

[32] We might, alternatively, exclude them if there is a welfare-preserving bijection between some lives in different distributions.

off than the B-people.[33] Average utilitarianism says, in the Egyptology cases above, that C is better than B, which is better than A, and that C' is worse than B', which is worse than A'. But the subdistributions over those affected by our choice are these:

$$A'' = (\Omega, \Omega, \Omega)$$

$$B'' = (10, 10, \Omega)$$

$$C'' = (10, 10, 10)$$

We can ignore the welfare of the ancient Egyptians because our choice does not affect them. The only people affected are my children. The average wellbeing in A'' is undefined, so it is incomparable with B'' and C'', which are as good as each other. So we have no more reason to choose any one of these subdistributions: I can permissibly have zero, two, or three children.

Average utilitarianism still implies the axiological Egyptology objection, but I agree with Carlson that the intuition is primarily deontic. Broome (194) objects that he cannot understand how goodness, in such cases, could be irrelevant to what we ought to do. But, according to independence of the unaffected, goodness is still relevant to what we ought to do: the goodness of the lives affected by our choice.

Average utilitarianism along with independence of the unaffected implies

Strong Deontic Neutrality: We have no (welfare-based) reason to add or not to

[33] By "average," I mean the arithmetic mean. By "the A-people," I mean everyone who ever lives in A. And by "better off," I mean higher lifetime wellbeing. The view says nothing about aggregation within lives.

add good lives in a way that affects no one else.

We saw above that A'' is incomparable with B'' and C''. This is because the average wellbeing of an empty distribution is undefined. And, in any case of mere addition, the relevant subdistribution in which we don't add lives is empty. We only acquire reasons to expand the population when doing so would affect other people.

I think that average utilitarianism captures the best of Bader's view because, if we accept independence of the unaffected, average utilitarianism provides a simpler account of the neutrality of creating happy lives. This account appeals to incomparability with the empty distribution. It rejects the mere addition principle without implying the weak repugnant (or weak deontic very repugnant) conclusion. And it doesn't require the equinumerosity constraint, dominance, or the conditional choiceworthiness relation.

I shall consider five objections to this average utilitarian strategy.

First, it might be objected that average utilitarianism still generates implausible reasons for or against adding people. Suppose, for example, you can have a child who will be better off than you, but doing so will make you worse off:

$$D = (4, \Omega)$$

$$E = (3, 6)$$

D is the subdistribution if you don't have a child, and E is the subdistribution if you do. Average wellbeing is greater in E than in D. So you ought to have children, even

though this makes things worse for you and better for no one. Similar reasoning shows that, according to average utilitarianism, you ought *not* to have children even in cases where doing so would be better for you and worse for no one.

This might seem a bad result because we take permissive views about the morality of procreation. I agree that it would be strange if we could be obligated to have children even though doing so would be worse for us, or obligated not to have children even though doing so would be better for us. But I think that such claims can be better explained in other ways than rejecting the average (and total!) utilitarian judgment that E is better than D. Many people independently think that we can justifiably pursue our own good (within limits) even when doing so fails to make things go impartially best (Scheffler 1994). And others think that the agent's own happiness is not, in itself, relevant to what she morally ought to do (Ross 1939), or that utilitarians should downplay the agent's happiness in some more sophisticated way (Sider 1993). We might think that agent-relativity shows up in other ways: I think, for example, that it may be permissible for me to have a child even if her existence (by some strange accident) prevents the existence of some stranger who would have a slightly better life. That permission is agent-relative because, in general, one ought to bring about a better rather than worse life. Therefore, although there may remain some counterintuitive implications about procreation even after adding independence of the unaffected, these intuitions might be more plausibly accommodated in other ways.[34]

Second, it might be objected that strong deontic neutrality doesn't fully vindicate

[34]These agent-relative suggestions do not endanger the impartiality of the axiology, because they only affect what we ought to do, not which distributions are better or worse.

our neutrality towards creating happy people. For example, it does not let us ignore changes in population size when deciding whether to save people's lives (Broome 2009). But this practice seems to me unjustifiable.[35] Average utilitarianism at least avoids the pitfalls of theories that try to justify this practice. For example, Broome supposes that global warming will kill a hundred million people and prevent a hundred million more from coming into existence. He imagines that the distributions are these:

$$F = (4, 4, \ldots, 4, 4, 4)$$

$$G = (4, 4, \ldots, 4, 2, \Omega).$$

Each location represents the welfare of a hundred million people. F is how things would've been without global warming. Broome's theory implies that we cannot assert (or deny) that F is better than G, because the value of the critical level is vague. Broome cannot say that global warming is bad. Average utilitarianism, by contrast, says that G is worse than F, and that we ought to prevent global warming in this scenario.

Third, independence of the unaffected might yield counterintuitive results when considering extinction. Suppose that, if humanity continues to reproduce, the distribution will be H, and otherwise it will be I:

$$H = (4, 4, \ldots, 4, 9, 9, \ldots, 9)$$

[35] Or, if it can be justified, it must be justified on very different grounds—e.g., related to our ignorance about how acts affect population size.

$$I = (4, 4, \ldots, 4, \Omega, \Omega, \ldots, \Omega)$$

Average utilitarianism says that H is better than I. But independence of the unaffected says to focus only on the welfare levels after 4—i.e., those of our descendants. And those subdistributions of H and I are incomparable. So we have no moral reason to continue to reproduce, even though the world will be better if we do. If some generation of humans could make themselves the last one, without affecting anyone else (including other species), they could permissibly do so.

I am not sure what to make of this conclusion. It does not strike me as obviously *wrong* for people not to reproduce, even if doing so would mean that humanity becomes extinct. But, to the extent that I find the conclusion counterintuitive, I think that the fault lies neither with average utilitarianism nor with independence of the unaffected. It may just be a deep mistake to suppose that present and past people's welfare could be unaffected by extinction. As Scheffler (2013) has argued, the value of our own lives and projects may depend on the future of humanity. It seems to me that the sting of the objection might be removed by Scheffler's insight. If Scheffler is right, then a human extinction scenario would look more like this:

$$I' = (2, 2, \ldots, 2, \Omega, \Omega, \ldots, \Omega)$$

Average utilitarianism says that this is worse than H. So we shouldn't let humanity go extinct. Maybe Scheffler is wrong, in which case I does reflect an extinction scenario relative to H. But then I would find it reasonable to deny that we have an obligation to continue our species.

Fourth, independence of the unaffected allows a sequence of permissible choices to yield a worse outcome. Recall the version of the mere addition paradox I gave on page 7. Suppose first that our options are these:

$$A = (100, \Omega, \Omega, \ldots, \Omega),$$

$$A+ = (100, 1, 1, \ldots, 1).$$

Independence of the unaffected tells us to ignore the first slot of people. Average utilitarianism then says that it's permissible to create the extra people, because the relevant subdistributions are incomparable. Suppose we do so. Suppose next that we can move from $A+$ to

$$Z = (2, 2, 2, \ldots, 2).$$

Independence of the unaffected is no longer relevant. We are now required to bring about Z, because it contains greater average wellbeing than $A+$. A permissible choice and an obligatory choice lead to Z, which is worse than where we started. If we were only choosing between A and Z, or between all three outcomes, we ought to have brought about A. And yet here we are.

This strikes me as undesirable, but not implausible. In fact, it captures my intuitions about the mere addition paradox pretty well. Intuitively, it is permissible, if we are in A, to bring about $A+$. And, intuitively, if we are in $A+$, we ought to bring about Z. But, intuitively, A is better than Z. It is unfortunate and confusing that a permissible

choice and an obligatory choice can together lead us to a worse outcome. But this conclusion does not strike me as impossible, repugnant, or less plausible than the conjunction of the premises that got us there.

Fifth, independence of the unaffected might be accused of being *ad hoc*; many arbitrary restrictions of average utilitarianism have been proposed (see McMahan 1981). But independence of the unaffected does not draw an *arbitrary* distinction—e.g., between persons living at different times. Even Sidgwick (1874 IV.1.1) restricts utilitarianism to "all whose happiness is affected by the conduct."

Moreover, total utilitarians may need to appeal to independence of the unaffected to get plausible results in infinite cases. Suppose that if you save some person's life, her lifetime wellbeing will be greater and there will be no other effects, but that there are infinitely many happy people. Your act does not increase the sum of wellbeing in the world. A partial solution to this "infinitarian paralysis problem" (Bostrom 2011) is to accept independence of the unaffected. Then, when your act does a finite amount of good, it still matters despite the infinite value of the universe.[36] So my claim isn't just that average utilitarians should accept this principle. I think everyone should accept it.

Independence of the unaffected may seem arbitrary to those who think we ought to make the history of the world go best. The history of the world includes unaffected people. But it seems to me that this view is essentially impersonal: it ascribes goodness to an entity—the world or its history—that isn't a person. I argue in Section 2.3 that average utilitarianism rests on person-affecting foundations, on which all bet-

[36]This doesn't address the infinite cases discussed in Section 2.3.2.

terness is betterness for someone. Average utilitarians care not about the goodness of the world or its history, but rather about the goodness of lives. Independence of the unaffected adds that, when deciding what to do, they should care about the goodness of the lives affected.

2.3 Balancing vs. Averaging

The main difference between Bader's view and average utilitarianism is the equinumerosity constraint. This constraint might be justified by Bader's method of balancing gains and losses, which only works in same-number cases. This method, which I stated on page 42, appeals to the differences between people's wellbeing, not to their sums. But it allows us to compare distributions with different members, because everyone's wellbeing can be measured on the same scale.

In this section, I discuss three reasons why Bader might reject average utilitarianism as inferior to balancing: first, average utilitarianism require sums of wellbeing to be meaningful quantities; second, average utilitarians cannot account for infinite cases; and third, average utilitarianism is arbitrary. I defend average utilitarianism against these objections.

2.3.1 Sums of Wellbeing

Average utilitarians evaluate a distribution by adding up different people's wellbeing and dividing the sum by the number of people. Bader might therefore object that average utilitarianism "constructs a special moral point of view by combin-

ing those of individuals into a single conglomerate viewpoint distinct from all of them" (Nagel 2012, 123). Bader's method of balancing eschews impersonal sums of different people's wellbeing.

We might wonder why differences between people's wellbeing are kosher while sums are not. We need differences to make interpersonal comparisons of wellbeing. But the difference between two quantities of wellbeing is a quantity of wellbeing, and it isn't someone's. Moreover, if our interpersonal scale of wellbeing has a meaningful zero and negative quantities (i.e., lives worth not living), then we can understand the sum of *x*'s wellbeing and *y*'s wellbeing as the difference between *x*'s wellbeing and the additive inverse of *y*'s wellbeing.

But Bader might say although sums of wellbeing are meaningful quantities, they are not axiologically significant. However, average utilitarians can determine which distributions contain better lives without summing different people's wellbeing. Consider an example:

$$A = (2, 8, \Omega)$$

$$B = (1, 8, 12)$$

Take the average of the differences between the first life in *A* and the lives in *B*:

$$\frac{(2-1)+(2-8)+(2-12)}{3} = -5$$

Then take the average of the differences between the second life in *A* and the lives

in B:

$$\frac{(8-1)+(8-8)+(8-12)}{3} = 1$$

Now take the mean of these average differences:

$$\frac{-5+1}{2} = -2$$

This means that, on average, the difference between the lives in A and the lives in B is -2. And indeed this equals the difference between A's average wellbeing (5) and B's (7).

This method, however, does not add up different people's wellbeing. It only, like Bader's method, adds up *differences* between people's wellbeing. Average utilitarianism, therefore, does not require meaningful or axiologically significant sums of different people's wellbeing. It is, of course, easier to compute the average difference between lives by taking the difference between the averages of them; that involves summing different people's wellbeing. But we can understand that as a shortcut for the moral arithmetic above. The two methods give the same results.[37] This is not mysterious. The sum of temperatures on the Celsius scale is not a meaningful quantity. But their average is, as is the average difference between two sets of temperatures.[38]

[37] At least, in all finite cases.

[38] Similarly for dates on a calendar.

2.3.2 Infinite Cases

Bader's balancing approach works well in infinite cases. It intuitively ranks A over B in the following case:

$$A = (2, 2, 2, 2, \dots)$$

$$B = (1, 1, 1, 1, \dots)$$

The ellipses after the last number means that the sequence is infinite. The series of differences between each A-life and each B-life diverges, so Bader says that A is better than B. Total utilitarianism, however, implies that A is not better than B, because each sum diverges, so neither sum is greater than the other.

Bader's view also gives good results when balancing gains and losses against distributions whose sums converge *conditionally*, such as C:

$$C = \left(1, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4}, \dots\right)$$

If we add C's terms in the order presented, the series converges to a finite number—$\ln 2$—but its corresponding series of absolute values

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4}, + \dots$$

diverges. Riemann's (1867) rearrangement theorem says that C's terms can there-

fore be rearranged to add up to any value whatsoever. It is, therefore, problematic for total utilitarianism to assign a value to C. Total utilitarianism would then violate impartiality: it would matter which people lead which lives, or when and where they live. So total utilitarians cannot say that A is better than C. Bader's method, by contrast, says that A is better than C, regardless of how C's terms are arranged. This seems right, because everyone in C is worse off than everyone in A, and all of the A-lives are worth living.

Bader might reject average utilitarianism because it cannot account for infinite cases. It asks us to divide an infinite sum of wellbeing by an infinite number of people. That is undefined. So average utilitarianism fares no better than total utilitarianism in infinite cases.

