Totalism

Tomi Francis

St John's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

in Philosophy

November 22, 2022

Abstract

Totalism

Tomi Francis

St John's College, University of Oxford

Submitted for the degree of *Doctor of Philosophy* in Philosophy

November 22, 2022

Word count: **63177**

Totalism in population axiology is the view that one population is better than another if and only if it contains greater total wellbeing. This thesis defends Totalism. I begin, in Chapter 1, by arguing for the principle of Anonymity, according to which any two populations with the same anonymous distribution of wellbeing are equally good. Chapter 2 argues against Prioritarian population axiologies on the basis that they imply the *desirability of welfare diffusion*: the implausible claim that it can be better for there to be less total wellbeing spread thinly between a larger total number of people. Chapter 3 argues against the popular intuition that adding good lives within a certain range cannot result in a better population. Chapters 4 and 5 provide new evidence for the unavoidability of the Repugnant Conclusion, the controversial implication of Totalism which holds that a smaller population of excellent lives can be worse than a much larger population of mediocre lives (and which is traditionally considered to be a damning objection to Totalism). Chapter 6 argues for a version of Totalism applicable to prudence in fission cases: cases where an individual "splits" into two or more successors, while retaining what matters in survival. Finally, Chapter 7 ties things together and summarises several direct arguments for Totalism. With the exception of Chapter 7, which makes extensive use of results and notation from earlier parts of the thesis, each chapter is entirely self-standing and can be read independently of the others.

Acknowledgements

I owe thanks to many people for making this thesis possible. At one extreme, if certain views about the fragility of identity and the chaotic nature of the world are correct, I should thank virtually every individual who was born on earth significantly before me, without whom I would not have existed (and therefore would not have written this thesis). But let me be more selective.

I am very grateful to my supervisors, Dr. Teruji Thomas and Prof. Hilary Greaves for their consistently high-quality feedback, advice and encouragement over the last four years.

I certainly would not have been able to write this thesis without financial support from the Arts and Humanities Research Council (grant AH/L503885/1), the Global Priorities Institute, the Forethought Foundation for Global Priorities Research, and Longview Philanthropy.

I have received valuable feedback and discussion on parts of this thesis from too many friends and colleagues to easily list here. I am grateful to them all, but special thanks go to Kacper Kowalczyk, who has taught me more about ethics than perhaps any other philosopher; Petra Kosonen, who has always been an exceptional source of new ideas and arguments; Johan Gustafsson, who has taught me virtually everything I know about decision theory; and Elliott Thornley, who has helped me with my philosophical difficulties more times than I can count. More than any other colleague, I have to thank Todd Karhu. Todd has read and provided comments on a truly supererogatory amount of my work, often at very short notice, and has given me more help and encouragement than anyone could have reason to expect.

I could not have done any of this without my family, especially my mother, Rosalind Francis, who put in an incredible amount of time and effort to raise me and my brothers.

Last, but certainly not least, I am grateful to Biqing Wang for keeping me company over the COVID-19 pandemic, giving me something to work for, and believing in me more than anyone else.

Contents

Introduction

Population axiology concerns the evaluation of outcomes in terms of their overall betterness, when these outcomes can differ with respect to the identities or the numbers of the people they contain. This thesis defends the following view in population axiology:

> *Totalism* For any populations X and Y, $X \succeq Y$ if and only if X contains at least as much total wellbeing as Y.

Totalism is a simple, elegant and powerful theory. Yet it has a number of extremely controversial implications. Among them are

(i) Differences in the identities of people make no evaluative difference. In particular, other things being equal, it makes no evaluative difference whether we make independently existing people better-off, or instead create better-off people in place of different, worse-off people.

(*Anonymity*)

(ii) Equality of wellbeing is not intrinsically valuable, nor do benefits to the worse-off get any axiological priority over benefits to the better-off. Pure, non-rank-switching transfers of wellbeing from better-off to worse-off people do not make an outcome better.

(*Indifference to Inequality*)

(iii) Creating happy people can make an outcome better, just as creating unhappy people makes an outcome worse.

(*Procreation Symmetry*)

(iv) Any population consisting of many excellent lives is worse than some much larger population consisting entirely of lives barely worth living.

(*The Repugnant Conclusion*)

Chapters 1–5 directly or indirectly defend these controversial premises.

Chapter 1, *Anonymity and Non-Identity Cases*, argues for Anonymity by considering what we should say about *non-identity cases*. In such cases, we can choose whether to create a better-off person or group of people, or instead create a different, worse-off person or group of people. Most of us believe that in non-identity cases it would be better to create the better-off person or group of people; I call this the *Non-Identity Principle*. Chapter 1 shows that Anonymity follows from the Non-Identity Principle. It also provides an argument for the Non-Identity Principle, for the benefit of those who do not find it initially compelling.

Chapter 2, *The Welfare Diffusion Objection to Prioritarianism*, argues against Prioritarian population axiologies on the basis that they imply the *desirability of welfare diffusion*: the implausible claim that it can be better for there to be less total wellbeing spread thinly between a larger total number of people. It shows that Prioritarian axiologies imply this claim if they satisfy

the *Mere Addition Principle*, on which adding good lives to a population cannot make things worse. Additionally, it argues that Prioritarians are indirectly committed to the Mere Addition Principle via their commitment to the unimportance of relativities between people's wellbeing levels. Indirectly, then, Chapter 2 argues for Indifference to Inequality. It also contains a direct argument for a restricted version of Totalism, which applies to populations containing only good or neutral lives. This argument is extended to support full-blown Totalism in Chapter 7.

Chapter 3, *In Favour of Making Happy People*, argues for Procreation Symmetry. It argues that the popular *Principle of Neutrality*, which holds that adding good lives within a certain range cannot result in a better population, is inconsistent with the *Absolute Value Principle*, which holds that every population consisting entirely of bad lives must be worse than every population consisting entirely of good lives. It makes the claim that because the Absolute Value Principle is more compelling than the Principle of Neutrality, we should reject the latter principle. Chapter 3 also treats the cases of non-transitive and option-set-dependent value, and considers whether we might have *moral obligations* to create happy people. It argues that we do have such moral obligations, *provided* that all else is equal. This does not immediately imply that in practice, we have a moral obligation to have children with expectedly good lives, because in such practical cases, all else is far from equal.

Chapters 4 and 5 provide new evidence for the unavoidability of the Repugnant Conclusion. Chapter 4, *Repugnance Without Mere Addition*, provides an argument which shows that an additive version of the Repugnant Conclusion follows from several compelling *interpersonal* principles: principles which directly concern comparisons of populations, such as the Absolute Value Principle. These premises are logically weaker and more compelling than those of any other arguments for the Repugnant Conclusion (often called "impossibility theorems" by those who find the Repugnant Conclusion incredible) I am aware of in the present literature. In return, the additive version of the Repugnant Conclusion proved in Chapter 4 is somewhat logically weaker than the non-additive versions featuring in some other impossibility theorems. (However, the additive version of the Repugnant Conclusion seems to me just as "repugnant" as its non-additive counterpart.)

Chapter 5, *Intrapersonal Arguments for the Repugnant Conclusion*, provides a stronger version of an *intrapersonal* argument for the Repugnant Conclusion recently given by Jacob Nebel (2019). Intrapersonal arguments derive interpersonal conclusions by appealing to principles regarding which prospects are better or worse for individuals, together with versions of the *Ex Ante Pareto* principle, which says that one prospect is better than another if it is better for each person. The strengthening concerns Nebel's most controversial premise, the *Probable Addition Principle*, which roughly says that giving a person an additional chance of existence at a neutral life is

equally good for them. In particular, it involves replacing this principle with the much more compelling *Conditional Value Principle*, which holds that a prospect which guarantees that a person will have a good life *if* they exist must be better for them than a prospect which guarantees that they have a bad life *if* they exist. It also shows that, given one extremely minimal condition on betterness for individuals, the Ex Ante Pareto Principle alone implies an intuitively "repugnant" conclusion.

Chapters 6 and 7 take a different tack. Rather than defending the controversial implications of Totalism, they argue for versions of Totalism directly. Chapter 6, *Prudence in Different-Number Fission Cases*, argues for *Fission Totalism*, which is a version of Totalism applicable to prudence in fission cases: cases where an individual "splits" into two or more successors, while retaining what matters in survival. It first provides an argument for Fission Totalism which involves an analogue of Harsanyi's (1955) Aggregation Theorem, which roughly says that (Fission) Totalism, restricted to cases involving the same number of people, follows from three principles: the Ex Ante Pareto principle, a minimal condition on rationality, and the principle of Anonymity. This argument establishes Fission Totalism where the total amount of well-being is measured on the scale generated by Expected Utility Theory: that is, where the scale is chosen such that by definition, it is best, when facing a case of risk, to maximise the expectation of this function. An additional argument establishes Fission Totalism on a conceptually distinct scale of wellbeing: the

life-years scale, where amounts of wellbeing are directly proportional to years of good life. Since these arguments are both sound only if the two conceptually distinct scales of wellbeing are co-extensive, Chapter 6 also contains another direct argument for the coincidence of these two scales of wellbeing; with this argument in place, only one of the two prior arguments need to succeed in order to establish Fission Totalism on both scales of wellbeing.

Finally, Chapter 7, *Concluding Arguments*, summarises the results of the previous chapters and provides several direct arguments for Totalism in population axiology.

Let me bring this introduction to a close by mentioning several standard, but arguable, assumptions I make during the course of this thesis. First, I shall assume throughout that it is always possible to make *ceteris paribus* comparisons of populations (or fission populations, in the case of Chapter 6). This assumption is often not even noticed, and when it is, it is usually taken to be so obviously correct that it can be passed over with little to no comment. I'm generally happy to make this assumption, as it makes my life much easier, but I'm not so sure that it should be taken for granted in a complete treatment of population axiology. The best argument for this assumption that I know of proceeds from the premise that, if two outcomes are equally good for each person, they must be equally good overall. I think that this premise is true, but it is logically stronger than necessary, and it is not exactly unquestionable. There might be other things that matter, eval-

uatively speaking, other than wellbeing. If there are, no obvious argument comes to my mind for the claim that, although non-wellbeing considerations matter, they *must* matter *independently* of wellbeing, so that we can safely separate wellbeing-related and non-wellbeing-related considerations.