Average utilitarianism, however, can be extended in a natural way to capture these judgments. Instead of nonsensically dividing an infinite sum by infinity, we can understand the "average" of an infinitely large welfare distribution as follows. Suppose that some distribution X is an infinite sequence of welfare levels (x_k). Recall distribution A. Now take the averages of the first n terms of the sequence:

$$\frac{2}{2}, \frac{2+2}{2}, \frac{2+2+2}{3}, \frac{2+2+2+2}{4}, \ldots$$

This form a new sequence, whose terms are called the *Cesàro averages* of (x_k):

$$2, 2, 2, 2, \ldots$$

This sequence converges to 2. The limit of X's sequence of averages, as n goes to infinity, is what I call the average wellbeing of X. That is,

The Limit of Cesàro Averages: The average wellbeing of $X = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} x_k$.

A's sequence of Cesàro averages converges to 2. B's sequence of Cesàro averages converges to 1. So A's average wellbeing is greater than B's.

If we are against sums of different people's wellbeing but, like Bader, countenance differences, we can instead take Cesàro averages of differences. The first life in A is better than each life in B by

$$1, 1, 1, 1, \ldots$$

This is a sequence of welfare differences. The sequence of its Cesàro averages is

$$1, 1, 1, 1, \ldots$$

This sequence converges to 1. That is, in my sense, the average difference between the first person's wellbeing in A and the wellbeing of each person in B. If we do this for everyone else in A, we get the following sequence of average differences:

$$1, 1, 1, 1, \ldots$$

This sequence's Cesàro averages, as before, converge to 1. In this sense, the *A*-people are, on average, better off than *B*-people by 1 unit, and twice as well off as the *B*-people. That is the intuitively correct verdict.

What about *C*?

$$C = \left(1, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4}, \dots\right)$$

Its average wellbeing is 0: its terms tend to 0, and the limit of Cesàro averages of a convergent sequence is just the limit of the original sequence (Katznelson 2004, 13). Moreover, since the limit of a convergent *sequence* (not the *series*—i.e., the sum of its terms) does not depend on how its terms are arranged, the limit of *C*'s Cesàro averages will always be 0. This is holds for all conditionally convergent series: their terms converge to 0, so their Cesàro averages do too.

Average utilitarianism assigns a value to *C*: 0. It, therefore, secures the intuitive result that *A* is better than *C*. So far, I have explained how average utilitarianism fares better than total utilitarianism and no worse than Bader's method of balancing in the cases he considers. Does it have any advantage over Bader's method?

I think so. Bader's method cannot usefully tell us *how much* better one infinite distribution is than another:

$$A = (2, 2, 2, 2, \dots)$$

$$B = (1, 1, 1, 1, \dots)$$

$$D = (5, 5, 5, 5, \ldots)$$

The sum of the differences between the *A*-people and the *B*-people is infinite, as is that of the differences between the *D*-people and the *A*-people. This doesn't let us say that *D* is better than *A* by *more* than *A* is better than *B*. Average utilitarianism lets us say this, because the average levels in *A*, *B*, and *D* are 2, 1, and 5, respectively. This is an important advantage when it comes to practical questions—e.g., how blameworthy one would be for choosing wrongly, or whether a gamble with even odds of *B* and *D* is preferable to having *A* with certainty. Average utilitarianism has this advantage over Bader's view.[39]

I have discussed only the infinite distributions that Bader considers. Average utilitarianism may give less plausible results in other cases. And the evidential value of such cases may be limited, given our poor intuitive grasp of infinity. So I shall turn to a different reason why Bader might reject average utilitarianism.

2.3.3 Foundations of Average Utilitarianism

Average utilitarianism is often dismissed as arbitrary. Why should care whether, on average, better lives are lived? Temkin (1993) argues that the fundamental utilitarian intuition is that more of the good is better. Average utilitarianism seems arbitrary because it does not satisfy this intuition. And, without this intuition, why be utilitarian?

Average utilitarianism is indeed not supported by the impersonal intuition that

[39] Bader might, of course, introduce new machinery to answer these questions.

more of the good is better. It is instead supported by

The Impartial Person-Affecting View: *A* is better than *B* iff *A* is better for the *A*-people than *B* is for the *B*-people.[40]

This view appeals to what I called an *interdimensional* comparison on page 22.[41]

Impartial person-affecting theorists care about making people happy, in the sense that they prefer that better lives are lived. Total utilitarians care about this too, indirectly. But impartial person-affecting theorists do not care about making happy people: they do not care about the number of good lives. Nor do they care about making *particular* people happy: they do not care which lives are whose. That separates impartial person-affecting theorists from *partial* person-affecting theorists, who care more about present, actual, or independently existing people (Arrhenius forthcoming, ch. 9).

Average utilitarianism can be understood as an interpretation of the impartial person-affecting view: *A* is better for the *A*-people than *B* is for the *B*-people iff the *A*-lives are, on average, better than the *B*-lives (cf. Parfit 1984, 396). Bader accepts the impartial person-affecting view but interprets it differently. He thinks that the *A*-lives are better than the *B*-lives iff they are better via his method of balancing, which requires equinumerosity. He then extends this method to different-number cases through the dominance relation defined on page 43. Average utilitarianism

[40] This view is closer to what Temkin (2012) calls a *wide* person-affecting view than to Parfit's view of the same name, which is framed in terms of benefits.

[41] We might add, with Sen (1970b), that interpersonal comparability of wellbeing must be imprecise. But I have been pretending that our scale of wellbeing is precise.

is just a simpler interpretation of the impartial person-affecting view. It captures the neutrality of creating happy people without appealing to conditional choiceworthiness relations or implying weak repugnant conclusions.[42]

Rawls (1999, 25) argues that total utilitarianism fails to respect "the plurality of distinct persons with separate systems of ends." It does this by adding up different people's wellbeing. Rawls—rightly, I think—does not accuse average utilitarianism of this mistake. Total utilitarianism views each person's wellbeing as a distinct *location* where the same quantity—*good*—is instantiated. Person-affecting theorists, by contrast, view each person's wellbeing as a distinct *dimension* of value: my good and your good are different quantities, which cannot be "fused" together as if they were the good of a single person (Rawls 24).[43] The person-affecting theorist maintains that the only goodness is goodness for people, and that the bearer of value is not the world, but the life. This seems to me a deep disagreement between average and total utilitarianism.

Average utilitarianism can be impartial without being impersonal. This is reflected in Harsanyi's (1953) Rawlsian argument for average utilitarianism. The argument says, roughly, that if we had an equal probability of living any life in a distribution but would certainly exist, we ought to prefer a distribution with higher average wellbeing for our own sakes, because this distribution would, in expectation, be better for us. Therefore it is better. This reasoning is impartial, because we cannot give preference to our own interests if we don't know which interests are ours. And yet the reasoning is not impersonal, because it does not appeal to any good that

[42] It is also, Bader points out, the transitive closure of his dominance relation.

[43] However, although these dimensions are distinct, they can be measured on the same scale.

isn't someone's.

The textbook objection to this argument is that assuming our existence means that our preference is not impartial. Parfit (1984, 392) makes this objection and attributes it to Barry (1977, 317), who attributes it to Kavka (1975, 240). But no one has substantiated it. Parfit compares the assumption of existence to assuming that we are men when choosing between principles that would harm women. But assuming our sex violates impartiality because it allows us to know that we would not, as Parfit says, "bear the brunt" of our choice. Nonexistence, however, is no brunt to be borne: it couldn't possibly be worse for us. Assuming ignorance of our sex makes the choice impartial because it requires us to care about the interests of the opposite sex. If some chosen principle harms women and we turn out to be women, then we will be disadvantaged. But nonexistent people have no interests for us to care about. If some chosen principle reduces the number of people and we turn out not to exist, we will not be disadvantaged. So it is innocuous to assume that we will certainly exist. Moreover, between whom or what could we be failing to be impartial by making this assumption? Assuming our sex fails to be impartial between people of different sexes—i.e., between men and women. But we cannot fail to be impartial between existing and nonexistent people, because there are no nonexistent people.[44]

Parfit presents another objection, which I discuss in Section 2.4. There are many other objections to this argument, most of which are not specific to assuming our own existence. I doubt that the thought experiment provides a non-question-

[44] I here ignore the view, shared by Arrhenius and Rabinowicz (2015) and Fleurbaey and Voorhoeve (2015), that existence can be better for someone than nonexistence.

begging *argument* for average utilitarianism. But I think it reveals an important aspect of average utilitarian reasoning—namely, that it is impartial without being impersonal.

2.4 The Asymmetry

I have argued that average utilitarianism better explains the neutrality of creating happy people than Bader's view. Average utilitarianism is supported by an impartial person-affecting view and is no less plausible than Bader's method of balancing. If we accept independence of the unaffected, as Bader should do in order to avoid the deontic Egyptology objection, then we obtain strong deontic neutrality: mere addition is always permissible and never obligatory.

The most decisive objections to average utilitarianism involve lives that are worth not living. For example:

$$A = (-10, \Omega, \Omega, \ldots, \Omega)$$

$$B = (-10, -9, -9, \ldots, -9)$$

Average utilitarianism says that B is better than A. That is absurd.

Bader has an independently justified account of the wrongness of bringing miserable lives into existence, which I mentioned on page 44. His account is deontic rather than axiological: he does not say that it is worse for such lives to exist than

not to exist. He can use this account to explain why it is wrong to choose *B*. Because Bader's asymmetry is deontic, he cannot say that *B* is *worse* than *A*. But this would not bother him, because on his view, the betterness relation between distributions *just is* a betterness relation between lives, and the *B*-lives are indeed better than the *A*-lives. The hellish objections to average utilitarianism need not bother Bader. So Bader should still become an average utilitarian.

However, many of us do not have a deontic account of the asymmetry up our sleeves, and we find it crazy to say that *B* is better than *A*. Nonetheless, I think it would be premature to reject the relevance of average wellbeing entirely, given the independently plausible asymmetry between good lives and bad lives: we are neutral about making happy people, but we are against making miserable people. We should not, therefore, expect to treat good lives and bad lives symmetrically. In the previous section, I explained how a version of average utilitarianism captures the first part of this asymmetry: the neutrality of bringing happy people into existence. In this section, I focus on the second part: the wrongness or badness of bringing miserable people into existence.

On page 67, I mentioned Harsanyi's argument for average utilitarianism. I rejected the textbook objection that it violates impartiality to assume our own existence. Parfit's other objection to the argument is that it gives absurd results in hellish cases, like *A* and *B* above. The argument asks us not to care about the existence of additional people whose lives are worth not living, if their existence improves our prospects from behind the veil of ignorance. But I wonder if the argument can be revised to account for this problem. The standard version of the argument asks us what ought to be preferred for the sake of anonymous person who will certainly

exist. This question is important because such a preference is impartial without being impersonal. But we can reasonably add that it is *morally wrong* to prefer *B* to *A*, or to choose a principle or policy that leads to *B*, even for the sake of an anonymous existent. It is wrong because many more people's lives are worth not living in *B*, and this is not outweighed by the improvement in an anonymous person's prospects.

Deontic constraints can sometimes affect which preferences we ought to have even for our own sakes. For example, it might be expectably better for someone if her wealthy grandfather died prematurely, leaving her with a massive inheritance; but it is wrong for her or anyone else to prefer, even for her sake, that her grandfather die prematurely. I do not have an account of *why* it is wrong to prefer, even for one's own sake, that more people have lives worth not living. But it is clearly wrong. And if we think the better outcome is the one that ought to be preferred from an impartial perspective, then the morally permissible preferences that we ought to have for the sake of an anonymous person might be a good guide to betterness.[45]

One way of drawing this asymmetry might be

Asymmetric Utilitarianism: If *A* has negative average wellbeing and *B* does not, then *A* is worse than *B*. If neither *A* nor *B* has negative average wellbeing, then the better one is the one with greater average wellbeing. If both *A* and *B* have negative average wellbeing, then *A* is better than *B* iff *A* contains greater *total* wellbeing than *B*.[46]

[45] This would require abandoning the hope that we can derive derive all moral principles from impartial self-interest. I have no such hope.

[46] This statement would need to be revised in infinite cases.

This theory draws an asymmetry between good lives and bad lives. It is independently plausible that there is *some* such asymmetry, although I have no argument for why it should be this one. The theory ranks all nonnegative distributions above all negative ones, and then ranks nonnegative distributions by their averages and negative distributions by their sums.[47]

Other versions of asymmetric utilitarianism are possible. For example,

Negativist Asymmetric Utilitarianism: If the lives in A are, on balance, worth not living, and the lives in B are, on balance, at least neutral, then A is worse than B. If the lives in both A and B are, on balance, at least neutral, then A is better than B iff the lives in A are, on average, better than the lives in B. If the lives in both A and B are, on balance, worth not living, then A is better than B iff B contains a greater sum of uncompensated suffering than A.

Following Parfit (1984, 408), I say that a life's suffering is *uncompensated* if the life is worth not living. A distribution's sum of uncompensated suffering is the total badness of its bad lives. Negativist asymmetric utilitarianism is just like asymmetric utilitarianism, but its total utilitarian component for bad distributions cares only about lives that are worth not living. Another variant is

Extreme Asymmetric Utilitarianism: If A contains a greater sum of uncompensated suffering than B, then A is not at least as good as B. If, in addition, A contains less total or average wellbeing than B, then A is worse than B. If,

[47] Parfit (1984, sec. 138) discusses views that treat happiness and suffering asymmetrically, but his objections only apply to views on which happiness has diminishing marginal value.

> however, A contains greater total and average wellbeing than B, then A is neither worse than nor at least as good as B: they are on a par. If A doesn't contain a greater sum of uncompensated suffering than B, then A is better than B iff the lives in A are, on average, better than the lives in B.

Extreme asymmetric utilitarianism says that a greater sum of uncompensated suffering (regardless of the overall wellbeing of the population) can never be outweighed by increases in total or average wellbeing.

All three views give the right result in the case on page 69: B is worse than A. But I shall stick with the simpler view above—plain-vanilla asymmetric utilitarianism. Let me mention three problems with the view.

First, asymmetric utilitarianism implies the sadistic conclusion, which I stated on page 15. It says that C, below, is better than D:

$$C = (10, -5, \Omega, \Omega, \ldots, \Omega)$$

$$D = (10, 1, 1, 1, \ldots, 1)$$

This view says that we only consider the averages because both distributions are good overall. So, if there are enough people in D, C is better.

Independence of the unaffected might soften this blow. It allows us, when choosing between C and D, to ignore the first person and focus only on the subdistributions below:

$$C' = (-5, \Omega, \Omega, \ldots, \Omega)$$

$$D' = (1, 1, 1, \ldots, 1)$$

D' is better than C' according to asymmetric (and standard average) utilitarianism. So we ought not to bring the miserable life into existence.

However, this strategy can be easily nullified:

$$C = (10, -5, \Omega, \Omega, \ldots, \Omega)$$

$$D'' = (11, 1, 1, 1, \ldots, 1)$$

Asymmetric utilitarianism implies that C is better than D''. We cannot appeal to independence of the unaffected in this case. Negativist asymmetric utilitarianism doesn't help either. Extreme asymmetric utilitarianism implies that C is worse than D'', because C has more uncompensated suffering and lower total wellbeing. But extreme asymmetric utilitarianism also implies that E and F, below, are on a par:

$$E = (-1, 9, 9, 9, \ldots, 9)$$

$$F = (1, 1, 1, 1, \ldots, 1)$$

E has more uncompensated suffering because someone's life is worth not living, but has greater total and average wellbeing than F, so extreme asymmetric utilitari-

anism says that E and F are on a par.[48] Asymmetric utilitarianism, therefore, seems either too averse or insufficiently averse to uncompensated suffering.

I find it more plausible that C is better than D'' than that E isn't better than F. Why might we think that C cannot be better than D''? We might think that miserable life in C makes it worse than any distribution containing only lives worth living.[49] But there must be limits to the wrongness or badness of bringing miserable lives into existence. Proponents of the procreative asymmetry owe an account of when we can permissibly create miserable lives. For example, we need to explain why our species can continue to reproduce, even though we know that we will certainly bring more miserable lives into existence by doing so. Asymmetric utilitarianism gives a crude account: we can add such lives whenever they bring about higher average wellbeing (if the affected lives are good overall) or total wellbeing (if the affected lives are bad overall) than the alternatives. Until we have a better account, I think it would be premature to regard this objection as decisive.

Second, asymmetric utilitarianism makes comparability intransitive. Suppose that your options are these:

$$G = (-1)$$

$$H = (1)$$

[48] It therefore violates Arrhenius's (forthcoming) *general non-extreme priority condition*.

[49] Alternatively, we might focus on the fact that everyone who exists in both distributions is better off in D''. But this is *partial* person-affecting reasoning: it does not apply if we change which lives are whose. Such reasoning may be attractive, but it faces major problems that I wish to avoid here (Temkin 2012, ch. 12). Or we might think that D'' is better than C because it contains much more wellbeing. This total utilitarian reasoning, however, seems to lead straight to the repugnant conclusion.

$$I = (\Omega)$$

Asymmetric utilitarianism says that G is worse than H and I. And it says that H and I are incomparable, because the average wellbeing in I is undefined. So you shouldn't choose G, but it's fine to choose I. This makes comparability intransitive: H is comparable with G, which is comparable with I, but H is incomparable with I.

How bad is this result? I assumed on page 10 that comparability was transitive only because it simplifies the distinction between incomparability and parity. But if we can independently distinguish between these relations—as Rabinowicz (2012), Chang (2002), and Carlson (2010) argue—then I would be happy to allow intransitive comparability. Otherwise, this seems to me a major problem with asymmetric utilitarianism.

Third, asymmetric utilitarianism implies negative analogues of the repugnant conclusion:[50]

$$-A = (-100, \Omega, \Omega, \ldots, \Omega)$$

$$-Z = (-1, -1, -1, \ldots, -1)$$

Imagine that each location represents the welfare of ten billion people. Asymmetric utilitarianism implies that $-A$ is better than $-Z$ because $-A$ contains more total wellbeing (assuming there are enough people in $-Z$). The negativist version has the

[50] See Carlson (1998).

same implication, as does the extreme version because $-Z$ contains a greater sum of uncompensated suffering and less total wellbeing. If the $-A$-lives are sufficiently horrible, this strikes me as repugnant, although much less so than the original repugnant conclusion.

There are likely many other problems with the theory. It may give strange results when average wellbeing hovers close to zero. Because it is a hybrid theory, it lacks the intrinsic plausibility of its components and may be subject to the objections to both. There may be more disastrous problems that I have not considered.

2.5 Conclusion

Asymmetric utilitarianism is a simple theory. It says that, conditional on lives not being miserable, it is better if better lives are lived; otherwise, it is better if there is more wellbeing. When combined with independence of the unaffected, it secures both parts of the procreative asymmetry and avoids the deontic Egyptology objection. It avoids the repugnant conclusion and gives intuitive results in infinite cases where total utilitarianism fails.

I do not endorse asymmetric utilitarianism. I think it would be better to have an independently justified account, such as Bader's, of the wrongness of creating miserable lives. If this account is deontic, rather than axiological, and is silent on the creation of happy lives, then we could combine it with average utilitarianism and independence of the unaffected to obtain the asymmetry. The theory would then have a simpler axiology. That is what I think Bader should do. But if such accounts of the asymmetry do not succeed, then I think asymmetric utilitarianism, despite

its problems, deserves serious consideration.

3 Total Utilitarianism without Repugnance

Recall the mere addition paradox:

$$A = (100, \Omega, \Omega, \ldots, \Omega),$$

$$A+ = (100, 1, 1, \ldots, 1),$$

$$Z = (2, 2, 2, \ldots, 2)$$

In the previous chapters, I considered views on which $A+$ is not at least as good as A. Critical-band utilitarianism says that $A+$ and Z are on a par with A. I agree with Broome that this appeal to parity is objectionably *ad hoc*. Bader's view says that $A+$ and Z are incomparable with A. This avoids Broome's objections to critical-band utilitarianism. But the view is problematic in other ways. It is designed to capture the intuitive neutrality of creating happy people. But this neutrality is more plausibly explained by average utilitarianism, which has its own problems. My conclusion from these two chapters is that incommensurability in population ethics is not plausibly explained by mere differences in population size, as Bader and proponents of critical-band utilitarianism propose.

In this chapter, I see what happens if we grant the mere addition principle: $A+$ is at least as good as A. In Section 3.1, I consider two views that appeal to lexical superiority in population ethics. In Section 3.2, I suggest that the objections to lexical superiority can be mitigated by appealing to parity.

3.1 Two Lexical Views

In order to avoid the conclusion that *Z* is better than *A*, these views must reject non-anti-egalitarianism, which I stated on page 8: they must say that *Z* is not better than *A*, even though *Z* contains greater total and average wellbeing and has it more equally distributed.

Parfit (2004) proposes such a view. According to

Parfit's Perfectionism: "[E]ven if some change brings a great net benefit to those who are affected, it is a change for the worse if it involves the loss of one of the best things in life." (19)

This is one version of

The Lexical View: There is some number of excellent lives whose existence would be better than any number of mediocre lives, even though the existence of each mediocre life would make things better, by some fixed amount that does not depend on the number or wellbeing of other people.

This last clause makes the lexical view relevantly different from *variable value views*, on which the contributive value of a life depends on how many other people exist and on how well off they are (Hurka 1983).

In this chapter, I defend a different version of the lexical view. It is simply a version of total utilitarianism:

Lexical Total Utilitarianism: *A* is better than *B* iff *A* contains a greater sum of wellbeing. But a distribution in which enough people's lives are excellent contains a greater sum of wellbeing than any distribution in which everyone's lives are, at best, just barely worth living.

This view passes the buck of avoiding the repugnant conclusion to our theory of wellbeing. It, therefore, differs from Parfit's perfectionism. Parfit doesn't think that there is *less wellbeing* in a world with fewer of the best things in life. Instead, he thinks that such a world is *worse*, even though it may contain much more wellbeing.

These views differ in their responses to the mere addition paradox. I said that if we grant the mere addition principle, then we can only avoid the conclusion *that Z is better than A* by rejecting non-anti-egalitarianism. Just so, for Parfit's perfectionism. Lexical total utilitarianism, however, entails non-anti-egalitarianism: any population with greater total wellbeing is better; equality has nothing to do with it. So it cannot avoid the conclusion that *Z* is better than *A*. Doesn't it, therefore, entail the repugnant conclusion?

Not so fast. To obtain the repugnant conclusion, we have to assume that any excellent life is better than a mediocre life by some scalar quantity of wellbeing. For example, we've been assuming since page 8 that a life at 100 is excellent and that a life at 1 is barely worth living. This means that an excellent life is 100 times better than a mediocre life. That is what our numbers mean. And that is why total utilitarianism seems to lead to the repugnant conclusion. If any excellent life is at most *n* times better than any mediocre life, where *n* is a real number, then we obtain the repugnant conclusion by adding at least *n* mediocre lives for every excellent life.

But it is reasonable to deny that there is any such n. We could, I suppose, stipulate that "100" is excellent and that "1" is mediocre. But then our scale would not reflect how good a life really is. We could not say, for example, that Z contains a greater sum of wellbeing than $A+$, even though the sum of the numbers assigned is greater.

Lexical total utilitarians reject that quantities of wellbeing are scalar quantities (as defined on page 11). This is because the real numbers obey

The Archimedean Property: For any scalar values x and y with $x > 0$, there is some natural number n such that $nx > y$.

There might be real-valued ratios between *some* quantities of wellbeing—e.g., between any two excellent lives or any two mediocre lives. Our version of the mere addition paradox above would then be operating within what Griffin (1988) calls a "pocket of cardinality" (97). But the repugnant conclusion tries to connect different pockets on a single cardinal scale, which lexical total utilitarians believe cannot be done.

That is why lexical total utilitarians require incommensurability, as I said on page 12. They deny that the dimensions of wellbeing can be measured on a single cardinal scale. For then the Archimedean property would kick in, leading to the repugnant conclusion. This is the sense in which Ross (2002, 150) thinks virtue to be incommensurable, but comparable, with pleasure.[51] This is a manifestation of incommensurability that need not violate completeness. But I extend it to violate completeness in Section 3.2.

[51] He later (1939) thinks they are incomparable, because they are good in different senses.

Many theorists would independently reject the assumption that quantities of well-being can be represented by single real numbers. Sen (1980), for example, argues that wellbeing is fundamentally a vector quantity—i.e., representable as a list of components. He thinks that even hedonists should accept this. Even if the vector's components can be added or otherwise reduced to a scalar value, that should be regarded as a special case and a substantive claim about wellbeing. Chipman (1960) takes a similar view and goes so far as to *define* utility as a lexical ordering represented by a vector. Like Sen, Chipman thinks that everyone should accept this definition, even if they think that the components of the vector (or what he calls the "lexical number") can be summed together to form a scalar quantity.[52] Sen's and Chipman's reasons have nothing to do with the repugnant conclusion. So lexical total utilitarianism's core move is not merely an *ad hoc* response to the problems of population ethics; this cannot be said of many other views.

Here is a toy model of lexical total utilitarianism. We represent each person's well-being as a list of components, whose first component carries the more important aspects of wellbeing (i), and whose second component carries the more trivial aspects (t).[53] What are the important things? What are the trivial things? I don't know. I mention some broad possibilities through this chapter, but I don't assume any of them.

Suppose that the values of the important component can be represented by integers, and that the values of the trivial components can be represented by real numbers, with no upper and lower bound on either dimension. The sum of (i_1, t_1) and (i_2, t_2)

[52] Chipman would say that Broome's (1991) definition is a special case.

[53] Here I follow Kitcher (2000).

is $(i_1 + i_2, t_1 + t_2)$. We now want a lexical ordering of these values. The standard lexical ordering is

$(i_A, t_A) \geq (i_B, t_B)$ iff either

(a) $i_A > i_B$, or

(b) $i_A = i_B$ and $t_A \geq t_B$.

This means that any increase along the important dimension makes things better overall, no matter how much worse things are along the trivial dimension. In Section 3.2, I provide a more plausible partial ordering. But the standard ordering suffices for now.

Say that a life is neutral if it has none of the important stuff ($i = 0$) and none of the trivial stuff ($t = 0$), that a life is barely worth living iff it has none of the important stuff ($i = 0$) and some of the trivial stuff ($t > 0$), and that a life is excellent only if it has some of the important stuff ($i > 0$). In this model, the existence of any number of excellent lives would be better than any number of lives that would be barely worth living.

We should not read too much into this toy model. The whole idea behind lexical total utilitarianism is that wellbeing is more complex than arguments for the repugnant conclusion assume. Lexical total utilitarians should not make the same mistake by thinking that wellbeing can accurately be represented by *two* numbers. We should, therefore, remember that the model is an oversimplification.

I am not the first to suggest that total utilitarianism can avoid the repugnant con-

clusion. This suggestion, or something like it, is mentioned by many others.[54] But many theorists seem either not to notice it or not to take it seriously: Cowen (1996), Sider (1991), Arrhenius (2000), Parfit (2004), Huemer (2008), and Temkin (2012) all assert that total utilitarianism entails the repugnant conclusion.[55]

This chapter makes two contributions to the literature on lexical total utilitarianism. First, I explain why lexical total utilitarianism is more plausible than Parfit's perfectionism. In Section 3.1.1, I argue that lexical total utilitarianism provides a better account the single-life repugnant conclusion, and in Section 3.1.2, I argue that it better explains the repugnance of the (many-life) repugnant conclusion. In Section 3.2, I suggest that lexical total utilitarians can avoid some of the objections to their view by appealing to parity, in a way that avoids Broome's objections discussed in Chapter 1. I think this view may identify the most plausible source of incommensurability in population ethics.