Second, except during Chapter 3, I shall generally assume that the at-least-as-good-as relation on populations is transitive and option-set-independent. That is, if $X \succeq Y$ and $Y \succeq Z$, then $X \succeq Z$; and any two populations can be evaluatively compared independently of the set of other feasible alternatives which are available. Both of these assumptions have been questioned by population ethicists in recent years. A full treatment of population axiology ought to explain why we should (or should not) accept these assumptions. I have not taken on this task in this thesis. Since these are very standard assumptions, I shall be happy enough if my conclusion is a conditional one: Totalism is true *if* the betterness relation is transitive and option-set-independent.

Lastly, throughout this thesis I only treat finite populations and prospects. There are three important kinds of infinity I ignore. The first is the possibility of there existing infinitely many people. The second is the possibility of there being infinite quantities of wellbeing enjoyed by single individuals. The third is the possibility of there being prospects of *infinite support*, i.e., prospects which assign positive probability to infinitely many outcomes. Worryingly, arguments from infinite ethics challenge some of the most important basic

foundational principles supporting Totalism, such as Anonymity, principles of rational choice in cases of risk, and Ex Ante Pareto principles.

This thesis focuses exclusively on finite cases for two reasons. First, it is notoriously hard to find acceptable theories of infinite ethics. Perhaps it is entirely intractable. Second, considering these problems in the depth they deserve would leave little or no space for finite population axiology, whereas considering them in a shallow way would be worse than doing nothing. It is better to acknowledge and then pass over a problem than to give a half-baked solution. Neither of these are good reasons to avoid considering the problems of infinite ethics. But perhaps they are good reasons for not treating these problems in one's DPhil thesis.

Chapter 1

Anonymity and Non-Identity Cases

Abstract

I argue for the principle of Anonymity, according to which two populations are equally good whenever they have the same anonymous distribution of wellbeing. I first show that, given transitivity of the at-least-as-good-as relation, Anonymity is entailed by the "Non-Identity Principle", according to which the consequence of bringing better rather than worse lives into existence is, all else equal, better. I then argue for the Non-Identity Principle on the basis that if it were false, it would follow that we fail to improve the world when we make existing people better off, while at the same time replacing worse-off future people with different better-off future people. Since this is very implausible, we should accept the Non-Identity Principle, and therefore Anonymity as well.

According to the principle of *Anonymity*, any two populations with the same anonymous distribution of wellbeing are equally good. Although it might appear innocuous, Anonymity is a substantive and powerful principle. It implies that differences in the identities of those who exist are evaluatively irrelevant and thereby simplifies population axiology significantly. Perhaps partly for this reason, Anonymity is often assumed in formal treatments of population axiology, sometimes without much in the way of supporting argument.[1] But Anonymity is not obviously true, because it is not obviously true that differences in identity make no all-things-considered evaluative difference.[2] If we are going to accept Anonymity, we had better have a good argument for doing so; my aim in this chapter is to provide one.

The crux of my argument for Anonymity is a compelling principle about value in the non-identity cases first discussed by Parfit (1984: ch. 17), in which we can either bring about the existence of a better-off person, or alternatively bring about the existence of a different, worse-off person. According to the "Non-Identity Principle", in such cases, the outcome in which the better-off person comes to exist is better overall. The main technical result of this chapter is that given some standard structural assumptions, the

[1]See for instance Blackorby and Donaldson 1984: 14, Blackorby et al. 2003: 346–47, Broome 2004: 135–136, and Asheim and Zuber 2014: 632.

[2]For example, Anonymity conflicts with evaluative versions of McMahan's (2013) "Weak Asymmetry" between comparative and non-comparative benefits, with Otsuka's (2018) view that it is in itself regrettable for people to be worse off than they might have been, and with various person-affecting views.

Non-Identity Principle implies Anonymity. After demonstrating this result, I shall argue that if the Non-Identity Principle were false, it would follow that replacing worse-off for better-off people, while at the same time improving the lives of existing people, often fails to make things better overall. Since such combinations of changes clearly do make things better overall, the Non-Identity Principle must be true.

Before we turn to the argument from the Non-Identity Principle, let me clarify the background framework in which we shall operate. I take the notion of a life to be primitive, where lives are understood to be individuated by the people living them, and the sorts of lives they are. I assume that there is an "at-least-as-good-as" relation on lives, which is transitive and reflexive; the "betterness", "equal goodness" and other evaluative relations on lives are defined from the at-least-as-good-as relation in the usual way. As shown by Arrhenius 2011, we can identify wellbeing levels with equivalence classes of lives under the equal goodness relation. In practice, I shall represent wellbeing levels by numbers, but this is for ease of presentation only.

The objects of comparison are populations, which I take to be finite sets of lives in which no person appears twice. The relation we are interested in is the at-least-as-good-as relation $\succeq$ on populations, with the derived relations of betterness, equal goodness and so on again being defined in the usual way. I shall allow that $\succeq$ may be incomplete: that is, there may be populations X and Y, neither of which is at least as good as the other. I shall,

however, make two important assumptions about $\succeq$: I shall assume that it is option-set-independent (in the sense that how two populations compare does not depend on any underlying option set in which they are considered) and also transitive.[3] Although I believe that we should accept choice-set-independence and transitivity, I shall not argue for them in this chapter. Those who doubt one or both of these assumptions, and also find Anonymity to be incredible, might take the lesson of this chapter to be that we had better reject transitivity or option-set-independent betterness in order to maintain the Non-Identity Principle without incurring Anonymity.

Anonymity, stated more precisely, is the claim that if two populations are *Anonymously Equivalent*, meaning that we can find a wellbeing-preserving bijection between the lives in each of the two populations, these two populations are equally good.

That is enough in the way of background; on to the argument from the Non-Identity Principle to Anonymity. Consider now a standard non-identity case, in which one must choose between bringing Eve into existence, with a better life, or instead bringing Adam into existence, with a worse (but still good) life. Imagine that nobody else is affected by this choice, and there are no other potentially morally relevant differences at play: Adam and Eve

[3]The option-set-independence of betterness is disputed by Roberts 2003, 2011, Voorhoeve 2013 and Frick 2014, 2022. The rejection of transitivity is discussed at length by Temkin 2012, and is argued for by Temkin 1987, 1996, and Rachels 1998, 2001, 2004. One standard defence of transitivity can be found in Broome 2004: 50–51.

are equally deserving, no impersonally valuable things are affected, and so on. (A *ceteris paribus* clause like this should be taken to apply to all further cases and principles appealed to in this chapter.)

One thing I think we should say about this case is that it would be better overall for Eve to exist with the better life, rather than for Adam to exist with the worse life. Note that this claim is purely evaluative. I am not claiming that it would be wrong to bring Adam into existence, or even that we would have a moral reason to bring Eve into existence. These two claims are also quite plausible, but are somewhat more controversial than the evaluative claim.[4]

It also seems that it would be better for Eve to exist rather than Adam regardless of whether there also exist any number of people, with any configuration of wellbeing levels, provided that these people are entirely unaffected by the choice. More generally, it seems that we should accept the

> *Weak Non-Identity Principle* If life l_1 is better than life l_2, then for any population X not including the l_1 or l_2 persons, $X + l_1$ is better than $X + l_2$.

The Weak Non-Identity Principle supports the stronger claim that, additionally, if lives l_1 and l_2 are equally good, then $X + l_1$ is equally as good as

[4]Boonin provides an influential defence of the view that we have no moral obligation to bring about better lives in non-identity cases. But even he admits that in his model non-identity case, "conceiving [the better off person] would produce significantly better consequences" (2014: 151).

$X+l_2$. This is because the Weak Non-Identity Principle implies that the balance between the populations $X+l_1$ and $X+l_2$ is sensitive to arbitrarily small improvements and deteriorations: it would be better to add a slightly better version of l_1 than to add l_2, and vice versa. This kind of sensitivity is widely (and correctly) held to be strong evidence of equal goodness; conversely, insensitivity to small improvements is "the mark of incommensurability" (Raz, 1986: 325–326) or symptomatic of "imprecise equality" (Parfit, 2016: 115). Broome (2004: 21) goes so far as to *define* equal goodness to hold whenever we have sensitivity to all improvements/deteriorations. We are thus justified in inferring from the Weak Non-Identity Principle the

> *Strong Non-Identity Principle* Life l_1 is at least as good as life l_2 if and only if for any population X not including the l_1 or l_2 persons, $X + l_1$ is at least as good as $X + l_2$.

(Henceforth just the 'Non-Identity Principle'.)

We shall now see that the Non-Identity Principle implies Anonymity. First note that the Non-Identity Principle requires that if a single person in one population is replaced by a different person at the same wellbeing level, the resulting population will be equally good. We can use this fact to show that Anonymously Equivalent populations must be equally good when they are disjoint (by which I mean that no person belongs to both populations). Given two disjoint, Anonymously Equivalent populations, one population can

be transformed into the other by simply replacing each person by their same-wellbeing counterpart along a bijection witnessing Anonymous Equivalence, one at a time. The Non-Identity Principle implies that each replacement results in an equally good population, and transitivity then yields that the first population is equally as good as the last.

This suffices to show that any two disjoint, Anonymously Equivalent populations must be equally good. Now suppose we have two arbitrary Anonymously Equivalent populations, A and B, which might not be disjoint. We may construct a third population C, which is Anonymously Equivalent to, and disjoint with, both A and B. Since each of A and B are equally as good as C by the previous argument, transitivity requires that A and B are equally as good as each other. The Non-Identity Principle therefore implies Anonymity, given transitivity. In fact, something stronger can be shown: the Non-Identity Principle implies the principle of *Anonymous Pareto*, according to which a population is better than another when the first is Anonymously Equivalent to a population which is weakly pareto superior to the second (i.e. better for some, and at least as good for all). The proof of this claim uses the same sort of idea as the proof of Anonymity. Namely, we replace each person by their better-off or equally well-off counterpart one at a time, and use transitivity to chain the betterness claims together.

On the face of it, it is surprising that the Non-Identity Principle implies Anonymity. The Non-Identity Principle only says that what we may call

"non-identity improvements" – replacements of worse-off people by different better-off people – make the world better. But the Non-Identity Principle does *not*, on the face of it, say that non-identity improvements make just as much evaluative difference as the provision of ordinary comparative benefits. Thus, it is apparently possible to marry the Non-Identity Principle to a "Two-Tier View" (Parfit, 2017), on which non-identity improvements matter, but they matter less than ordinary comparative benefits (evaluatively speaking). Yet the Two-Tier View implies that the Non-Identity Principle is true, and Anonymity is false, so how could the Non-Identity Principle imply Anonymity? The answer is that, as Parfit shows, the Two-Tier View is cyclic (or option-set-dependent), and therefore intransitive (or option-set-dependent). One might nevertheless wonder whether there might be some transitive, option-set-independent population axiology which satisfies the Non-Identity Principle but not Anonymity. The preceding argument confirms that there is no such population axiology.

I believe most readers will find the Non-Identity Principle rather plausible. But perhaps some will find it more palatable to reject the Non-Identity Principle than to accept Anonymity. Can anything be said against this position? Yes. The consequences of rejecting the Non-Identity Principle are more implausible than it might at first appear. At least, this is true if we accept the principle of *Same-Person Anonymity*, according to which Anonymously Equivalent populations are equally good if they contain exactly the same

people. Same-Person Anonymity is the uncontroversial part of Anonymity. It implies no contentious claims about how the importance of benefiting people compares to the importance of replacing people; it merely states that it does not make an evaluative difference who gets what, when everyone exists regardless of which population comes about. It is a very plausible principle, so I shall take it for granted in what follows.[5]

Suppose that the (Weak) Non-Identity Principle is false, so that for some population X, a better life l_1 and a worse life l_2, $X + l_1$ is not better than $X + l_2$. If this is true, then presumably it remains true in at least one case where X contains a life at a wellbeing level w strictly between the l_1 and l_2 levels. Let us assume that A is such a population. Write A' for a population that differs from A only in that one person at w is now at the wellbeing level of l_1, and write l_1' for a life in which the l_1 person has wellbeing level w. By assumption, we have that $A+l_1$ is not better than $A+l_2$. We also have, from Same-Person Anonymity, that $A + l_1$ is equally as good as $A' + l_1'$. Given transitivity, these two claims imply that $A' + l_1'$ is not better than $A + l_2$. That is, we have reached the implausible conclusion that it is sometimes not better overall if existing people are made better off, while at the same time

[5]Philosophers who are partial to partiality can take heart that the following argument can, at the cost of a little extra complication, be reconstructed to make do with weaker principles for trading off wellbeing between people instead of Same-Person Anonymity. In the table representing populations A, B and C on the next page, we merely need to reduce Adam's wellbeing in population C to a level slightly greater than 60. The details are left to the reader.

worse-off people are replaced by different better-off people.

It is easier to see what is going on if we consider a more concrete case. Suppose that, because we reject the Non-Identity Principle, we believe that A is not better than B in the case illustrated by the table below, where Ω represents non-existence:

	Adam	Steve	Eve
A	60	80	Ω
B	60	Ω	40
C	80	60	Ω

Intuitively, C is better than B. Same-Person Anonymity implies that A and C are equally good. Transitivity then implies that A is better than B, in line with the Non-Identity Principle. If we want to reject the Non-Identity Principle in this instance, we must therefore say that C is not better than B, despite appearances. But compare B and C directly. We can imagine going from C to B by making two changes. First, we make Adam worse off by twenty units of wellbeing. Second, we replace Steve with the worse-off Eve. The first change clearly makes the world worse overall.[6] In this particular case, it is implausible to claim that if the second change occurs as well, the

[6]This claim might be false on a suitably extreme egalitarian view, on which decreases in wellbeing are not worse overall when they sufficiently reduce inequality. I think that views like this are not particularly plausible, but it is worth noting that if we accept such a view, it might follow that the first change does not make things worse overall, but on the other hand it would seem that the second change does make things worse overall, because it brings us from an equal to an unequal population. So we would still face a similar argument to the effect that the combined change must be for the worse, since if we combine a bad change with a change that only makes someone worse off, the combined change must make things worse overall.

combined effect of both changes is not to make the world worse overall.[7] To avoid this implausible claim, we should accept the Non-Identity Principle.

The preceding argument may also be applied to cases in which multiple worse lives could be replaced by multiple better lives. If the Non-Identity Principle were false, so that replacements of worse for better lives sometimes fail to make the world better overall, it would seem natural to think that replacements of many worse for many better lives would likewise fail to make the world better overall. But this view has absurd implications. Consider two possible futures. In the first future, the present seven billion inhabitants of Earth will each enjoy 80 units of wellbeing, and there will exist seven billion future individuals, each with 60 units of wellbeing. In the second future, the present inhabitants will instead have 60 units of wellbeing, while seven billion future individuals, who are non-identical to the future individuals in the first future, will have only 40 units of wellbeing. If we suppose that it does not make the world better to replace any number of lives at level 40 for the same number of lives at level 80, an argument exactly analogous to that of the preceding paragraph shows that, given Same-Person Anonymity

[7]There are similarities between my argument here and Broome's (2004; 2005) "greediness" objection to the Intuition of Neutrality. Some readers might worry that criticisms of Broome's argument, such as those levelled by Rabinowicz (2009) or Frick (2017), might apply equally to my own argument. But this is not the case. My argument does not appeal to the general claim that a bad thing plus a neutral thing must be bad, or that a bad thing plus a thing which is not bad must be bad. It instead appeals to the more specific claim that replacing better-off people with different worse-off people cannot swallow up the badness of making existing people worse off. This specific claim is plausible even if the more general claims are false.

and Transitivity, the second future must not be worse than the first. But we should not accept this conclusion. If a change both makes seven billion present people worse off, and also makes seven billion future people worse off than the seven billion who would otherwise have existed, and affects nobody else, it is clear that this change makes things worse overall. This remains clear even when it is stipulated that the change in question would alter the identities of all future people.

Let us take stock. Suppose we accept transitivity and choice-set-independence. We must then accept the Non-Identity Principle, or face what appear to be compelling counterexamples to our view. Since the Non-Identity Principle implies Anonymity, we must accept Anonymity as well. If we do so, this amounts to substantial progress. There are, broadly speaking, two basic questions in population axiology. The first is the question of how to compare populations which differ with respect to the number of people they contain. The second is the question of how to compare populations which differ with respect to the identities of the people they contain. Anonymity provides a full answer to the second question, namely that such differences in the identities of people make no all-things-considered evaluative difference.

Chapter 2

The Welfare Diffusion Objection to Prioritarianism

Abstract

According to the Welfare Diffusion Objection, we should reject Prioritarianism because it implies the "desirability of welfare diffusion": the claim that it can be better for there to be less total wellbeing spread thinly between a larger total number of people, rather than for there to be more total wellbeing, spread more generously between a smaller total number of people. I argue that while Prioritarianism does not directly imply the desirability of welfare diffusion, Prioritarians are nevertheless implicitly committed to certain principles for comparing different-number populations which, together with the Prioritarian same-person axiology, imply the desirability of welfare diffusion.

2.1 Introduction

This chapter is about what I shall call the "Welfare Diffusion Objection" to Prioritarianism. According to Prioritarianism, "benefiting people matters more the worse off these people are" (Parfit, 1997: 213). To be more precise, Prioritarianism of the sort I shall discuss holds that

(i) Benefits to the worse off matter more in an *axiological* sense: they do more to make an outcome better than same-sized benefits to the better-off. That is, I am talking about what Parfit (1997: 213) calls *Telic* rather than *Deontic* Prioritarianism.

(ii) The worse off are those who have less *wellbeing*, rather than (for example) those who have access to fewer resources.

(iii) The sort of wellbeing in question is a person's wellbeing during her entire life, rather than during a part of her life or a moment in time.

(iv) Benefits to the worse off matter more only because these people are at a lower absolute level, and not because they are at a lower level relative to other people. Prioritarianism, as I use the term, thus rules out Egalitarianism.

Prioritarianism of this sort implies that it would be better for ten people to have slightly less than 50 units of wellbeing each – for concreteness, let's

say 49 units each – than it would be for five people to have 100, and the remaining five people to have nothing. This is rather plausible.

It would be much less plausible to say that it would be better for ten people to have 49 units of wellbeing each, rather than for five people to have 100 while *nobody else exists*. In that case, we would be saying that it can be better to increase the number of people and spread a *lesser* total amount of whatever makes life worth living thinly among these people, even if in the smaller alternative population there would be perfect equality and everyone would be individually better off. As Ingmar Persson (2011, 2012) puts it, we would be committed to the "desirability of welfare diffusion". The Welfare Diffusion Objection to Prioritarianism holds that (i) Prioritarians cannot avoid saying that welfare diffusion is desirable, and that (ii) since this is very implausible, we should reject Prioritarianism.

Some philosophers have argued that Prioritarianism implies the desirability of welfare diffusion outright (Persson, 2011, 2012; Holtug, 2010). The most obvious argument for this claim starts from the observation that Prioritarianism can be construed as recommending populations with greater total *priority-weighted* wellbeing, where the priority weighting function is some strictly concave, strictly increasing function of wellbeing which maps the neutral level to zero.[1] Since larger populations with less total wellbeing can

[1]We shall discuss *Critical Level* versions of Prioritarianism, which do not map the neutral level to zero, in §2.3.

nevertheless have greater priority-weighted wellbeing, Prioritarianism of this sort says that such populations are better (Persson, 2011, 2012). Another argument goes as follows: when we bring additional people into existence with lives worth living, these people are thereby benefited. Since Prioritarians care about benefiting people, they should say that bringing such people into existence makes the world better. And if they say that, they will have to say that welfare diffusion can be desirable (as we shall see in §2).

There are problems with both arguments. The first argument takes Prioritarianism to require us to compare different-number populations by their *sums* of priority-weighted wellbeing. But why think that Prioritarians need to compare sums of wellbeing, rather than aggregating in some other way, such as taking the average? Unless a case is made for aggregation by summation, the first argument is incomplete. The second argument assumes that, because creating people with good lives benefits these people, Prioritarians need to say that doing so makes the world better. But this argument seems to assume *Existence Comparativism*: the claim that a good (bad) life can be better (worse) for a person than non-existence. But this claim is very controversial, and so a Prioritarian may respond by simply denying Existence Comparativism.[2] To be fair, the second argument might be construed in another way: it might be argued that, even if creating a person with a good

[2]Many philosophers reject Existence Comparativism. See for instance Narveson (1967), Parfit (1984: 489), Broome (1999: 168), Bykvist (2007) and Bader (2022b).

life does not make that person better off, we can still intelligibly say that the person is thereby *benefited* (Parfit, 1984; McMahan, 2013). But now there is a new problem: why must a Prioritarian say that benefiting people *in this way* makes the world better? They might instead claim, quite reasonably on the face of it, that benefiting people makes the world better only when the people who receive benefits are thereby made better off.