3.1.1 Single-Life Repugnance

I begin with McTaggart's version of

The Single-Life Repugnant Conclusion:

> Take a life which in respect of knowledge, virtue, love, pleasure, and intensity of consciousness, was unmixedly good, and possessed any

[54] See, for example, Griffin (1988), Crisp (1992), Portmore (1999), Kitcher (2000), Thomas (manuscript), and Carlson (manuscript).

[55] Arrhenius (forthcoming) doesn't consider this theory a version of total utilitarianism because it implies that wellbeing cannot be represented on a single ratio scale. But this is an unduly restrictive definition: it implies that Mill, as many interpret him, couldn't have been a total utilitarian. Also, see Carlson (2007) for a simple model of additive measurement that violates the Archimedean property.

> finite degree of goodness you choose. Suppose this life prolonged for a million years if you like without its value in any way diminishing. Take a second life which had very little consciousness, and had a very little excess of pleasure over pain, and which was incapable of virtue or love. The value in each hour of its existence, though very small, would be good and not bad. And there would be some finite period of time in which its value would be greater than that of the first life, and another period in which it would be a million times greater. (1927, 452–53)

McTaggart calls the longer life an "oyster-like" life. He accepts the single-life repugnant conclusion, although he notes that it would "be repugnant to certain moralists."

It is, I think, a desideratum of a solution to the repugnant conclusion that it can be extended in a natural way to solve the single-life repugnant conclusion (or vice versa). There are, of course, important differences between the two cases. For example, there is someone for whom a longer life is better, but there might be no one for whom a larger population is better. I don't want to deny that this difference is important, or to claim that these two repugnant conclusions should be avoided in the exact same way. But there are important connections between them, too.

Here is one such connection. If the single-life repugnant conclusion is true, that would make the many-life version much harder to reject. Suppose that the single-life repugnant conclusion is true: for any life of finite length at some very high welfare level, there is some oyster-like life that, because it would be much longer, would be better. Then consider some very high welfare level a, for which it seems

that any sufficiently large population at level *a* would be better than any number of people living at some much lower level *z*. The single-life repugnant conclusion implies that a sufficiently long oyster-like life could be at some level as high as, or higher than, *a*. Now consider

The Mundane Conclusion: For any number of oyster-like lives, however long, there is some much larger number of oyster-like lives whose existence would be better, even though they are, because they are each much shorter, barely worth living.

This claim does not strike me as repugnant. But the mundane conclusion and the single-life repugnant conclusion (along with some other plausible assumptions) entail the repugnant conclusion: for any number of lives at any welfare level, there is some number of lives at some much lower (but still positive) welfare level whose existence would be better. Take some finite population of lives at any welfare level, however high. This population would be worse than some population of very long oyster-like lives. This follows from the single-life repugnant conclusion and

The Weak Impartial Pareto Principle: For any populations *A* and *B* of the same size, if everyone in *A* is better off than everyone in *B*, then *A* is better than *B*.

Because each oyster-like life would be better than each life of very high quality, the weak impartial Pareto principle implies that the population of oyster-like lives is better. And this population would, in turn, be worse than some much larger number of shorter oyster-like lives, each of which would be barely worth living. So,

by transitivity, for any finite population of lives at any welfare level, however high, there is some larger population whose existence would be better, even though each of its members lives a very short oyster-like life.

In order to reject the standard repugnant conclusion, we should reject either the single-life repugnant conclusion, the mundane conclusion, or the weak impartial Pareto principle. The single-life repugnant conclusion is, I think, the most implausible of these claims.

Here is one way of highlighting the implausibility of rejecting the mundane conclusion while accepting the single-life repugnant conclusion: we increase the size of the oyster-like population in cheesy ways that also reduce the lifespan of its members. The strategy here depends on one's views about personal identity, and not all views will permit such a strategy. I shall illustrate it with a variation on a case due to Parfit (1984, 299):

Amoeba *Z*: In *A*, the oyster-like beings each live for *n* years (where *n* is enough to make their lives very well worth living, by the single-life repugnant conclusion). In *Z*, the oyster-like beings reproduce by dividing, like amoebae. They divide at regular intervals much shorter than *n*—so short that each life is just barely worth living. Because they divide so often, there are vastly more beings in *Z* than there are in *A*.

It seems to me that amoeba *Z* could easily be better than *A*. Although the division process makes each life, from one fission to another, much shorter and less worth living, that doesn't seem to make *Z* much worse. We could even imagine that each

post-fission descendant is psychologically and physically extremely similar to its ancestor. At any single time, the only difference between *A* and amoeba *Z* would be that the latter contains many more minds experiencing the same simple pleasure. This makes the mundane conclusion seem very plausible to me. So I think that the single-life and many-life repugnant conclusions stand or fall together. I don't have a non-question-begging argument for the mundane conclusion that would convince theorists who are already convinced of some conflicting population axiology. But it is clearly not *repugnant*, and that datum needs to be explained.

Lexical total utilitarianism explains it as follows. The mundane conclusion is true, because this version of *Z* contains a greater sum of trivial goods than *A*. The single-life repugnant conclusion is false, because the oyster-like life fails to instantiate the more important dimensions of wellbeing. That is why it is barely worth living. Extending such a life by millions of years cannot make it excellent. In our toy model, extending this life just multiplies 0 by a large number. And the repugnant conclusion is false, because no number of oyster-like lives could contain a greater sum of wellbeing than a population of excellent lives. No real number reflects how much better an excellent life is than an oyster-like life.

I don't mean to imply that the aggregation of goods within each life is exactly the same as the aggregation of goods across lives. There may, for example, be interaction effects between goods. McTaggart asks us to suppose, for instance, that the first life in his comparison is good with respect to "knowledge, virtue, love, pleasure, and intensity of consciousness." But what if we instead split these goods up, so that each of five people enjoyed one of these things and had none of the others? It would not be clear to me that each of these people's lives would be better than

any oyster-like life, however long. Nor would it be clear to me that those lives taken together would be better than any number of oyster-like lives. But this suggests to me not that the lexical diagnosis is wrong, but rather that we need to be careful about what makes some lives lexically better than others. What makes lives most worth living may involve organic unities or patterns within a life. Parfit (1984, 502), for example, suggests that what makes someone's life go best is wanting and taking pleasure in certain (rational, loving, and aesthetic) activities. The right kind of combination of desire, pleasure, and activity is, on this view, much more valuable than each of these things taken separately.

Griffin (1988, 34f.) argues, more generally, that wellbeing is not an additively separable function of various goods. He suggests a *global view* of wellbeing (Crisp 1992), on which we cannot determine the value of a life simply by looking at goods in isolation. This view is compatible with lexical total utilitarianism. Lexical total utilitarianism requires that the value of a distribution be an additively separable function of the wellbeing of its members. But it does not require the wellbeing of its members to be an additively separable function of various goods (e.g., pleasures and accomplishments). Lexical total utilitarianism may seem incompatible with the global view because I have been representing the goodness of lives and distributions as lists with multiple components. But I haven't said what the components are or how lives can be better or worse in terms of them. The important dimensions could be global properties of lives. That seems, for example, to be Broad's (1938) solution to the single-life repugnant conclusion: the shorter life's "elaborate temporal pattern" makes it like "the performance of an opera," whereas the oyster-like life "is like a single note played . . . on a single very simple instrument such as a

tin-whistle" (688).

What is the perfectionist diagnosis of the single-life repugnant conclusion? For simplicity, let's assume that knowledge, virtue, and love—the things in the shorter life imagined by McTaggart—are perfectionist goods (i.e., some of the best things in life). Whether or not Parfit can appeal to such perfectionist goods in order to avoid the single-life repugnant conclusion depends on whether Parfit sees perfectionist goods as ingredients of welfare or as impersonal ideals (or both). By *ingredients of welfare*, I mean the kinds of things that make someone's life go well, and are good for her. By *impersonal ideals*, I mean the kinds of things that are supposed to make worlds good, without necessarily making them good for anyone. If the perfectionist goods are merely impersonal ideals, then Parfit's perfectionism cannot explain why we should reject the single-life repugnant conclusion. That's because the single-life repugnant conclusion is a judgment about the goodness of lives. The most that Parfit could say is that a world in which the shorter life exists would be better than one in which any oyster-like life exists. But it seems to me not just that it's better if the shorter life *exists*, but also that the shorter life is a better *life*: it has a higher lifetime wellbeing, and if one could live either the shorter life or the oyster-like life, one should prefer to live the shorter one for one's own sake.

Suppose, then, that Parfit sees his perfectionist goods as ingredients of welfare, and as being lexically superior to the mild pleasures of the oyster-like life. He can then avoid the single-life repugnant conclusion by appealing to his perfectionist theory. But then his view would be compatible with lexical total utilitarianism. Perfectionism would simply identify what makes some quantities of wellbeing lexically greater than others. Combining a perfectionist theory of wellbeing with the struc-

ture of lexical total utilitarianism would allow Parfit to avoid the single-life and many-life repugnant conclusions, without violating non-anti-egalitarianism.[56]

Parfit may have several reasons for rejecting the lexical total utilitarian construal of perfectionism. One reason might be to make his lexical view compatible with many different theories of wellbeing. For example, although some hedonists follow Mill in holding that pleasures of certain qualities are lexically superior to others, many would reject this view. And Parfit seems to think, or at least to hope, that the solutions to the problems of population ethics do not depend on our views about wellbeing. But it seems to me that the plausibility of different methods of aggregating "the amount of whatever makes life worth living" (Parfit 1984, 387) may depend on what one thinks makes life worth living. Some views of wellbeing—e.g., views that value only the nonfrustration of preferences—have clear implications for population ethics.[57] We shouldn't reject an otherwise attractive population axiology just because it is less plausible according to certain theories of wellbeing. Indeed, it is progress to discover that different theories stand or fall together.

Parfit might reject my diagnosis of the single-life repugnant conclusion because he distinguishes between *quality of life* and *quantity of wellbeing* (manuscript). He gives an example in which the best things in *x*'s life are of a higher quality than the best things in *y*'s, but where *x*'s life is worse because it is much shorter. This distinction might be what Parfit has in mind in his diagnosis of the single-life repugnant conclusion: "The Century of Ecstasy would be better for me in an essentially qualitative way. Though each day of the Drab Eternity would have some value for me,

[56] This would also make an additional non-welfarist construal of perfectionism redundant.

[57] See, e.g., Fehige's contribution to Fehige and Wessels (1998).

no amount of this value could be as good for me as the Century of Ecstasy" (2004, 18).[58]

However, I do not understand Parfit's distinction between quality of life and quantity of wellbeing. In Parfit's example, the *things* in *x*'s life are of a higher quality than the *things* in *y*'s. But *x*'s *life* is worse, so she has lower *lifetime* wellbeing. It seems to me that the quality of someone's life is just how good or bad it is: if *x*'s life is worse than *y*'s, then it is of lower quality. If *x*'s life is also shorter, then it might be true that, during each period of *x*'s life, *x* enjoy a greater quality of life. But this is a judgment about momentary wellbeing, not about lifetime wellbeing. I cannot intelligibly distinguish between the quality of an *entire* life and the quantity of *lifetime* wellbeing. Similarly, it seems to me that if no amount of a drab eternity's value could be as *good* as a century of ecstasy, then no amount of this value could be *as much value* for someone as a century of ecstasy. I do not understand the view that a century of ecstasy would be better, but would have less value, than a drab eternity. Quantities of wellbeing *just are*, as I understand them, qualities of life.

There are, of course, other responses to the single-life repugnant conclusion. We might, for example, think that the shorter life is better because pleasure has diminishing marginal value for the person experiencing it. I do not find this very

[58] Parfit also says, in endnote 3, "These remarks assume that the quality of life is higher if people's lives go better, and that each life goes better if it contains a greater quantity either of happiness or of whatever else makes life worth living. 'Quality' thus means 'quantity, per life lived'. In Section 5 below," where he discusses the single-life repugnant conclusion, "I drop this assumption, thereby simplifying the contrast between quality and quantity. (If this note is puzzling, ignore it.)" Similarly, McTaggart thought that those who would reject the single-life repugnant conclusion would be motivated by "a conviction that quality is some thing which is inherently and immeasurably more important than quantity" (453). Kagan (1994) distinguishes between a person's wellbeing and her quality of life. I agree that persons are importantly distinct from their lives. But even Kagan (2009, 271) now rejects his earlier conclusion.

plausible, but I shall not argue against it here. My claim in this section is merely that it is reasonable to take a lexical view about quantities of wellbeing in light of the single-life repugnant conclusion. This gives lexical total utilitarianism an independent motivation that perfectionism lacks. If perfectionist goods are impersonal ideals, then perfectionism seems not to explain the single-life repugnant conclusion. If they are ingredients of welfare, then Parfit should accept lexical total utilitarianism.[59]

3.1.2 Repugnance

What is repugnant about the repugnant conclusion? According to Parfit, the repugnant conclusion is hard to accept because, in *Z*, "people's lives are barely worth living, and most of the good things in life are lost" (208). But, as Ryberg (1996) argues, this diagnosis seems not to fully explain the repugnance of the repugnant conclusion.