More broadly, both preceding arguments might be rejected because they mistakenly assume that Prioritarianism involves any commitments regarding comparisons of different-number populations in the first place. Prioritarianism might instead be taken to be a theory of same-person comparisons only.[3] Understood this way, Prioritarianism would seem to have nothing at all to say about whether wellbeing should be aggregated by summation or in another way, or about whether so-called "existential benefits" make the world better, or about any other population-ethical matter. As Derek Parfit put it:

> Like the Principles of Personal Good, or Pareto Principles, the Prioritarian Principles that I have considered cannot be applied to cases in which, in the different possible outcomes, different people would exist. When we consider these cases, we need other principles.

[3]By "same-person" comparisons, I mean comparisons involving the same people in each population under consideration. In contrast, "different-number" comparisons involve populations which have different numbers of people.

Parfit (2012: 440)

If Prioritarianism is understood in this way, is the Welfare Diffusion Objection toothless? I shall argue that it is not. Prioritarianism of this sort cannot imply the desirability of welfare diffusion outright. But that does not mean that Prioritarians avoid the Welfare Diffusion Objection entirely. One of the main distinctive features of Prioritarianism is that it is concerned with people's absolute wellbeing levels, rather than their wellbeing levels relative to other people (Parfit, 1997: 214). This concern for absolute levels, I shall argue, means that there are some, fairly weak, population-ethical principles which Prioritarians ought to accept. And it turns out that these principles are enough to bridge the gap from a purely same-person Prioritarian theory to the desirability of welfare diffusion.

Let me be more specific. Given transitivity, Prioritarians cannot avoid the desirability of welfare diffusion if they accept the Mere Addition Principle, which says that additions of good lives do not make the world worse. I demonstrate this in §2.2. The Mere Addition Principle is pretty plausible in its own right – note that it says additions of good lives make the world *not worse*, not that they make the world at least as good – but it is particularly plausible for Prioritarians. I show in §2.3 that Mere Addition is strongly supported by the principle of Separability, which encodes the Prioritarian concern for people's absolute levels. I also show that, given some minor fur-

ther assumptions which most Prioritarians would accept, the Mere Addition Principle can replaced by a weaker principle which only says that creating *very* good lives never makes things worse.

The upshot, I think, is that Prioritarians need to either bite the bullet and learn to live with welfare diffusion, take radical measures like the denial of transitivity, or stop being Prioritarians. §2.4 will be of interest particularly to those who find the third option most palatable (or those who never accepted Prioritarianism in the first place): it demonstrates that, holding fixed transitivity and the Mere Addition Principle mentioned earlier, we can only avoid the desirability of welfare diffusion by accepting Totalism – the view that one population is better than another iff it has greater wellbeing. Or, to be more precise, we need to accept that Totalism applies when comparing populations involving good or neutral lives: the argument from the undesirability of welfare diffusion does not rule out giving priority to those with bad lives. After presenting this argument, in §6.7, I consider and reply to several objections to my earlier claims.

2.2 The Mere Addition Argument for Welfare Diffusion

We begin with a few technical preliminaries. As was implicit earlier, I shall use real numbers to represent wellbeing levels. Positive numbers represent lives worth living, negative numbers represent lives worth not living, and zero represents neutral lives. Greater numbers represent better lives.

There remains the question of how ratios of the differences between wellbeing levels are to be understood. I shall use the following scale: for some fixed good quality of life q, a life is at $x \geq 0$ units of wellbeing if and only if it is equally as good as a life which lasts for x years, and is at some constant good quality q. A life is at $-x$ if and only if one would rationally be indifferent between a 50-50 gamble yielding the $-x$ life and a life at x, or a neutral life for sure.[4]

I assume that there are infinitely many possible people. A *population* is a finite set of people with associated wellbeing levels. Populations represent the distributions in which precisely these people exist, with lives at the respective wellbeing levels, and nobody else exists. $p_i[w_i]$ denotes the population consisting of just person p_i at level w_i; similarly, if X is a set

[4]Readers who doubt that additional years of good life have constant marginal value presumably have some other scale of wellbeing in mind; these readers are invited to consider that scale. I discuss what happens if we get our scale from Expected Utility Theory in §2.5.2.

of possible people, $X[w_i]$ denotes the population containing the X-people at level w_i, and nobody else. Populations are disjoint when they have no persons in common. If X and Y are disjoint populations, we may write $X + Y$ for the population which consists of the X-people at their respective levels, the Y-people at their respective levels, and nobody else.[5] (When I write $X + Y$, I am always assuming that X and Y are disjoint; quantifiers should be understood to be restricted accordingly.)

We shall be interested in the at-least-as-good-as relation, denoted by $\succeq$. I take this to be a transitive binary relation on populations.[6] This is not a completely innocent assumption. Some philosophers, most notably Larry Temkin (1987, 1996, 2012) and Stuart Rachels (1998, 2001, 2004), have argued that the at-least-as-good-as relation is not transitive. Others believe that the at-least-as-good relation is option-set-dependent, and hence cannot be understood to be a binary relation on populations.[7] I'm sceptical of both positions, but I won't argue against them in this chapter. Let's set them aside for the time being.

That is enough in the way of background. Let's see how the core commitment of Prioritarianism, when conjoined with the Mere Addition Principle, implies the desirability of welfare diffusion. To begin with, let's set these

[5]More precisely, $X + Y$ is the set-theoretic union of X and Y.

[6]Recall that a binary relation R is *transitive* iff whenever aRb and bRc, we have aRc.

[7]See Frick (2014, 2022). Cusbert (2017) suggests that Temkin's Essentially Comparative View (2012) can be understood as implying option-set-dependent betterness, rather than intransitivity within choice sets.

principles out more precisely. (Two of these principles unavoidably look a bit complicated when stated precisely, but they are easy to understand by examining Figures 1 and 2.)

We shall understand the core commitment of Prioritarianism to be the principle of

> **Strong Pigou-Dalton** Let $w^- < w$ be any wellbeing levels, and let a be any quantity of additional wellbeing. There is a small positive quantity of wellbeing ϵ' such that for any possible persons p_i and p_j, and any disjoint unaffected background population U, if $0 \leq \epsilon \leq \epsilon'$ then

$$U + p_i[w] + p_j[w^- + a - \epsilon] \succ U + p_i[w + a] + p_j[w^- + \epsilon]$$

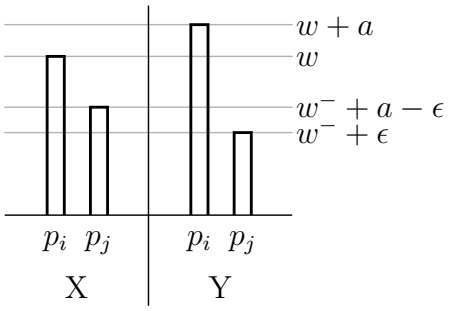


Figure 2.1: *
$X \succ Y$

Figure 2.2: Strong Pigou-Dalton

Strong Pigou-Dalton says that slightly smaller benefits to the worse off

make the world better than slightly larger benefits to the better off. Put another way, transfers of wellbeing from the better-off to the worse-off make the world better even when they are slightly "leaky", resulting in a small loss of total wellbeing. Prioritarians cannot reject this principle.

Next, we have the Mere Addition Principle:

> **Mere Addition** For any populations X and Y, if Y consists only of lives worth living, then $X + Y$ is not worse than X.

As I mentioned earlier, Mere Addition does not imply that an addition of lives worth living is *at least as good*, or *better*, than no addition at all. Principles of that sort are suspect because they conflict with the Evaluative Procreation Asymmetry, according to which bringing lives worth living into existence never makes the world better.[8] Mere Addition, as stated here, faces no such objection.

Finally, we need a principle which guarantees that welfare diffusion is not desirable. This shall be:

> *Different-Number Egalitarian Dominance* Let X and Y be any populations. If
>
> (i) X is a perfectly equal non-empty population of lives worth living;

[8]Proponents of the Procreation Asymmetry include Frick (2014, 2017, 2020) and Roberts (2011). McMahan (2009, 2013) suggests we should accept a weaker version of the Asymmetry. As it happens, I think that we should reject the evaluative version of the Procreation Asymmetry; Broome (2004, 2005) provides the best argument I know of to that effect.

(ii) each person in X is better off than each person in Y;

(iii) each person in X exists in Y (and is therefore better off in X than in Y);

(iv) X has greater total wellbeing than Y,

then X is at least as good as Y.

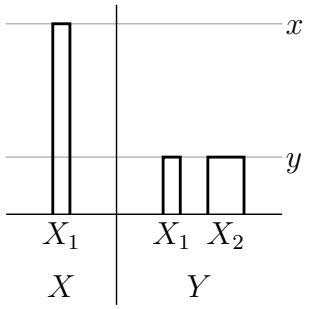


Figure 2.3: *
If $x|X| > y|Y|$, *then* $X \succeq Y$

Figure 2.4: Different-Number Egalitarian Dominance

Different-Number Egalitarian Dominance tells us that smaller populations with greater total wellbeing and perfect equality of wellbeing are at least as good as larger populations with lesser total wellbeing. It is restricted to cases where the people in the smaller population also exist in the larger population. This restriction ensures that everyone who exists in the smaller population is better off, even if Existence Comparativism is false. Different-Number Egalitarian Dominance is therefore compatible with narrow person-

affecting views, on which an outcome can only be better than another if it is better for some particular person.

Different-Number Egalitarian Dominance encodes avoidance of the Welfare Diffusion Objection: population axiologies which do not satisfy Different-Number Egalitarian Dominance say that welfare diffusion is sometimes desirable, or at least say that sometimes, welfare diffusion is not undesirable.[9] So population axiologies which do not satisfy Different-Number Egalitarian Dominance are open to the Welfare Diffusion Objection.

Putting these three principles together, it can be shown that

> **Proposition 1** *No population axiology satisfies Strong Pigou-Dalton, Mere Addition and Different-Number Egalitarian Dominance.*

> *Proof.* Let p_i and p_j be any two possible people. Apply the Strong Pigou-Dalton principle with an empty unaffected background population, higher wellbeing level 50, lower wellbeing level 0, and 50 units of potential additional wellbeing. This tells us that there is some small positive quantity of wellbeing ϵ such that

$$p_i[50] + p_j[50 - \epsilon] \succ p_i[100] + p_j[\epsilon]$$

[9] Different-Number Egalitarian Dominance could be false without smaller populations with greater total wellbeing being *worse* than larger populations with lesser total wellbeing. It would be enough for two such populations to be incomparable. But since it seems to me that the claim that the two populations are incomparable is not much more plausible than the claim that the larger population is better, I shall ignore this distinction going forward.

Different-Number Egalitarian Dominance implies that

$$p_i[100] \succeq p_i[50] + p_j[50 - \epsilon]$$

Transitivity then implies that

$$p_i[100] \succ p_i[100] + p_j[\epsilon]$$

Which contradicts Mere Addition. □

To get this argument through, we applied Different-Number Egalitarian Dominance to the case of comparing a *single-person* population with greater total wellbeing to a larger population with slightly lesser total wellbeing. Different-Number Egalitarian Dominance might seem suspicious in exactly these kinds of cases: it might be reasonable, for example, to think that it would be better for there to be ten billion people, at wellbeing level one hundred, than for there to be one person at level one hundred billion.

However, this objection can be brushed aside, because Proposition 1 still goes through even if we weaken Different-Number Egalitarian Dominance so that it applies only to populations of size n or larger (no matter how large n is). If X_i and X_j are disjoint sets of n possible people each, observe that by applying Strong Pigou-Dalton and transitivity n times, we can show that[10]

[10]Strictly speaking, this claim should be (and can be) proved by induction on n. The proof is routine, and is omitted for brevity.

$$X_i[50] + X_j[50 - \epsilon] \succ X_i[100] + X_j[\epsilon]$$

From here, following the same strategy as in the proof of Proposition 1, it is easy to apply Different-Number Egalitarian Dominance and transitivity in order to show that

$$X_i[100] \succ X_i[100] + X_j[\epsilon]$$

which contradicts Mere Addition.

Since Prioritarians cannot deny Strong Pigou-Dalton, the upshot of Proposition 1 is that they must choose between Mere Addition and Different-Number Egalitarian Dominance.

In the next section, I shall argue that Prioritarians cannot reasonably reject Mere Addition. I shall also argue that they are implicitly committed to the principle of Separability, which says that unaffected people can be ignored when comparing populations. Separability in turn provides strong support for Mere Addition, and even stronger support for a weaker version of Mere Addition which, if we slightly strengthen our Pigou-Dalton principle, implies the desirability of welfare diffusion.

2.3 Mere Addition and Separability

What distinguishes Prioritarianism from Egalitarianism? Parfit (1997: 214) answers: while Egalitarians are concerned with relations between people's wellbeing levels and the wellbeing levels of others, Prioritarians are solely concerned with people's absolute wellbeing levels. Consider a situation in which one person is at 100 and another is at wellbeing level 0, when both could instead have been at wellbeing level 50. Egalitarians decry this situation because it involves inequality: the less well-off person is not as well off as the better-off person. Prioritarians decry the same situation on different grounds: for them, the situation is regrettable not because it avoids inequality, but because although there is nothing bad about the first person being at level 100 rather than level 50, it would have been better for the second person to be at level 50 rather than level 0, even though the two potential benefits are of the same size.

The Prioritarian concern for people's absolute wellbeing levels provides immediate support for the Mere Addition Principle.[11] Consider the application of this principle in the proof of Proposition 1. Mere Addition there

[11]It might be objected that if we take the restriction of Prioritarianism to same-person cases seriously, we cannot say that Prioritarians should accept different-number principles like Mere Addition. But my claim is not that Prioritarianism *itself* supports Mere Addition. My claim is instead that a sole concern for people's absolute wellbeing levels – an important pre-theoretic intuition that underpins Prioritarianism – supports Mere Addition.

implied that $p_i[100] + p_j[\epsilon]$ is not worse than $p_i[100]$. Egalitarians can reject this claim. They can say that, because the existence of p_j introduces inequality, it would be better if only p_i were to exist. But Prioritarians cannot say the same thing. Since they are concerned only with people's absolute levels of wellbeing, they cannot appeal to relations between p_i's wellbeing and p_j's wellbeing when both exist. On the face of it, Prioritarians can say that it is bad for p_j to exist only if existence is bad *for* p_j. But that cannot be the case: while p_j has only a low positive wellbeing level, a low positive level still represents a life worth living, though perhaps only barely.

We can make another, more precise, argument from the sole concern with absolute levels to the Mere Addition Principle. A sole concern for people's absolute wellbeing levels is captured by the principle of

> **Separability** Let X, Y and Z be any populations. X is at least as good as Y if and only if $X + Z$ is at least as good as $Y + Z$.

Separability is widely accepted by Prioritarians. Indeed, Adler and Holtug (2019: 104) take a version of Separability to a defining feature of Prioritarianism. The version of Separability they are talking about is restricted to same-person cases, while mine is unrestricted (and needs to be).[12] But a sole concern for people's absolute wellbeing levels supports the unrestricted version of Separability just as well as it supports the restricted version. If

[12]My thanks to an anonymous reviewer for pointing this out.

unrestricted Separability is false, then the relative contributive values of populations X and Z depend not only on the absolute levels of the persons involved in X and Y, but also on the status of the unaffected people in population Z. Thus, one cannot deny unrestricted Separability without thereby expressing a concern for more than just people's absolute wellbeing levels.

Separability is hard to square with the negation of Mere Addition. If we deny Mere Addition, we think that sometimes it is worse to add lives worth living to the world. If we accept Separability as well, then we will have to infer that it is *always* worse to add such lives to the world. This claim is implausible in its own right, and it can also be shown to be incompatible with a very compelling principle, namely the

> **Absolute Value Principle**[13] If X is a population consisting solely of lives worth living, and Y is a population consisting solely of lives worth not living, then X is better than Y.

At least, this is so if we accept

> **Non-Absolute Priority** For any positive quantity of wellbeing x, there is some sufficiently small positive quantity of wellbeing ϵ such that for any persons p_i and p_j, and any disjoint unaffected background

[13]This principle is sometimes called "Priority for Lives Worth Living" (see for instance Blackorby et al., 2005: 135). I avoid this name because it is suggestive of Prioritarianism, whereas the Absolute Value Principle is satisfied by many non-Prioritarian population axiologies (such as Totalism and the Average view).

population U,

$$U + p_i[\epsilon] + p_j[\epsilon] \not\succ U + p_i[x] + p_j[-\epsilon]$$

Non-Absolute Priority says that we should not give absolute priority to those who are slightly below the neutral level, over those who are slightly above the neutral level. The opposite view, Absolute Prioritarianism, says that those who are below the threshold of a life worth living are to be prioritised absolutely over those who are above the threshold. On this view, it would be better to spare one person from a pinprick which would push them just barely below the neutral level, rather than to spare trillions of people from a greater harm which would not push them below the neutral level. Since most people do not find this kind of view very plausible, I shall not discuss it further.[14]

Given Non-Absolute Priority and Separability, the negation of Mere Addition is inconsistent with the Absolute Value Principle. If Mere Addition is false, there is some case in which an addition of a life at positive wellbeing level x is worse than no addition at all.[15] By Separability, adding a person (let's say p_2) at level x is therefore *always* worse than no adding no one. Now

[14] See Crisp (2003) for a critical discussion of this kind of Absolute Prioritarianism.

[15] Strictly speaking, this does not follow from the negation of Mere Addition, since Mere Addition could be false because some addition of *multiple* lives worth living is worse, while additions of individual lives worth living are always incomparable with no addition at all. In practice this does not matter, because the instance of Mere Addition appealed to in the proof of Proposition 1 concerned an addition of a single life.

consider the following three populations.

A_1 $\quad p_1[-\epsilon]$

B_1 $\quad p_1[-\epsilon] + p_2[x]$

C_1 $\quad p_1[\epsilon] + p_2[\epsilon]$

By the negation of Mere Addition and Separability, B_1 is worse than A_1. By the Absolute Value Principle, A_1 is worse than C_1. Transitivity then implies that B_1 is worse than C_1. This contradicts Non-Absolute Priority.

I find the Absolute Value Principle utterly compelling, so I find this argument for Mere Addition decisive. But not everyone accepts the Absolute Value Principle. Critical Level (or Critical Range) Prioritarians, for example, believe that it can be worse (or not better) for there to be many lives that are positive, but below a "critical level" $x^* > 0$, rather than for there to be fewer lives at a negative wellbeing level.[16]

It turns out, however, that even Critical Level Prioritarians do not avoid the Welfare Diffusion Objection. We can adapt the previous argument to make do with a weaker version of the Absolute Value Principle, which Critical Level Prioritarians would accept. We can then only get a weaker version of Mere Addition out, but it will be enough for our purposes. This weaker Absolute Value Principle says that

[16]See Blackorby et al. (1995, 2005) for a discussion of critical level views.

Weak Absolute Value Principle There is a positive wellbeing level a such that if X is a population consisting solely of lives which are at least at wellbeing level a, and Y is a population consisting solely of bad lives, then X is better than Y.

Critical Level Prioritarians say that a large number of lives barely worth living can be worse than a smaller number of lives worth not living. But they do not say that a large number of *excellent* lives can be worse than a smaller number of negative lives. The former claim is pretty implausible, but one might perhaps reluctantly accept it in order to avoid the Repugnant Conclusion. The latter claim is much more implausible, and cannot be justified on this basis.

We also need a slightly different, but still very plausible, Non-Absolute Priority condition. The condition we shall use is

Non-Absolute Priority 2 For some sufficiently small positive quantity of wellbeing ϵ', if $\mathcal{W}$ is any bounded interval of non-negative wellbeing levels, there is a sufficiently large positive quantity of wellbeing δ' such that given any unaffected background population U, if $\epsilon \leq \epsilon'$, $\delta \geq \delta'$ and w_i, w_j are in $\mathcal{W}$, then

$$U + p_i[w_i - \epsilon] + p_j[w_j + \delta] \succ U + p_i[w_i] + p_j[w_j]$$

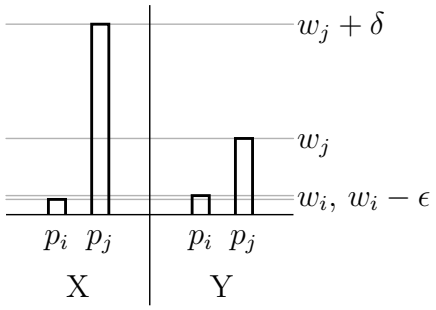


Figure 2.5: *
$X \succ Y$

Figure 2.6: Non-Absolute Priority 2

This principle looks a little complicated, but essentially it just says that it is always better to provide a sufficiently large benefit to a better-off person, rather than a very small benefit to a worse-off person. The required size of the benefit to the better-off person can increase as the gap between the two increases (which is why we quantify over $\mathcal{W}$). Note also that Non-Absolute Priority 2 only requires us to avoid giving absolute priority to people with wellbeing levels which are, at worst, only slightly negative.