Ryberg's argument appeals to a version of *Z* in which the best things in life are not lost. He considers a world in which people's lives are mostly mediocre, but they contain both glimmers of excellence and some sprinkles of great evil. Ryberg suggests, for example, that they might contain "an experience of Mozart's music but also the experience of undergoing a serious operation without anaesthetic" (208). Ryberg says that the worst things in life "have the same numerical value as higher positive values (surely it is reasonable to think that there are such higher negative

[59] Reflection on the nature of welfare may lead us to lexical total utilitarianism in other ways. List (2004) presents an analogue of Arrow's theorem for multidimensional theories of wellbeing. We might become lexical total utilitarians because we accept interpersonal but not interdimensional comparisons of wellbeing.

values)" (ibid.). He imagines that these great evils cancel out the higher goods, making their lives just barely worth living. So it is a version of *Z*. And presumably it would be repugnant to conclude that this version of *Z* is better than *A*. But Parfit's diagnosis of the repugnant conclusion seems not to apply to this case. In this version of *Z*, most of the good things in life are not lost. They are just thinly spread and counterbalanced by evils.

I have doubts about Ryberg's appeals to one-off tastes of higher goods. If the best things in life were things like *listening* to Mozart, then it would be clear that these things could, in principle, be enjoyed in tiny doses. But those doses would not do much to make life worth living. It is more plausible to value the *appreciation* of Mozart. Appreciation—at least, the kind that it is plausible to assign great value—requires more than mere listening, and even more than enjoyment, but also recognition of why the thing is worth enjoying. Such recognition may require a background of knowledge and other experiences—e.g., having listened to Mozart's predecessors and successors. And if one has the kind of knowledge and experience that makes one able to appreciate something in a valuable way, then one has already achieved some (perhaps lesser) perfections.

However, even if we reject Ryberg's particular claims about his version of *Z*, we could imagine a sufficiently similar version of *Z*. Some Holocaust survivors, for instance, have claimed that their lives were, on the whole, barely worth living (Hayes 2015, 779). Imagine a person who, after years in a concentration camp, is liberated and has a good future: she starts a new family, experiences some of the best things in life, and is, when not remembering her past, happy. It is reasonable to think that at least some lives of this kind are barely worth living; some people who have led

such lives have thought so. If they are right, then Ryberg's version of *Z* might be populated by many such lives. It is repugnant to conclude that this world is better than a much smaller one in which there are no concentration camps. But this repugnance cannot be explained by the absence of the best things in life, because the best things in life are present in this *Z*, and perhaps in greater quantities than in *A*.

I don't think this objection to Parfit's perfectionism is decisive: Parfit could perhaps give a different diagnosis of Ryberg's case. But a unified explanation would be more satisfying.

Lexical total utilitarianism offers a different diagnosis. On this theory, the repugnant conclusion is false because, no matter how big *Z* is, there is less wellbeing in *Z* than in *A*. The repugnance lies in the conception of wellbeing required for *Z* to contain more wellbeing than *A*. The value of a life cannot be reduced to a single real number. When explaining why total utilitarianism entails the repugnant conclusion, Parfit (1984) compares *Z* to "a heap of bottles each containing only a single drop" of milk (338). But wellbeing is, in this way, unlike milk.

This diagnosis also applies to Ryberg's version of *Z*. Ryberg characterizes the lives in his *Z* as being, overall, neutral in the important ways—because they don't contain a surplus of either higher positive or negative values—and good in the trivial ways. That is why the lives are just barely worth living. Lexical total utilitarianism implies that no population of such lives could contain more wellbeing than a population of lives that go well in the important ways.

We might, however, suspect that other versions of the repugnant conclusion will slip through the cracks of lexical total utilitarianism. This would be true if the

most important dimensions of wellbeing, or the higher goods, can come in very small degrees. For we could then imagine a vast population of people who barely instantiate the important goods. Suppose, with Ross, that virtue is good for its own sake and lexically superior to pleasure.[60] Now consider a world in which billions of people are highly virtuous. Is there some number of people whose existence would be better, even though each person is only slightly virtuous? The affirmative answer seems no less repugnant than the initial repugnant conclusion. But, on one way of aggregating virtue, there is more virtuous activity in the more populous world.

One way of answering this objection would be to posit lexical superiorities within dimensions or goods. We might think, for example, that no number of barely virtuous people would instantiate or manifest as much virtue as a sufficient number of truly noble people. Many have made similar claims about other goods or evils. Gurney (1887), for example, claims that no duration of moderate pain would be worse than torture. Some might hold analogous views about the goodness of knowledge. On one view of this kind, knowledge is good for its own sake, but some kinds of knowledge are lexically superior to others. For example, some fundamental physical laws or moral truths may be more worth knowing than any number of facts about grains of sand on beaches.

The objection could also be avoided by views on which the more important dimensions of wellbeing are discrete. The objection is strongest for values that are structurally like the real numbers. For then the Archimedean property kicks in to generate repugnant conclusions. If the values along the important dimension are

[60] I here ignore that Ross's notion of goodness is not a notion of wellbeing; it is a property of facts, not of lives.

discrete, like the integers, then it is easier to avoid such conclusions. The simplest case, discussed by Kitcher (2000), would be a binary dimension. For example, the lives that are most worth living might be ones that are *meaningful* or *free*.[61] Although people can be more or less free and have more or less meaningful lives, we might care most about the binary question—i.e., whether we are free, or whether our lives have meaning. (This may be like the way in which some people care much more about not becoming bald than about having less hair.) We might think that a world filled with enough free agents or meaningful lives, if they are sufficiently happy, contains more of what makes live worth living than a world filled with any number of unfree agents living meaningless lives, however happy they are.

As these remarks suggest, I don't think that lexical total utilitarianism's avoidance of seemingly repugnant conclusions is a *fait accompli*. Much depends on our theory of what makes life worth living. If, for example, we are hedonists who reject the lexical priority of any pleasures or pains, then we cannot avoid the repugnant conclusion by appealing to lexical total utilitarianism. But I am not confident that population axiology and the theory of wellbeing can be successfully pursued independently of each other.

3.2 The Lexical Threshold

I have argued that lexical total utilitarianism better explains the repugnance of both the single- and many-life repugnant conclusions. It also entails the eminently plausible non-anti-egalitarian principle. It has other advantages, too, which I shall

[61] On meaning, meaningfulness, and worth, see Audi (2005), Frankfurt (1999 ch. 7), Smuts (2013), and Wolf (2010). Mulgan (2006) proposes a lexical view that gives priority to freedom.

not discuss: it strikes at a weak spot in Arrhenius's (forthcoming) impossibility theorems,[62] its structure can accommodate lexical intuitions in fixed-population cases,[63] and it might avoid negative analogues of the repugnant conclusion.

In this part of the chapter, I discuss three of the most pressing objections to lexical views. The first, pressed by Jensen (2008), is that lexical views are committed to an implausibly strong form of lexical priority, on which *any* number of lives at some level would be better than any number of lives at some other level. The second, pressed by Arrhenius and Rabinowicz (2005), is that they must assign lexical priority to lives that are only *slightly* better than the lives to which they are superior. And the third, pressed by Huemer (2010), is that they have paradoxical implications in cases of risk.

I argue that all three objections can be mitigated by accepting a *threshold* version of lexical total utilitarianism, which appeals to parity.

3.2.1 Collapse

The lexical view leaves open the following question: is there some number of excellent lives whose existence would be worse than *some* number of mediocre lives? Arrhenius and Rabinowicz (2005) distinguish between two kinds of lexical superiority:[64]

[62] Carlson (manuscript) and Thomas (manuscript) show that Arrhenius's case for assuming that wellbeing levels are "discrete" ignores non-Archimedean theories of wellbeing. Without this assumption, many of his arguments fail.

[63] For example, that no number of lollipop licks would outweigh the badness of one person's torture (Temkin 2012, 34).

[64] The earliest discussion of this distinction that I know of is by Laird (1936, 255), although not with these names. Griffin (1988) emphasizes it under the names "trumping" and "discontinuity."

Strong Superiority: The *X*s are strongly superior to the *Y*s iff any number of *X*s would be better than any number of *Y*s.

Weak Superiority: The *X*s are weakly superior to the *Y*s iff there is some number of *X*s whose existence would be better than any number of *Y*s.

The lexical view only requires weak superiority, which is enough to avoid the repugnant conclusion. And strong superiority is implausible: it seems absurd to think that even a single excellent life would be better than any number of mediocre lives.

This section is about a problem that arises if we accept weak but not strong superiority. I suggest that we solve the problem by rejecting completeness.

One of the major advantages of the lexical view is its compatibility with plausible independence principles about the contributive value of lives. This is, I think, its main advantage over variable value views. The lexical view is compatible with

Separability of People: For any populations *A* and *B*, *A* is better than *B* iff, for any population *X*, adding *A* to *X* would be better than adding *B* to *X*.[65]

We can think of *X* as a context of preexisting (or independently existing) people to which we can add *A* or *B*. *A* is better than *B* iff it's better to add *A* rather than *B* to any such context.

One reason to reject separability would be if we viewed populations as organic unities (for example, Broad 1938, II:692). If populations were organic unities, then

[65] See Broome (2004 ch. 13) for a more precise formulation.

which of two expansions to the population would be better would plausibly depend on what the rest of the population looks like. I have no objection to the view that populations are organic unities, and I have no argument for separability. Separability is usually supported by appealing to versions of the Egyptology objection, but I am only troubled by deontic versions of this objection, which can be avoided by independence of the unaffected. Nonetheless, the most plausible version of the lexical view would accept separability. For if separability is rejected, then the lexical view seems no more, and perhaps less, plausible than variable value views (e.g., Sider 1991). Moreover, total utilitarianism requires separability, and I have argued that the most plausible lexical view is just a version of total utilitarianism. My aim in this chapter is to develop the most plausible lexical view. So I shall assume separability.

Jensen (2008) argues, however, that if we assume separability, then weak superiority collapses to strong superiority. He proves that if *at least as good as* is transitive and complete, then separability and weak superiority entail strong superiority. Suppose that the A-lives are weakly but not strongly superior to the B-lives. For example, suppose that 100 A-lives wouldn't be better than 1 million B-lives, but that 200 A-lives would be better than any number of B-lives. Separability implies that 200 A-lives wouldn't be better than 1 million B-lives plus 100 A-lives, and that 1 million B-lives plus 1 A-lives wouldn't be better than 2 million B-lives. If *at least as good as* is transitive *and complete*, then *not better than* would be transitive. So we could conclude that 200 A-lives wouldn't be better than 2 million B-lives, contrary to what we have supposed.

If we reject completeness, however, *not better than* becomes intransitive: our

slightly improved Bach on page 10 isn't a greater genius than Einstein, who isn't a greater genius than Bach, but slightly improved Bach is a greater genius than Bach. So we can maintain separability and weak superiority without strong superiority.

Nonetheless, we obtain a weaker result. We can distinguish between two kinds of *noninferiority*:

Strong Noninferiority: The *X*s are strongly noninferior to the *Y*s iff, for any numbers *n* and *m*, the existence of *n X*s would not be worse than the existence of *m Y*s.

Weak Noninferiority: The *X*s are weakly noninferior to the *Y*s iff, for some *n*, the existence of *n X*s would not be worse than the existence of any number of *m Y*s.

Strong noninferiority says that no amount of some lower good can outweigh any amount of some higher good. Weak noninferiority says that some amount of some higher good cannot be outweighed by any amount of some lower good.

Separability, weak superiority, and transitivity entail strong noninferiority.[66] Suppose that the *A*-lives are *not* strongly noninferior to the *B*-lives: there is some number *m* of *B*-lives whose existence would be better than some number *n* of *A*-lives. We can then show by mathematical induction that, for any natural number *q*, the existence of *qm B*-lives would be better than *qn A*-lives. The base case, in which $q = 1$, is given: we have supposed that *m B*-lives would be better than *n A*-lives. The inductive step is that, for any natural number *q*, if *qm B*-lives would be better than

[66] The proof is similar to Jensen's.

qn *A*-lives, then $(q+1)m$ *B*-lives would be better than $(q+1)n$ *A*-lives. To prove the inductive step, assume that qm *B*-lives would be better than qn *A*-lives. By separability, $(q+1)m$ *B*-lives would be better than qn *A*-lives plus m *B*-lives: it is better to add qm *B*-lives to a population of m *B*-lives than it is to add qn *A*-lives to that same population. Moreover, the base case implies, by separability, that qn *A*-lives plus m *B*-lives would be better than $(q+1)n$ *A*-lives. By transitivity, $(q+1)m$ *B*-lives would be better than $(q+1)n$ *A*-lives. This proves the inductive step. So, by mathematical induction, for any q, qm *B*-lives would be better than qn *A*-lives. This means that the *A*-lives cannot be weakly superior—or even weakly noninferior—to the *B*-lives. For then there would be some number q such that qn *A*-lives would be better—and, therefore, not worse—than any number, including qm, of *B*-lives. And we have just shown that, without strong noninferiority, this is impossible. Therefore, given separability, weak superiority requires strong noninferiority.

If we were to assume completeness, then strong noninferiority would be unsustainable. For if the existence of any number of *A*-lives would not be worse than the existence of any number of *B*-lives, and if *not worse than* implies *at least as good as*, then, for any n and m, n *A*-lives would be at least as good as m *B*-lives. Take $n = 1$. If a single *A*-life would be better than any number of *B*-lives, then we have strong superiority. If a single *A*-life would be *just as good* as m *B*-lives, then we are in trouble. According to lexical total utilitarianism, if the *B*-lives are worth living, then for any m, $m + 1$ *B*-lives would be better than m *B*-lives. But if $m + 1$ *B*-lives would be better than m *B*-lives, and if m *B*-lives would be just as good as a single *A*-life, then $m + 1$ *B*-lives would be better than a single *A*-life. This contradicts strong noninferiority, according to which no number of *B* lives would be better than any number

of A-lives. So the only option consistent with strong noninferiority when an A-life is at least as good as m B-lives is for the A-life to be better. Therefore, given completeness and separability, strong noninferiority would collapse to to strong superiority.