Let's now see how these principles imply a weaker version of Mere Addition. Consider the following three populations, where a is a sufficiently high wellbeing level for the Weak Absolute Value Principle to apply, ϵ represents an arbitrarily small quantity of wellbeing, and a^+ is some arbitrarily good wellbeing level:

$A_2 \quad p_i[-\epsilon]$

B_2 $\quad p_i[-\epsilon] + p_j[a^+]$

C_2 $\quad p_i[a] + p_j[a]$

By applying Non-Absolute Priority 2 finitely many times, it can be shown that B_2 is better than C_2.[17] The Weak Absolute Value Principle implies that C_2 is better than A_2. Transitivity then implies that B_2 is better than A_2.[18] Given Separability, this constitutes a proof of

> *Weak Mere Addition* There is some positive wellbeing level a such that for any population X, and any population Y consisting of lives at level a, $X + Y$ is not worse than X.[19]

This weaker version of the Mere Addition Principle is enough to commit the Prioritarian to the desirability of welfare diffusion. To show this, we

[17]Let ϵ be a small enough quantity of wellbeing that Non-Absolute Priority 2 applies, and let a be large enough that the Weak Absolute Value Principle applies. Let n be the smallest number greater than $\frac{a+\epsilon}{\epsilon}$, and let $e = \frac{a+\epsilon}{n}$; we then have $e \leq \epsilon$. We can apply Non-Absolute Priority 2 repeatedly to show that

$$p_i[a] + p_j[a] \prec p_i[a-e] + p_j[a+\delta_1] \prec \cdots \prec p_i[a-ne] + p_j[a+\sum_{i=1}^{n}\delta_i]$$

Writing a^+ to stand for $a + \sum_{i=1}^{n}\delta_i$, the last population is equal to

$$p_i[-\epsilon] + p_j[a^+].$$

[18]We only really need the weaker conclusion that B_2 is not worse than A_2.

[19]In fact, we have proved something stronger: $X+Y$ is better than X. I shall not use this stronger claim, but it's worth noting, for those who may disagree that we are intuitively committed to the *undesirability* of welfare diffusion rather than its mere *non-desirability*, that we could use the stronger claim to establish that Prioritarians are committed to the claim that welfare diffusion can make things *better*, rather than merely not making things worse.

need a same-person Prioritarian principle which is slightly different to Strong Pigou-Dalton. We can call this

> *Priority-Utility Trade-off* For any positive wellbeing level a, there are greater wellbeing levels b and c, with $c > b$, a set N of possible people of size n and a possible person p_i such that
>
> (i) $N[b] + p_i[b] \succ N[c] + p_i[a]$
>
> (ii) $nc > (n+1)b$

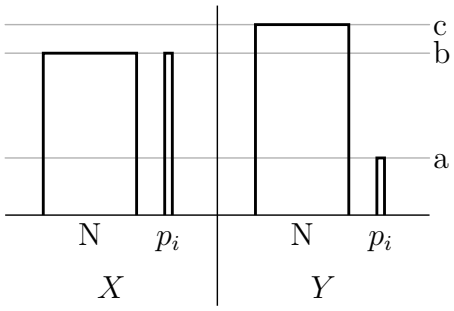


Figure 2.7: $*$
$X \succ Y$

Figure 2.8: Priority-Utility Trade-Off

As the name suggests, Priority-Utility Trade-Off requires that we give priority to the worse off in such a way that we sometimes sacrifice a significant amount of total utility. More precisely, it says that for any level of wellbeing a, we can find a large number of people at a higher level of wellbeing c such that, rather than having the unequal outcome in which one person at level

a and the large number at level c exist, it would be better if instead all of these people existed at an intermediate wellbeing level b, even though this would mean a sacrifice of more than a units of total wellbeing. For example, although wellbeing level 100 is the level of a very good life by contemporary standards, Prioritarians (and Egalitarians) will presumably think that, rather than having one person at level 100 and another person at level 1001, it would be better if instead both were at level 500, even though this would come at the cost of more than 100 units of total wellbeing. As far as I can see, Prioritarians cannot reject this principle.

It can now be shown that

Proposition 2 *No population axiology satisfies Priority-Utility Trade-off, Weak Mere Addition and Different-Number Egalitarian Dominance.*

Proof. Let a be a wellbeing level witnessing Weak Mere Addition (that is, additions at level a are never worse). Priority-Utility Trade-off implies that there are wellbeing levels $b > c$, both of which are greater than a, a set N of possible people of size n and a possible person p_i such that $nc > (n+1)b$ and

$$N[b] + p_i[b] \succ N[c] + p_i[a].$$

Now compare the population $N[c]$ with $N[b] + p_i[b]$. $N[c]$ has total

wellbeing nc, which (from Priority-Utility Trade-Off) is greater than $(n+1)b$, which is the total wellbeing of $N[b]+p_i[b]$. Furthermore, $N[c]$ is a perfectly equal population of good lives, each person in $N[c]$ exists in $N[b]+p_i[b]$, and each person in N is better off in the former population than in the latter. Different-Number Egalitarian Dominance therefore implies that

$$N[c] \succ N[b] + p_i[b].$$

By transitivity, we then have

$$N[c] \succ N[c] + p_i[a],$$

which contradicts Weak Mere Addition. □

We also have:

Corollary *No population axiology satisfies Separability, Non-Absolute Priority 2, the Weak Absolute Value Principle, Priority-Utility Trade-off and Different-Number Egalitarian Dominance*

Since the first four of these principles are satisfied by all plausible versions of Prioritarianism, the upshot is that no plausible version of Prioritarianism avoids the Welfare Diffusion Objection. Note also that even Prioritarians who deny Separability do not necessarily escape the Welfare Diffusion Objection. Proposition 2 does not appeal to Separability directly: I have used

Separability only to support (Weak) Mere Addition. But (Weak) Mere Addition is independently very plausible, and would be hard to deny even for those who do not find Separability particularly compelling.

2.4 A Related Argument for Totalism

The arguments of §2.2 and §2.3 are not only of interest to Prioritarians and their critics. Proposition 1, which we used to establish that Prioritarians cannot accept Mere Addition without leaving themselves open to the Welfare Diffusion Objection, can be repurposed into an argument for Totalism. Recall that Proposition 1 shows that no population axiology satisfies Mere Addition, Different-Number Egalitarian Dominance and Strong Pigou-Dalton. Strong Pigou-Dalton is controversial (albeit still intuitively compelling) because it says that benefits to the worse off matter *more* than benefits to the better off. A weaker Pigou-Dalton principle, which only says that benefits to the worse off matter *at least as much* as benefits to the better off, is accepted by virtually everyone:

> **Weak Pigou-Dalton** Let p_i and p_j be any two possible people, and let U be any disjoint unaffected background population. If w^- is a lower wellbeing level than w, then for any positive quantity of additional

wellbeing a,

$$U + p_i[w^- + a] + p_j[w] \succeq U + p_i[w^-] + p_j[w + a]$$

Assume also a stronger version of Mere Addition, which says that additions of good *or neutral* lives cannot make the world worse. Call this principle Mere Addition*. Finally, consider a slightly stronger version of Different-Number Egalitarian Dominance, which applies to neutral as well as good lives and drops the requirement that those who exist in the population of better-off people must also exist in the population of worse-off people:

> *Different-Number Egalitarian Dominance** Let X and Y be any populations. If
>
> (i) X is a perfectly equal non-empty population of good or neutral lives;
>
> (ii) each person in X is at least as well off as each person in Y;
>
> (iii) X has greater total wellbeing than Y,
>
> then X is at least as good as Y.

We shall now see that these three principles together imply

> *Totalism for Good Populations* Suppose that non-empty populations X and Y contain only lives that are neutral or good. Then X is at least

as good as Y if and only if $T(X)$ is at least as great as $T(Y)$ (where $T(X)$ denotes the total wellbeing of population X).

That is, we have

Proposition 3 *Every population axiology which satisfies Mere Addition*, Weak Pigou-Dalton and Different-Number Egalitarian Dominance* also satisfies Totalism for Good Populations.*[20]

Proof. Given transitivity, it is sufficient to show that every population consisting only of good or neutral lives is equal in value to a singleton population containing one person at the total wellbeing level. Different-Number Egalitarian Dominance* then requires these singleton populations to be ranked according to total wellbeing, and transitivity extends this ranking to all other populations with only good or neutral lives.

Let C be an arbitrary non-empty population of good or neutral lives. Let p_i be some person in C, and let C' be the set of people in C, except for p_i. Define populations A and B to be:

$A \quad p_i[T(C)]$

$B \quad p_i[T(C)] + C'[0]$

[20]Huemer (2012) provides a similar argument. His argument assumes a stronger version of the Mere Addition Principle, which implies that additions of good lives must render an outcome *at least as good*. This Mere Addition Principle is justified by an appeal to Existence Comparativism.

C is obtainable from B by means of a series of pure transfers of well-being from better-off to worse-off, taking wellbeing from p_i each time. Weak Pigou-Dalton therefore implies that C is at least as good as B. Different-Number Egalitarian Dominance* implies that A is at least as good as C. Applying transitivity, we find that A is at least as good as B. But Mere Addition* implies that B is not worse than A. It follows that A and B must be equally good.[21] Recalling that A is at least as good as C and that C is at least as good as B, we can conclude that A and C are equally good too. □

2.5 Objections and Replies

2.5.1 The Repugnant Conclusion

Proposition 3 strikes me as a good argument for Totalism for Good Populations. But is it *sound*? Here's one reason to think not: Totalism for Good Populations implies the Repugnant Conclusion, and many philosophers think the Repugnant Conclusion is false.[22] One might worry that this makes the

[21] We have $A \succeq B$ and $B \not\succ A$. Since (by definition) $B \prec A$ iff $A \succeq B$ and $B \not\succeq A$, we have that $B \succeq A$, hence $A \sim B$.