Proponents of the lexical view, therefore, have the following options. If they accept separability and transitivity, they must accept strong noninferiority. If they also accept completeness, they must accept strong superiority. And if they reject separability, they have little advantage over variable value views. It seems to me that their best option is to accept separability and strong noninferiority, but to reject completeness: strong superiority seems absurd.

Is strong noninferiority at all plausible? I think its plausibility largely depends on one's views about wellbeing and on the kinds of lives that are excellent. Many would independently accept that the best lives *can't* be lived in isolation. On these views, the best things in life involve things like mutual love, friendship, moral virtue, participation in shared projects, and the achievement of shared goals. And these things plausibly require the existence of many other people who participate in these things. On such views, strong noninferiority may not be too bad: it would be, for sufficiently low numbers of people, a vacuous consequence of her theory of wellbeing.

I have suggested that lexical total utilitarians avoid strong superiority by rejecting completeness. But I said on page 84 that lexical total utilitarians impose the following ordering:

Standard Lexical Total Utilitarianism: A is at least as good as B iff either

(a) $i_A > i_B$, or

(b) $i_A = i_B$ and $t_A \geq t_B$,

where i_A is the sum of each A-person's wellbeing in the important dimension(s), and t_A is the sum of each A-person's wellbeing in the trivial dimension(s). This ordering, however, implies strong superiority. If we reject strong superiority, we can instead impose a *lexical threshold*:

Single-Threshold Lexical Total Utilitarianism: A is at least as good as B iff either

(a) $i_A - i_B > \Delta$, or

(b) $i_A \geq i_B$ and $t_A \geq t_B$.[67]

Here Δ represents a lexical threshold: it is the amount of some higher good needed to outweigh any amount of the lower good.[68] The standard lexical ordering is obtained in the special case where $\Delta = 0$. I assume that Δ is a finite value that doesn't vary with the population or other features of the distribution.[69] The lexical threshold makes it possible that neither of two populations is at least as good as the other, because one might have less of the trivial stuff but not sufficiently more of the important stuff to exceed the lexical threshold. This might, for example, be a population of one person whose life is great in every way that doesn't require the existence

[67] Thanks to Teru Thomas for helping me formulate this.

[68] Mulgan (2006) uses "lexical threshold" differently, to mark the lower bound of lexically superior lives.

[69] There might, however, be one threshold for ranking populations and another for evaluating individual lives.

of other people. This population might be on a par with, not better than, a vast population of oyster-like lives.

However, the partial ordering above has counterintuitive consequences when some small gain along the important dimension is not enough to overcome the lexical threshold. Such a gain cannot outweigh any loss along the trivial dimension, however great or small. The "however great" side of this coin is, at least, more plausible than the analogous implication of the standard lexical ordering, which implies strong superiority. But the "however small" side has no appeal. Suppose, for example, that A is better in the important ways than B, but not by enough to exceed the lexical threshold (e.g., our one-person population), and that B is barely better in the trivial ways (e.g., one short oyster-like life). The view under consideration says that A is not at least as good as B. But that seems wrong.

We can fix this problem by imposing an additional threshold δ on the trivial dimension. We can represent B's trivial gain as $0 < t_B - t_A < \delta$. We might say that A's slight edge over B along the important dimension ($0 < i_A - i_B < \Delta$) outweighs this slight loss along the trivial dimension. But, if the trivial loss were much greater, so that it exceeded δ, then A would no longer be better than B, nor would A be worse. We can define δ as the greatest quantity along the trivial dimension that would be outweighed by a quantity of exactly Δ along the important dimension: $(\Delta, 0) > (0, \delta)$, but $(\Delta, 0) \not> (0, \delta + \varepsilon)$, for any $\varepsilon > 0$. We might then formulate the partial ordering as follows:

Multi-Threshold Lexical Total Utilitarianism: A is at least as good as B iff either

(a) $i_A - i_B > \Delta$, or

(b) $i_A \geq i_B$, and

  (i) $t_A \geq t_B$, or

  (ii) $\dfrac{i_A - i_B}{t_B - t_A} > \dfrac{\Delta}{\delta}$.

Condition (a) says that if A is better than B in the important ways by more than Δ, then A is at least as good as B, no matter how much better B is in the trivial ways. This secures weak superiority. (b) then states the two other ways in which A might be at least as good as B. They both require A to be at least as good in the important ways, thereby securing strong noninferiority. (b.i) says that if A is also at least as good in the trivial ways, then A is at least as good as B. This secures the tradeoff constraint, because A and B will never be on a par unless each is better than the other in some way—i.e., unless there is a tradeoff between the important and trivial dimensions. (b.ii) matters when A is better than B in the important ways by less than Δ, but worse than B in the trivial ways. It asks us to compare the ratio of the differences along each dimension to the ratio of each dimension's threshold. If the ratio of the important gain to the trivial loss exceeds the ratio of Δ to δ, then A is at least as good as B. This avoids the untoward consequence of single-threshold lexical total utilitarianism: even small gains along the important dimension can outweigh minuscule losses along the trivial dimension.

Threshold lexical total utilitarians accept weak superiority and strong noninferiority. They reject completeness, which allows them to avoid the collapse to strong superiority without violating separability. They think that dimensions of wellbeing are incommensurable, first, because they cannot be measured on any scale of

values that obeys the Archimedean property, and second, because some tradeoffs between higher and lower goods result in parity. This kind of parity meets the tradeoff constraint: it holds only when one distribution is worse in the trivial ways and is only marginally better in the important ways. Threshold lexical total utilitarians think that some goods cannot substitute for, or compensate for the losses of, other goods.

I argue below that threshold lexical total utilitarians can at least partially resolve two of the most vexing problems for lexical views.

3.2.2 Marginal Differences

Lexical superiority is most plausible when there are differences in kind, not merely of degree. But if the difference between excellent lives and mediocre lives is one of degree, then lexical views may be implausible. They seem forced to draw a sharp cutoff between excellence and mediocrity. But Parfit (2004, 20) points out that there are "fairly smooth continua" between excellence (e.g., Mozart) and mediocrity (e.g., muzak).

We might think that if there is a fairly smooth spectrum from one good to another, then there is no difference in kind between them, because the slightest difference in degree could not generate a difference in kind. But that would be tantamount to the conclusion of a Sorites paradox. There is a fairly smooth spectrum ranging from the hairy to the bald. It seems impossible that a slight change in the amount or distribution of hair could generate a difference in kind—e.g., from the hairy to the not-hairy, or from the not-bald to the bald. But that argument should not lead

us to conclude that there is no difference between the hairy and the bald. Soritical reasoning should not lead us to conclude that there is no difference in kind between Mozart and muzak, or between the lives of flourishing human beings and those of happy oysters or lizards.

We might, however, think that although there may be differences in kind between items on a spectrum of marginal differences, the only axiologically relevant differences are the differences in degree on which the kind-differences supervene. On this view, we should not give lexical weight to the seemingly arbitrary thresholds at which a life becomes excellent, an insight becomes profound, some pain becomes agony, or some creative work constitutes a work of genius. Goodness is a function only of the comparative, degree-based properties in virtue of which things have these vague, absolute properties.

Commonsense morality, however, gives great weight to properties with borderline cases.[70] For example, it may be morally wrong to harvest one innocent person's vital organs in order to save two lives, but morally obligatory to do so for the sake of a million lives. We may have a duty to rescue a nearby child at little cost to ourselves, but no duty to donate nearly all our resources to save a greater number of children on another continent. It may be vague whether some act of consent was informed and freely given, and therefore, sufficient to make some sexual act morally permissible. It may be vague what one knows or intends, and yet the differences between knowledge and ignorance, intent and foresight, may determine which actions are negligent, which are reckless, which are warranted, which are blameworthy, and which make one liable to be harmed.

[70] I borrow these examples from Alexander (2008).

These examples, however, are deontic ones. It might be objected that although binary judgments about permissibility and wrongness may depend on such properties, axiological ones about goodness may not. But many conceptions of the good will confront similar cases. Lives containing both goods and evils may be on the borderline between worth living and not worth living, and plausible non-hedonic components of wellbeing—e.g., knowledge, love, and achievements—have borderline cases. I also suggested on page 34 that population size can be vague because of vagueness in personal identity, and that it can be vague whether something is painful. It seems that we will inevitably have to give great weight in our axiology to vague conditions, so this is problem is not unique to lexical views.

The problem, however, is not the existence of borderlines, but rather their significance for lexical superiority. Arrhenius and Rabinowicz (2005, 136) prove the following:

> Suppose that *at-least-as-good-as* is a complete and transitive relation on the domain. Then, in any finite sequence of objects in which the first element is weakly superior to the last element, there exists at least one element that is weakly superior to its immediate successor.

Arrhenius (2005) finds lexical superiority counterintuitive because he cannot see how something could be even weakly superior to something that is "only marginally worse" (108).

However, it must be shown that we can get from the superior object to the inferior object in finitely many steps that are only *marginally* for the worse. The lexical

theorist might claim that this is not true of her higher and lower goods. She might claim that, in a finite sequence, some of the steps are more significant than they might seem. For example, Griffin (1988, 339) suggests that "we might wish to stop the slide . . . at that point along the line where people's capacity to appreciate beauty, to form deep loving relationships, to accomplish something with their lives beyond just staying alive . . . all disappear." The step from people with these capacities to people without these capacities is, arguably, not a *slight* worsening.

Arrhenius and Rabinowicz, however, might object that we could replace this all-or-nothing step with a longer, finite sequence of lives where each step is less obviously significant. Capacities come in degrees. So we might start with people who have all of Griffin's capacities and gradually reduce them until they disappear. But if our theory assigns lexical superiority to discrete elements—e.g., personal capacities or loving relationships—we might reasonably care much more about the binary presence or absence of those elements than about their gradual waxing or waning. (We should also care about the quality of these things. But increases or decreases in quality might only matter above some threshold where the good exists.) And this binary question may be vague. If the lexical superiority of certain objects in a sequence depends on the lexical superiority of some discrete attribute, then it ought to be vague where the lexical superiority sets in.

This may still seem incredible. Even if it's vague which element in a sequence is weakly superior to its immediate successor, it may seem absurd that there could be such an element. Fortunately, *threshold* lexical total utilitarianism allows us to further weaken this result, by rejecting completeness. Arrhenius and Rabinowicz prove the following:

> Suppose that *at-least-as-good-as* is a transitive relation. If the first element in a finite sequence of objects is [weakly noninferior] to the last element, then there must exist some element in that sequence that is [weakly noninferior] to its immediate successor. (137, n. 16)

This is not, I think, a shocking result. Noninferiority between immediate neighbors along a sequence is still a cost of the lexical view (especially given our earlier result that, given separability, weak noninferiority implies strong noninferiority). But it is far easier to bear than if we had assumed completeness.

The combination of vagueness and incompleteness suggests a somewhat smoother picture than the lexical view may have initially seemed. Excellent lives are weakly superior to mediocre lives. Had we assumed completeness, there would have to be some life along a finite spectrum from excellent lives to mediocre ones that is weakly superior to its immediate successor. If we reject completeness, this life may instead be only *noninferior* to its successor, and it may be vague which lives are noninferior to their successors.

Broome objects to combinations of vagueness and parity. We saw in Section 1.4 that Broome's objection to this combination fails, but that it might survive in a limited form. The multidimensional collapsing principle stated on page 36 would rule out combinations of vagueness and parity due to the imprecise weights of conflicting dimensions. This might endanger appeals to parity that meet the tradeoff constraint. Is this a problem for threshold lexical total utilitarianism?

It depends on the breadth of the multidimensional collapsing principle. The vagueness suggested here is not explained by vagueness in the *weights* of conflicting di-

mensions: one dimension is non-vaguely more important, or weightier, than the other. The vagueness is multidimensional in *some* sense: it only arises because there are multiple dimensions of wellbeing. But the vagueness lies in the value of the lexical threshold(s), and Constantinescu's proposal is restricted to vagueness in the relative weights of conflicting dimensions. If we had an argument for the multidimensional collapsing principle, we could examine whether this argument rules out the threshold lexical total utilitarian's combination of vagueness and parity. But I had no argument for the principle. The closest I had was a question about the same phenomenon—imprecisely weighted dimensions of value—giving rise to two suspiciously similar relations. And there are two distinct phenomena here. Parity arises because some losses cannot be compensated by any gains; the higher goods are nonsubstitutable. But the magnitude of an important loss needed to outweigh any trivial gains is vague, due to vagueness in the lexical threshold(s). So I suspect that threshold lexical total utilitarianism avoids both the arbitrariness and vagueness objections.

3.2.3 Uncertainty

In general, lexical views face problems in cases of uncertainty. Such problems have been most developed in the context of population ethics by Huemer (2010), so I focus on his discussion.

Huemer argues that lexical views in population ethics give "paradoxical" results in cases of risk (333). His argument is aimed at Parfit's perfectionism, but we can here interpret perfectionism as a lexical view about wellbeing, which we can plug

into lexical total utilitarianism. This interpretation is harmless, because Huemer's objection should apply to any lexical view. Huemer offers the following case:

The Two-Charity Case:

> Suppose you have a large sum of money that you intend to donate to a charitable cause. Your only concern is that the money should do the most good. Two causes have attracted your attention. On the one hand, you could donate the money to a poverty relief organization, which you know will result in a large improvement to the welfare of many people who presently have only barely worthwhile lives. On the other hand, you could donate the money to an art school to enable it to expand its operations searching for and supporting unrecognized artistic talent. In the latter event, there is a probability p that, as a result of your support, a new artistic masterpiece would be produced. It seems that, on Parfit's Perfectionist view, if $p = 1$, it would be best to donate to the art school, regardless of how many people would be aided by the poverty relief organization. (338)

Huemer then asks,

Huemer's Question: For what values of p would this remain true?

Before discussing possible answers to this question, let me mention some problems with the example that we may need to fix. First, alleviating poverty has some probability of causing someone to produce a new artistic masterpiece. If we do enough

good for enough people, this probability could exceed some low values of p. Second, alleviating poverty may bring about other perfectionist goods and reduce great suffering, which could plausibly outweigh the production of a single artistic masterpiece. I am, therefore, not confident that Parfit's perfectionism would give the verdict that Huemer assumes it would, even when $p = 1$. And I find it hard to control for these factors by mere stipulation. It is easiest to secure Huemer's verdict if we instead suppose that the relief organization alleviates some minor discomfort of, or bestows some minor pleasure on, some vast number of people. The relief organization might, for example, supply minuscule tubes of anti-itch ointment or tasty lollipops. This makes it more reasonable to donate to the art school when $p = 1$, and to think that lexical views would recommend doing so.

Return to Huemer's question. Huemer considers three possible answers:

The Risk-Intolerant View: For any $p > 0$, it would be best to donate to the art school, regardless of how many people would be aided by the poverty relief organization.

The Risk-Tolerant View: For any $p < 1$, there is some number of people who would be aided by the poverty relief organization such that it would be best to donate to that organization.

The Threshold View: There is some probability $0 < p < 1$ above which it would be best to donate to the art school, regardless of the number of people aided by the poverty relief organization, and below which it would be better to donate to the poverty relief organization.

Huemer then points out a problem for each view. The problem for the risk-

intolerant view is that it seems to imply that it would be best for us to donate all our charitable resources to (not-so-starving) artists—even those with a vanishingly small, but nonzero, probability of churning out a masterpiece—and none to improving the welfare of the least well off. The problem for the risk-tolerant view is that it would make the lexical view irrelevant to practical deliberation, because we are never certain that some action would lead to the creation of a masterpiece. The problem for the threshold view is more complicated. It has to do with conjunctions. Suppose, in

The Three-Charity Case, that there are two art schools. You know that each one has a probability slightly less than p of producing a masterpiece. If you were to donate to them *both*, the probability of a masterpiece being produced would exceed p.

Huemer points out that, according to the threshold view, it would be best for you to donate to both art schools. But each donation, considered separately, would be worse than a donation to some effective poverty relief organization. This violates Huemer's principle that

Two Wrongs Don't Make a Right: If it is inappropriate for S to do A whether or not S does B, and it is inappropriate for S to do B whether or not S does A, then it is inappropriate for S to do A and B.[71]

[71] This should be restricted to finite sets of options, in order to avoid the problems raised by Arntzenius, Elga, and Hawthorne (2004). The ("whether or not") clauses are needed to avoid problems of the kind raised by Jackson (1985, 189).

Huemer leaves the notion of "inappropriateness" very broad, to encompass both moral wrongness and choosing an option that is worse than some available alternative. The threshold view implies that two wrongs make a right, because each of two options taken separately is worse than some alternative, but taking both options collectively is better than all alternatives.

Huemer's objection is that the lexical view is either too restrictive (because it requires us to chase the tiniest risk of a lexically superior outcome), irrelevant to practical deliberation (because lexical superiority is relevant only when we are certain), or otherwise paradoxical (because it implies that two wrongs make a right).

I know of one response to this problem in the literature. Crisp (1992) argues that utilitarians should not try to maximize expected utility. He recommends instead that we live virtuously. But what do the virtues recommend in this case, and why? I agree that, in many cases, the virtues may be the best guide to maximizing wellbeing. But there may be at least some cases in which maximizing expected wellbeing is the best way to maximize wellbeing, both for oneself and for others. This seems plausible when we know the probabilities of all the possible outcomes, when we have lots of time to make a decision, and when certain goods (e.g., those associated with loving relationships or personal projects) are not at stake. And we can imagine a version of Huemer's case in which those conditions obtain.

Threshold lexical total utilitarianism suggests a different response. We might call it

The Weak Threshold View: There is some probability $0 < p < 1$ above which it would be best to donate to the art school, regardless of the number of people

aided by the poverty relief organization, and below which it might not be worse to donate to the poverty relief organization.

Assuming that it's not wrong (or "inappropriate") to choose an option that isn't worse than one's alternatives, the weak threshold view does not imply that two wrongs make a right. The weak threshold view also avoids Huemer's objections to the risk-tolerant and risk-intolerant views. The weak threshold view doesn't imply that the best option is the one with the highest probability, however tiny, of bringing about some higher good. Nor does it have the risk-tolerant view's implication that lexical superiority only matters in cases of certainty.

The threshold view may seem arbitrary, because it's not clear where the value of p comes from. But threshold lexical total utilitarians can derive p from their lexical threshold Δ, which I introduced on page 105. For example, suppose that $\Delta = 5$ on some appropriate scale for the important dimension. Suppose next that some act A has a probability p of realizing some important gain of $i = 10$ (e.g., one masterpiece produced), whereas B will certainly realize a trivial gain of $t = 10,000$ (e.g., ten thousand itches relieved), and that these are one's only options and their only effects. In this case, it's better to do A as long as $p > 0.5$: for then $p(i) > \Delta$. Otherwise, A still isn't *worse*, but might not be better, than B.[72] Threshold lexical total utilitarianism, therefore, leads naturally to the weak threshold view.

Huemer might object, however, that the weak threshold view violates the analogous principle that

[72] A is still better, according to the multi-threshold view on page 106, if $\frac{p(10)}{10,000} > \frac{5}{\delta}$.

Two Rights Don't Make a Wrong: If it is appropriate for *S* to do *A* whether or not *S* does *B*, and it is appropriate for *S* to do *B* whether or not *S* does *A*, then it is appropriate for *S* to do *A* and *B*.

In the three-charity case, let *A* = *not donating to the first art school*, and let *B* = *not donating to the second art school*. The weak threshold view seems to allow us to do *A*, considered on its own, and to do *B*, considered on its own, but not to do *A* and *B*. For if we donate to neither art school, then we miss an opportunity to create a masterpiece with probability exceeding *p*.

The weak threshold view does not, however, imply that two rights make a wrong. For the antecedent is not satisfied in the three-charity case. If we donate to one art school, then we ought to donate to the other, and vice versa. If we don't donate to one of the art schools, then we have no *additional* obligation to donate to the other one, but we violate our obligation to donate to them both (or, we do less good than we can). So we can deny that it's appropriate not to donate to each art school whether or not one donates to the other. The wrong is not made by two rights.

The weak threshold view likely faces other objections. And lexical total utilitarianism may generate other problems in cases of uncertainty, beyond Huemer's objection. But I hope to have shown that, by appealing to parity, threshold lexical total utilitarianism is more plausible in cases of uncertainty than the standard version.

4 Conclusion

I have considered three manifestations of incommensurability in population ethics: parity, incomparability, and lexical superiority.

I began by considering critical-band utilitarianism, which appeals to parity. I found this theory unsatisfactory in several ways: it implies the weak repugnant and sadistic conclusions and is objectionably *ad hoc*, because it violates the tradeoff constraint. I claimed that Broome's theory, too, is *ad hoc*, because it predicts vagueness where we have little reason to expect it.

I, therefore, turned to Bader's view. Bader's appeal to incomparability is more extreme but less arbitrary than critical-band utilitarianism's appeal to parity. But Bader's view implies the weak deontic very repugnant conclusion and is subject to the deontic Egyptology objection. Bader can avoid the latter by accepting independence of the unaffected, but that opens the door to a simpler account of the neutrality of bringing happy people into existence—average utilitarianism—which avoids the weak deontic very repugnant conclusion.

I suggested that average utilitarianism can eschew sums of wellbeing, gives plausible results in infinite cases, and is impartial without being impersonal. I suggested that the other half of the asymmetry might be obtained either through deontic constraints or via asymmetric utilitarianism. The resulting theory says that we should act in such a way that those affected by our act, conditional on having lives worth living, live better lives. I concluded that asymmetric utilitarianism is problematic, but still worth taking seriously.

I then considered two views that accept the mere addition principle, both of which appeal to lexical superiority. I argued that lexical total utilitarianism better explains the repugnance of the single- and many-life repugnant conclusions than Parfit's perfectionism. I then suggested that lexical total utilitarians appeal to parity, in the form of a lexical threshold, to avoid the collapse to strong superiority, to mitigate the significance of seemingly marginal differences, and to avoid paradoxical implications in uncertain cases. Lexical total utilitarianism holds that important losses cannot be compensated for by any trivial gains. Its violation of completeness is explained by incommensurable dimensions of wellbeing, in a way that avoids Broome's objections to critical-band utilitarianism.

Lexical total utilitarianism, however, is a *structural* solution to the repugnant conclusion. Its plausibility depends on whether a reasonable theory of wellbeing fits that structure. This makes it a moving target: we cannot always say whether some implication of the theory is repugnant, because we don't know what the important and trivial dimensions are. I have, however, suggested several possibilities that might yield plausible results. I suggested that lexical total utilitarianism may be plausible if the things that make life most worth living are combinations of goods, global properties of lives, discrete or binary properties, and goods that require the existence of many people.

What I find most attractive about lexical total utilitarianism is its diagnosis of the repugnant conclusion's repugnance. The lexical total utilitarian says that the repugnant conclusion is repugnant because it oversimplifies what makes life worth living. Many philosophers compare the paradoxes of population ethics to Arrow's

(1951) theorem in the theory of social choice.[73] The solution to Arrow's theorem, in the context of social welfare aggregation, is to require more information about each person's good: we cannot get by with merely ordinal information about what each person prefers (Sen 1970a). We need a richer framework of wellbeing. Lexical total utilitarianism extends this insight to variable-population cases: we cannot get by with merely scalar information about each person's good, because no single cardinal scale can accommodate the complexities of what makes life worth living and the vast differences between lives of different qualities. On this view, wellbeing is not like milk, because the dimensions of wellbeing are incommensurable.

I believe that incommensurability can, indeed, help us solve the problems of population ethics, but not in the ways that many have expected.

[73] See Cowen (1996), Kitcher (2000), and Arrhenius (forthcoming).

References

Alexander, Larry. 2008. "Scalar Properties, Binary Judgments." *Journal of Applied Philosophy* 25 (2): 85–104. doi:10.1111/j.1468-5930.2008.00401.x.

Arntzenius, Frank, Adam Elga, and John Hawthorne. 2004. "Bayesianism, Infinite Decisions, and Binding." *Mind* 113 (450): 251–83. doi:10.1093/mind/113.450.251.

Arrhenius, Gustaf. 2000. "An Impossibility Theorem for Welfarist Axiologies." *Economics and Philosophy* 16 (02): 247–66. http://journals.cambridge.org/abstract_S0266267100000249.

———. 2005. "Superiority in Value." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 123 (1/2): 97–114. doi:10.2307/4321574.

———. forthcoming. *Population Ethics*. Oxford, UK: Oxford University Press. http://people.su.se/~guarr/FGkurs/FG%20%20Kurs%202013.pdf.

Arrhenius, Gustaf, and Wlodek Rabinowicz. 2005. "Millian Superiorities." *Utilitas* 17 (02): 127–46. doi:10.1017/S0953820805001494.

———. 2015. "The Value of Existence." In *The Oxford Handbook of Value Theory*, edited by Iwao Hirose and Jonas Olson. Oxford University Press. //www.oxfordhandbooks.com/10.1093/oxfordhb/9780199959303.001.0001/oxfordhb-9780199959303-e-23.

Arrow, Kenneth Joseph 1921-. 1951. *Social Choice and Individual Values*. Monograph (Yale University. Cowles Foundation for Research in Economics). New York : London: New York : Wiley.

Audi, Robert. 2005. "Intrinsic Value and Meaningful Life." *Philosophical Papers* 34 (3): 331–55. doi:10.1080/05568640509485162.

Bader, Ralf M. manuscript. "Neutrality and Conditional Goodness."

Barry, Brian. 1977. "Rawls on Average and Total Utility: A Comment." *Philosophical Studies* 31 (5): 317–25.

Bennett, Jonathan. 1978. "On Maximising Happiness." http://www.earlymoderntexts.com/jfb/maxhap.pdf.

Blackorby, Charles, Walter Bossert, and David Donaldson. 1996. "Quasi-Orderings and Population Ethics." *Social Choice and Welfare* 13 (2): 129–50. http://link.springer.com/article/10.1007/BF00183348.

Blackorby, Charles, Walter Bossert, and David J. Donaldson. 2005. *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. 39. Cambridge University Press.

Bostrom, Nick. 2011. "Infinite Ethics." *Analysis and Metaphysics*, no. 10: 9–59.

Broad, C. D. 1938. *An Examination of McTaggart's Philosophy*. Vol. II. Cambridge: Cambridge University Press.

Broome, John. 1991. "'Utility'." *Economics and Philosophy* 7 (01): 1–12. doi:10.1017/S0266267100000882.

———. 1999. *Ethics Out of Economics*. Cambridge [England] ; New York: Cambridge University Press.

———. 2004. *Weighing Lives*. Oxford ; New York: Oxford University Press.

———. 2007. "Reply to Qizilbash*." *Philosophy and Phenomenological Research* 75 (1): 152–57. http://onlinelibrary.wiley.com/doi/10.1111/j.1933-1592.2007.00064.x/full.

———. 2009. "Reply to Rabinowicz." *Philosophical Issues* 19 (1): 412–17. http://onlinelibrary.wiley.com/doi/10.1111/j.1533-6077.2009.00175.x/full.

Carlson, Erik. 1998. "Mere Addition and Two Trilemmas of Population Ethics." *Economics and Philosophy* 14 (02): 283. doi:10.1017/S0266267100003862.

———. 2004. "Broome's Argument Against Value Incomparability." *Utilitas* 16 (2): 220–24. doi:10.1017/S0953820804000548.