[22] According to the Repugnant Conclusion, for any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better, even though its members have lives that are barely worth living (Parfit, 1984: 388). Although the Repugnant Conclusion has traditionally been regarded as a decisive counterexample to Totalism and other population axiologies, this traditional view no longer enjoys near-unanimity; see Zuber et al. (2021).

arguments of §2.2 and §2.3 suspect: these arguments assume (or near enough assume) principles which jointly entail the Repugnant Conclusion.[23]

I do not think that such thoughts should be of much comfort to Prioritarians. If the Repugnant Conclusion is false, then so is at least one of the premises of Proposition 3. But can Prioritarians reasonably believe that one of these premises is false? I think not.

Consider first Mere Addition. This principle, while independently plausible, is made more plausible by the Prioritarian focus on people's absolute wellbeing levels. If the best way to avoid the Repugnant Conclusion is to reject Mere Addition, Prioritarians are in a tough spot: unlike others, such as Egalitarians, they have no natural explanation for how Mere Addition could fail.

Next consider Weak Pigou-Dalton. On the face of it, this principle is integral to Prioritarianism: surely if you believe that same-sized benefits to the worse-off matter more, you must believe that they make the world at least as good as same-sized benefits to the better-off.

Maybe that's a little too quick, though. Let's imagine for the sake of argument that Prioritarians can deny the letter of Pigou-Dalton principles while maintaining their spirit. Suppose, for instance, that a Prioritarian

[23]Different-Number Egalitarian Dominance is a premise of Proposition 1. Mere Addition* is almost identical to Mere Addition, which is a premise of Proposition 1. Weak Pigou-Dalton is closely related to Strong Pigou-Dalton, which is the last premise of Proposition 1.

can adopt something like Parfit's (2004) Perfectionism, and say that pure transfers of wellbeing from better-off to worse-off people can make the world worse when they involve the loss of the best things in life, but not when they do not. This view may help Prioritarians avoid the Repugnant Conclusion, but it will not help them avoid the Welfare Diffusion Objection. The problem is that if we can apply Pigou-Dalton whenever the best things in life are *not* lost, we can still construct cases where Different-Number Egalitarian Dominance will be violated. For example, by applying an argument of the same form as the proof of Proposition 1, Prioritarians could likely be pushed to accept that a population consisting of ten billion people, each at level 100, would not be better than a population of eleven billion people, each at level 90.[24] The latter population could retain the best things in life, since 90 years of good quality life leaves plenty of space for perfectionist goods. But the claim that the latter population is not better than the former still violates Different-Number Egalitarian Dominance, and still leaves Prioritarians open to the Welfare Diffusion Objection.

The third possibility is that the Repugnant Conclusion is to be avoided by denying Different-Number Egalitarian Dominance*. But this idea is, on face

[24]Begin with a population A of ten billion at 100. Add one billion at level 1, and raise the A-people to level 101, resulting in population B. Finally, let C consist of all eleven billion people at 90. The Prioritarian will presumably judge that C is better than B (or, if not, the numbers may be adjusted as necessary). Mere Addition implies that B is not worse than A. By applying transitivity, we find that C is not worse than A, contradicting Different-Number Egalitarian Dominance.

of it, wrong-headed. The point of Different-Number Egalitarian Dominance* is to say that smaller populations are better than larger ones when each person in the smaller population is better off than each person in the larger population, provided that the smaller population *also* has greater total well-being. Those who wish to avoid the Repugnant Conclusion will presumably accept this claim, along with the stronger claim that the smaller population is often better even if it has *less* total wellbeing. At any rate, denying Different-Number Egalitarian Dominance is a way of conceding the Welfare-Diffusion Objection, not avoiding it.

2.5.2 The Cardinalisation Objection

I have argued that Prioritarianism is open to the Welfare Diffusion Objection. But does the Welfare Diffusion Objection really have any intuitive bite? That depends on the way we determine the scale of wellbeing. I have generally worked with a scale based on years of good life, with one unit of wellbeing corresponding to one year of good quality life. But we could generate the scale of wellbeing in another way: by appealing to utility functions generated by Expected Utility Theory.[25] If our wellbeing scale is generated in this way, it is

[25]I am assuming that the axioms of Expected Utility Theory are satisfied for the prudential betterness relation on prospects; if they are not, it is not possible to represent the betterness relation as maximising the expectation of a real-valued utility function. See Morgenstern and Von Neumann (1944), Savage (1954) or Fishburn (1982), among many others, for proof of the equivalence between a relation on gambles satisfying the axioms of Expected Utility Theory and the representability of this relation by a utility function.

at best unclear that we can have sensible intuitions regarding total amounts of wellbeing.[26] On the face of it, it's reasonable to think that since the utility functions which generate our wellbeing scale are mere representation devices, spat out by a complex mathematical theorem, we lack any intuitive grasp of that wellbeing scale. But if our intuitions regarding total quantities of wellbeing are baseless in this way, then it seems the Welfare Diffusion Objection is nothing to worry about, since it appeals to precisely these sorts of baseless intuitions. Call this the "Cardinalisation Objection".

I do not think that the Cardinalisation Objection is of much help to Prioritarians. The reason is that it seems to me that the Objection works just as well against Prioritarianism itself: if we lack sensible intuitions concerning the moral importance of quantities of wellbeing, then we have no reason to accept that same-sized benefits to the worse-off matter more. The Cardinalisation Objection can therefore only render the Welfare Diffusion Objection impotent by rendering it unnecessary.

Another way of responding to the Cardinalisation Objection is to simply brush it aside. Whatever the merits of cardinalising wellbeing via Expected Utility Theory, we can still talk about other wellbeing scales for which our intuitions regarding quantities of wellbeing are not baseless, such as the "years of good life" scale. At least on the face of it, Prioritarians are committed to Strong Pigou-Dalton and Different-Number Egalitarian Dominance on the

[26] This line of thought is suggested by Greaves (2015).

years-of-good-life scale. So the Welfare Diffusion Objection seems to be a live problem for Prioritarianism on the years-of-good-life scale, even if it is not a problem on the scale generated by Expected Utility Theory. And that's fine: one problem is enough.

2.5.3 Quarantine

Yet another response to the Welfare Diffusion Objection is to attempt to "quarantine" the problem, along with all other difficulties associated with different-number cases. Parfit (2004) summarises this strategy with an analogy:

> It's very difficult to formulate acceptable welfarist theories that could apply to cases that involve infinite quantities of such things as suffering and happiness. That's a worry, but it doesn't undermine our confidence in the theories that can handle cases with only finite quantities.
>
> Parfit (2004: 257)

Similarly, one might think that although it is difficult to formulate acceptable theories of different-number comparisons, this should not undermine our confidence in theories like Prioritarianism, which only apply to same-person comparisons.

It seems to me that there is a more promising and a less promising interpretation of the quarantine strategy. On the more promising interpretation, when *all* theories applicable to domain D face severe difficulties when extended to the larger domain D', in some cases this should not decrease our confidence in the theories applicable to D. On the less promising interpretation, if theory T applicable to domain D faces some particular difficulty whenever it is extended to domain D', this should not make us sceptical of T, even if some other theory T' applicable to D can be extended to D' without facing a similar difficulty.

Since some non-Prioritarian views (for example, Totalism) *can* avoid the Welfare Diffusion Objection when extended to different-number cases, only the less promising interpretation of the quarantine strategy could help to defend Prioritarianism. Yet Parfit's analogy does not help to make the quarantine strategy seem plausible on this interpretation. Consider two theories, T and T', applicable to the evaluation of populations involving finite quantities of suffering and happiness. If some extension of T can deal with populations involving infinite quantities of suffering and happiness in an acceptable way, and no extension of T' deals with such populations in an acceptable way, this *does* seem to provide strong support for T over T'.

In the present case, the fact that Prioritarianism cannot be extended to different-number cases without incurring the Welfare Diffusion Objection is a strike against it. Granted, population axiology is notoriously difficult, all

theories of different-number cases have their intuitive difficulties, and these difficulties need to be balanced against each other in the final analysis. But the Welfare Diffusion Objection still weighs against Prioritarianism, even if other objections weigh against other theories.

2.6 Concluding Remarks

I have argued that the Welfare Diffusion Objection poses a significant challenge to Prioritarianism, even if Prioritarianism is to be thought of as a theory of same-person comparisons only. Importantly, I have assumed throughout this chapter that the at-least-as-good-as relation is both option-set-independent and transitive. If Prioritarians want to avoid the Welfare Diffusion Objection, their best bet may be to do so by challenging one or the other of these assumptions. They might say that in order to restrict Prioritarianism to same-person choices only, transitivity needs to be similarly restricted so that it does not allow one to chain together betterness claims which come from Prioritarian comparisons of same-person choices and non-Prioritarian comparisons of different-number choices.[27] Or, they might say that the Prioritarian weighting of benefits applies only to gains and losses which would render people better or worse off than they otherwise would have been. Given Existence Comparativism, the resulting version of Prior-

[27]Thanks to an anonymous reviewer for pointing out this possible response.

itarianism would either have to be intransitive (if losses and gains depend on the two outcomes being evaluatively compared) or option-set-dependent (if losses and gains depend on the set of relevant alternatives).[28] Taking any such path would be a significant departure from traditional ways of thinking about value, and would open the Prioritarian to other objections, like the prospect of susceptibility to value pumps.[29] Perhaps such objections can be answered, or perhaps they cannot.[30] Another possibility in the vicinity is to abandon axiological or "telic" Prioritarianism, but retain a deontic version of Prioritarianism. Because the "ought-to-bring-about-rather-than" relation is less obviously transitive than the "better-than" relation, intransitive deontic Prioritarianism would seem to be easier to defend than its intransitive telic counterpart.

There are several options for those who are not prepared to give up transitivity or option-set-independent betterness. One is to simply bite the bullet and accept the desirability of welfare diffusion. While this position seems to me unattractive, it might fairly be said that every transitive population axiology takes one unattractive position or another. The desirability of welfare

[28]Otsuka (2022: 538) suggests that Parfit would have endorsed a similar view.