———. 2007. "Higher Values and Non-Archimedean Additivity." *Theoria* 73 (1): 3–27. http://onlinelibrary.wiley.com/doi/10.1111/j.1755-2567.2007.tb01185.x/abstract.

———. 2010. "Parity Demystified." *Theoria* 76 (2): 119–28. doi:10.1111/j.1755-2567.2010.01063.x.

———. manuscript. "On Some Impossibility Theorems in Population Ethics." Unpublished manuscript.

Carter, Alan. 1999. "Moral Theory and Global Population." In *Proceedings of the Aristotelian Society*, 289–313. JSTOR. http://www.jstor.org/stable/4545311.

Chang, Ruth. 2002. "The Possibility of Parity." *Ethics* 112 (4): 659–88. doi:10.1086/339673.

———. 2013. "Incommensurability (and Incomparability)." Edited by Hugh

LaFollette. *The International Encyclopedia of Ethics*. Blackwell Publishing Ltd.

———. 2014. *Making Comparisons Count*. Routledge.

Chipman, John S. 1960. "The Foundations of Utility." *Econometrica* 28 (2): 193. doi:10.2307/1907717.

Constantinescu, Cristian. 2012. "Value Incomparability and Indeterminacy." *Ethical Theory and Moral Practice* 15 (1): 57–70. doi:10.1007/s10677-011-9269-8.

Cooper, Neil. 1995. "Paradox Lost: Understanding Vague Predicates." *International Journal of Philosophical Studies* 3 (2): 244–69.

Cowen, Tyler. 1996. "What Do We Learn from the Repugnant Conclusion?" *Ethics*, 754–75. http://www.jstor.org/stable/2382033.

Crisp, Roger. 1992. "Utilitarianism and the Life of Virtue." *The Philosophical Quarterly* 42 (167): 139. doi:10.2307/2220212.

Egre, Paul, and Nathan Klinedinst. 2010. *Vagueness and Language Use*. Palgrave Macmillan.

Eliaz, Kfir, and Efe A. Ok. 2006. "Indifference or Indecisiveness? Choice-Theoretic Foundations of Incomplete Preferences." *Games and Economic Behavior* 56 (1): 61–86. doi:10.1016/j.geb.2005.06.007.

Fehige, Christoph, and Ulla Wessels. 1998. *Preferences*. Berlin; New York: W. de Gruyter. http://site.ebrary.com/id/10585223.

Fleurbaey, Marc, and Alex Voorhoeve. 2015. "On the Social and Personal Value of Existence." In *Weighing and Reasoning: Themes from the Philosophy of John Broome*, edited by Iwao Hirose and Andrew Reisner. Oxford University Press. http://personal.lse.ac.uk/VOORHOEV/On%20the%20social%20and%20personal%20value%20of%20existence%20version%20to%20submit.pdf.

Frankfurt, Harry G. 1999. *Necessity, Volition, and Love*. Cambridge, U.K. ; New York: Cambridge University Press.

Fredrickson, Barbara L., and Marcial F. Losada. 2005. "Positive Affect and the Complex Dynamics of Human Flourishing." *The American Psychologist* 60 (7): 678–86. doi:10.1037/0003-066X.60.7.678.

Goodrich, Jimmy. 2014. "Imprecision in Population Ethics." Undergraduate Thesis, Rutgers University-New Brunswick.

Griffin, James. 1979. "Is Unhappiness Morally More Important Than Happiness?" *The Philosophical Quarterly* 29 (114): 47. doi:10.2307/2219182.

———. 1988. *Well-Being*. Oxford University Press. http://www.oxfordscholarship.com/view/10.1093/0198248431.001.0001/acprof-9780198248439.

Grinsell, Timothy Wood. 2013. "Avoiding Predicate Whiplash: Social Choice Theory and Linguistic Vagueness." In *Proceedings of SALT*, 22:424–40. http://elanguage.net/journals/salt/article/download/22.424/3479.

Gurney, Edmund. 1887. *Tertium Quid: Chapters on Various Disputed Questions*. https://archive.org/details/tertiumquidchap00gurngoog.

Harsanyi, John C. 1953. "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking." *Journal of Political Economy* 61. https://ideas.repec.org/a/ucp/jpolec/v61y1953p434.html.

Hayes, Peter. 2015. *How Was It Possible?: A Holocaust Reader*. U of Nebraska Press.

Heyd, David. 1988. "Procreation and Value Can Ethics Deal with Futurity Problems?" *Philosophia* 18 (2): 151–70. http://www.springerlink.com/index/1222K814631X546U.pdf.

Huemer, Michael. 2008. "In Defence of Repugnance." *Mind* 117 (468): 899–933. http://mind.oxfordjournals.org/content/117/468/899.short.

———. 2010. "Lexical Priority and the Problem of Risk." *Pacific Philosophical Quarterly* 91 (3): 332–51. doi:10.1111/j.1468-0114.2010.01370.x.

Hurka, Thomas. 1983. "Value and Population Size." *Ethics*, 496–507. http://www.jstor.org/stable/2380627.

Hyde, Mr Dominic. 2012. *Vagueness, Logic and Ontology*. Ashgate Publishing, Ltd.

Jackson, Frank. 1985. "On the Semantics and Logic of Obligation." *Mind* 94 (374): 177–95.

Jensen, Karsten Klint. 2008. "Millian Superiorities and the Repugnant Conclusion." *Utilitas* 20 (03). doi:10.1017/S0953820808003154.

Kagan, Shelly. 1994. "Me and My Life." *Proceedings of the Aristotelian Society* 94 (n/a): 309–24.

———. 2009. "Well-Being as Enjoying the Good." *Philosophical Perspectives* 23 (1): 253–72. doi:10.1111/j.1520-8583.2009.00170.x.

Kamp, Hans, and Galit Sassoon. 2015. "Vagueness." In *Cambridge Handbook of Formal Semantics*, edited by Maria Aloni and Paul Dekker. Cambridge University Press.

Katznelson, Yitzhak. 2004. *An Introduction to Harmonic Analysis*. Cambridge University Press.

Kavka, Gregory S. 1975. "Rawls on Average and Total Utility." *Philosophical Studies* 27 (4): 237–53.

Kennedy, Christopher. 2007. "Modes of Comparison." In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, 43:141–65. Chic Ling Society. http://cls.metapress.com/index/913k657037422361.pdf.

———. 2011. "Vagueness and Comparison." *Vagueness and Language Use*, 73–97. http://semanticsarchive.net/Archive/WU1NDI3Y/vaguenessandcomparison.pdf.

Kitcher, Philip. 2000. "Parfit's Puzzle." *Noûs* 34 (4): 550–77. doi:10.1111/0029-4624.00278.

Klein, Ewan. 1980. "A Semantics for Positive and Comparative Adjectives." *Linguistics and Philosophy* 4 (1): 1–45. http://link.springer.com/article/10.1007/BF00351812.

Laird, John. 1936. *An Enquiry Into Moral Notions*. New York,Ams Press.

Lewis, David. 1988. "Vague Identity: Evans Misunderstood." *Analysis* 48 (3): 128–30. doi:10.1093/analys/48.3.128.

List, Christian. 2004. "Multidimensional Welfare Aggregation." *Public Choice* 119 (1-2): 119–42. doi:10.1023/B:PUCH.0000024168.00362.af.

McMahan, Jefferson. 1981. *Problems of Population Theory*. JSTOR. http://www.jstor.org/stable/2380707.

McTaggart, J. Ellis. 1927. "The Nature of Existence, Volume II." https://ia600400.us.archive.org/29/items/natureofexistenc02mctauoft/natureofexistenc02mctauoft.pdf.

Morton, Adam. 1994. "Two Places Good Four Places Better." *Proceedings of the Aristotelian Society, Supplementary Volumes* 68: 187–98.

Mulgan, Tim. 2006. *Future People: A Moderate Consequentialist Account of Our Obligations to Future Generations*. Oxford; New York: Clarendon Press ; Oxford University Press.

Nagel, Thomas. 2012. *Mortal Questions*. Cambridge University Press.

Narveson, Jan. 1973. "Moral Problems of Population." *The Monist* 57 (1): 62–86.

Ng, Yew-Kwang. 1989. "What Should We Do About Future Generations?" *Economics and Philosophy* 5 (02): 235–53. doi:10.1017/S0266267100002406.

Ok, Efe A. 2002. "Utility Representation of an Incomplete Preference Relation." *Journal of Economic Theory* 104 (2): 429–49. https://ideas.repec.org/a/eee/jetheo/v104y2002i2p429-449.html.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford [Oxfordshire]: Clarendon Press.

———. 2004. "Overpopulation and the Quality of Life." In *The Repugnant Conclusion*, 7–22. Springer. http://link.springer.com/content/pdf/10.1007/978-1-4020-2473-3_2.pdf.

———. manuscript. "Can We Avoid the Repugnant Conclusion."

Portmore, Douglas W. 1999. "Does the Total Principle Have Any Repugnant Implications?" *Ratio* 12 (1): 80–98. doi:10.1111/1467-9329.00078.

Qizilbash, Mozaffar. 2007. "The Mere Addition Paradox, Parity and Vagueness*." *Philosophy and Phenomenological Research* 75 (1): 129–51. http://onlinelibrary.wiley.com/doi/10.1111/j.1933-1592.2007.00063.x/full.

Rabinowicz, Wlodek. 2008. "Value Relations." *Theoria* 74 (1): 18–49. doi:10.1111/j.1755-2567.2008.00008.x.

———. 2009. "Broome and the Intuition of Neutrality." *Philosophical Issues* 19 (1): 389–411. http://onlinelibrary.wiley.com/doi/10.1111/j.1533-6077.2009.00174.x/full.

———. 2012. "Value Relations Revisited." *Economics and Philosophy* 28 (02): 133–64. doi:10.1017/S0266267112000144.

Raffman, Diana. 2011. "Vagueness and Observationality." In *Vagueness: A Guide*, edited by Giuseppina Ronzitti, 107–21. Dordrecht: Springer Netherlands. http://www.springerlink.com/index/10.1007/978-94-007-0375-9_5.

Rashdall, H. 1902. "The Commensurability of All Values." *Mind*, New Series, 11 (42): 145–61. http://www.jstor.org/stable/2248411.

Rawls, John. 1999. *A Theory of Justice*. Cambridge, Mass.: Belknap Press of Harvard University Press.

Riemann, Bernhard. 1867. *Ueber Die Darstellbarkeit Einer Function Durch Eine Trigonometrische Reihe*. Edited by Richard Dedekind. In der Dieterichschen Buchhandlung.

Ross, W. D. 1939. *Foundations of Ethics*. Oxford University Press.

———. 2002. *The Right and the Good*. Edited by Philip Stratton-Lake. Oxford: Clarendon Press.

Ryberg, Jesper. 1996. "Parfit's Repugnant Conclusion." *The Philosophical Quarterly* 46 (183): 202. doi:10.2307/2956387.

Sassoon, Galit. 2013a. *Vagueness, Gradability and Typicality: The Interpretation of Adjectives and Nouns*. BRILL.

———. 2013b. "A Typology of Multidimensional Adjectives." *Journal of Semantics* 30 (3): 335–80. doi:10.1093/jos/ffs012.

Scheffler, Samuel. 1994. *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*. Rev. ed. Oxford : New York: Clarendon Press ; Oxford University Press.

———. 2013. *Death and the Afterlife*. Oxford University Press.

Schoenfield, Miriam. 2012. "Imprecision in Normative Domains." MIT.

Sen, Amartya. 1970a. *Collective Choice and Social Welfare*. San Francisco: Holden-Day San Francisco.

———. 1970b. "Interpersonal Aggregation and Partial Comparability." *Econometrica* 38 (3): 393. doi:10.2307/1909546.

———. 1980. "Plural Utility." *Proceedings of the Aristotelian Society*, New Series, 81 (ArticleType: research-article / Full publication date: 1980 - 1981 / Copyright © 1980 The Aristotelian Society): 193–215. doi:10.2307/4544973.

———. 1997. *On Economic Inequality*. Enlarged edition. Oxford: Oxford University Press.

Sider, Theodore. 1993. "Asymmetry and Self-Sacrifice." *Philosophical Studies* 70 (2): 117–32. doi:10.1007/BF00989586.

Sider, Theodore R. 1991. "Might Theory X Be a Theory of Diminishing Marginal Value?" *Analysis* 51 (4): 265–71. http://analysis.oxfordjournals.org/content/51/4/265.full.pdf.

Sidgwick, Henry. 1874. *The Methods of Ethics*. Seventh Edition. Indianapolis: Hackett Pub. Co.

Smuts, Aaron. 2013. "Five Tests for What Makes a Life Worth Living." *The Journal of Value Inquiry* 47 (4): 439–59. doi:10.1007/s10790-013-9393-x.

Temkin, Larry S. 2012. *Rethinking the Good: Moral Ideals and the Nature of Prac-*

tical Reasoning. Oxford Ethics Series. Oxford ; New York: Oxford University Press.

Temkin, Larry S. 1993. *Inequality*. New York: Oxford University Press.

Thomas, Teruji. manuscript. "Some Possibilities in Population Axiology." Unpublished manuscript.

van Rooij, R. 2011. "Measurement and Interadjective Comparisons." *Journal of Semantics* 28 (3): 335–58.

von Neumann, John, and Oskar Morgenstern. 2007. *Theory of Games and Economic Behavior*. Edited by Ariel Rubinstein and Harold William Kuhn. 60th Anniversary Commemorative edition. Princeton, N.J. ; Woodstock: Princeton University Press.

Wolf, Susan R. 2010. *Meaning in Life and Why It Matters*. The University Center for Human Values Series. Princeton, N.J: Princeton University Press.