[29]A particularly compelling money/value pump for cyclic theories has recently been provided by Gustafsson and Rabinowicz (2020). Gustafsson (nd) further claims that intransitive, acyclic theories are also vulnerable to value pumps, although this argument is less secure than the value pump argument against cyclicity. Value pumps would, on the face of it, appear to be effective against option-set-dependent theories which involved cycles of betterness among pairwise choices.

[30]Perhaps by defending a decision theory involving resolute choice (McClennen, 1985) or some other unorthodox decision theory (Ahmed, 2017).

diffusion is not clearly more implausible than other controversial positions in population axiology, such as acceptance of the Repugnant Conclusion.[31] It also may be that adopting a version of Prioritarianism which implies the desirability of welfare diffusion has payoffs elsewhere. For instance, unlike Totalism, Total Prioritarianism has the plausible implication that it would be worse to create a number of people at wellbeing level $-x$ and the same number of people at x than it would be to create nobody at all (Holtug, 2010: 255).[32] Total Prioritarianism also implies, rather plausibly, that it can be better for there to be more total negative wellbeing spread thinly among a larger number of people than for there to be less total negative wellbeing spread thickly among a smaller number of people (Holtug, 2010: p256–257).

How bad would it be for a Prioritarian to accept the desirability of welfare diffusion? A precise answer cannot be given without the details of how much priority is to be given to the worse-off, but let me give a rough answer anyway. I expect that many Prioritarians will believe that it is better to bring two people from level 1 to level 20 than it is to bring one person from level 20 to level 100. If that is true, and Mere Addition is also true, then it would not be worse for there to be thirty billion people at level 20 than it would be for there to be ten billion people at level 100. I find that difficult to believe.

[31]However, the "Super-Repugnant Conclusion" (Holtug, 2010: ch. 9), which might be difficult to avoid for a Prioritarian who accepts the desirability of welfare diffusion, is more implausible than the Repugnant Conclusion.

[32]By "Total Prioritarianism", I mean the view that ranks populations according to their total priority-weighted wellbeing.

Another option is to deny the Mere Addition principle. As I argued extensively in §3, this option should not be taken by a Prioritarian. But one could abandon Prioritarianism for this reason, and instead accept Egalitarianism. This seems to me a reasonable response to the arguments of this chapter: it is not crazy to claim that it can be worse to add people with lives worth living when (and because) doing so would introduce significant inequality. That said, it's worth mentioning that Proposition 2 shows that rejecting Mere Addition alone is not enough to avoid the Welfare Diffusion Objection: one would also need to reject Weak Mere Addition, which may be difficult even for an Egalitarian.

A final option is to accept all of the premises of Proposition 3, taking the desirability of welfare diffusion to rule out Prioritarianism, and denial of Mere Addition to rule out Egalitarianism. One would then be left with Totalism for Good Populations. The obvious next step is to accept unrestricted Totalism, but one is not actually forced to this position. One can, compatibly with the premises of Proposition 3, give priority to the worse off whenever the worse off are below the neutral level at the outset. There is more to be said for this restricted version of Prioritarianism than it might at first seem. Roger Crisp (2003: 755), noting the apparent absurdity of prioritising the rich over the super-rich, claims it is plausible that "when people reach a certain level, even if they are worse off than others, benefiting them does not, in itself, matter more". If he is right, then there is some threshold after

which considerations of priority no longer apply. Plausibly, such a threshold should be non-arbitrary. If so, what better candidate could there be than the neutral level of wellbeing?

Chapter 3

In Favour of Making Happy People

Abstract

According to the Evaluative Principle of Neutrality, there is a "neutral range" of wellbeing levels, corresponding to lives worth living, such that if all else is equal, creating people at wellbeing levels within this range never makes an outcome better. In this chapter, I argue against the Evaluative Principle of Neutrality by showing it to be incompatible with two highly plausible principles for comparing different-number populations – principles that are more plausible than the Evaluative Principle of Neutrality itself. I then show that these arguments can be made to work even if we allow that betterness may be option-set-dependent. Finally, I argue, more tentatively, that if the Evaluative Principle of Neutrality is false, we have moral reason to create happy people, and are morally obligated to do so when all else is equal. The practical upshot of these claims is not that we are morally obligated to have children, but that we may have stronger reasons to prevent human extinction than most of us are inclined to believe.

3.1 Introduction

Almost all of us are in favour of making people happy. But most of us are not in favour of making happy people – we are neutral about that.[1] If our neutrality about creating happy people is appropriate, this suggests three principles. First, creating happy people does not make the world better overall. Second, we have no moral reason to create happy people. Third, we are not morally obligated to create happy people. The first principle, stated more precisely, is the

[1] Narveson 1973: 80

> *Evaluative Principle of Neutrality* There is a range of good wellbeing levels (call it the "neutral range") such that if all else is equal, creating people at wellbeing levels within the neutral range never makes an outcome better.[2]

By wellbeing levels, I mean *lifetime* wellbeing levels, which correspond to how well an entire life goes, all things considered. A wellbeing level is "good" if and only if a life at that level of wellbeing is "worth living", or "good for" the person living it, all things considered.[3]

I shall assume that if the Evaluative Principle of Neutrality is true, the neutral range associated with it is fairly wide. To be more precise about what I mean, we shall need a way of denoting wellbeing levels by numbers. Throughout this chapter, I shall use the following scale: wellbeing level n corresponds to a life of some constant good momentary wellbeing level — perhaps the average momentary wellbeing level of good lives in developed

[2]This principle is a variation on Broome's (2004; 2005) *Intuition of Neutrality*. The main difference is that the neutral range is required to contain lives worth living. The upshot of this difference is that if there is a range of wellbeing levels which are neither personally good nor personally bad, and it makes an outcome neither better nor worse to add lives at these wellbeing levels, but it *does* make an outcome better to add lives which are personally good, then the Principle of Neutrality will be true, but the Evaluative Principle of Neutrality will be false. See Gustafsson 2020 for a view like this.

[3]I am appealing here to Parfit's notion of a life (or an outcome) being non-comparatively "good for" a person, which is supposed to be compatible with the claim that this life (or outcome) is not better for her than non-existence (1984: Appendix G). If you don't think that this kind of talk makes sense, you might instead replace talk of "good wellbeing levels" with talk of wellbeing levels at or above some particular threshold, where it intuitively seems permissible (but perhaps not obligatory) to create lives at the threshold quality if all else is equal. The same goes, changing what needs to be changed, for bad lives.

countries today — and which lasts for n years.[4] (Most of the arguments of this chapter won't depend on our having this particular wellbeing scale in mind, though.) I shall assume that the "neutral range" associated with the Evaluative Principle of Neutrality includes all good wellbeing levels up to level 60. That is, I am interested in versions of the Evaluative Principle of Neutrality which are strong enough to say that it would not be overall better to create the sorts of high-quality lives which can today be expected for those born into fortunate circumstances.

Many people who accept the Evaluative Principle of Neutrality also accept a stronger claim, namely the

> *Strong Evaluative Principle of Neutrality* If all else is equal, creating people with good lives, no matter how good their lives are, never makes an outcome better.

(The Strong Evaluative Principle of Neutrality is one half of the popular Evaluative Procreation Asymmetry.)[5] The reason I discuss the weaker Evaluative Principle of Neutrality, rather than the Strong version, is that the Strong version is less intuitively compelling. We are inclined to think that it would not be better for there to exist additional good lives of the sort we

[4]We need not assume that the "n years of good life" notation covers all possible good wellbeing levels. Relatedly, using this notation does not require us to assume that the at-least-as-good-as relation on lives is complete.

[5]The Procreation Asymmetry is discussed by McMahan (1981, 2009, 2013), Frick (2014, 2017), Roberts (2011), and many others.

are familiar with: lives of fairly high quality which, if all goes well, may last eighty years or so. Our intuitions are less clear when it comes to additions of *amazing* lives which involve continual bliss, creative excellence and valuable relationships, and which last for at least ten thousand years. While it might turn out that every theoretical rationale for the weaker claim also supports the stronger claim, I shall not assume that this is the case.

The structure of the chapter is as follows. In §3.2, I provide arguments against the Evaluative Principle of Neutrality which assume that the betterness relation is transitive across option sets. In §3.3, I extend these arguments to cover the possibility that betterness is option-set-dependent, so that transitivity applies only within option sets. In §3.4, I consider whether we can infer from the negation of the Evaluative Principle of Neutrality that we sometimes have moral reasons or obligations to create happy people. I tentatively argue that we can, or that at any rate the arguments of §3.3 can be adapted to support these claims. I also argue that if we do have moral reasons or obligations to create happy people, it (surprisingly) does not obviously follow that we are often obligated to have children. The best-supported revisionary implication is instead this: our moral reasons to prevent human extinction are significantly stronger than most of us are antecedently inclined to believe.

3.2 Neutrality For Orthodox Population Axiology

Throughout this section, I shall make two standard assumptions about betterness. First, I assume that there is a binary at-least-as-good-as relation on *populations*, which can be recovered from the more fundamental at-least-as-good-as relation on outcomes. (By a population, I just mean a finite set of people with associated wellbeing levels.) One population is at least as good as a second if and only if an outcome which instantiates the first population is at least as good as an outcome which instantiates the second population, if all else is equal. Our first assumption is, effectively, that this definition does indeed give us a well-defined relation. Since populations do not contain information about the alternatives available in the outcomes they instantiate, I am effectively assuming that the at-least-as-good-as relation is independent of the set of alternatives which could otherwise be chosen. This assumption, while intuitively attractive, is somewhat contentious, so I shall show how we can do without it in §3.3.

My second assumption is that the at-least-as-good-as relation on populations is transitive. Given our first assumption, transitivity then applies *across*, as well as *within*, option sets. In particular, we have

Transitivity Across Binary Option Sets For any populations X, Y and

Z, if X is at least as good as Y when these are the only two options, and Y is at least as good as Z when these are the only two options, X must be at least as good as Z when these are the only two options.

Armed with Transitivity Across Binary Option Sets, we can deploy an adaptation of John Broome's (2004; 2005) "greediness" argument against the Evaluative Principle of Neutrality. Consider populations P, Q and R illustrated by the table below, where Ω represents non-existence:

	One hundred people	One hundred different people
P	40	Ω
Q	60	40
R	40	60

Since R is obtained from P by adding lives within the neutral range, the Evaluative Principle of Neutrality implies that R is not better than P. Q and R are plausibly equally good, given that we are impartial between the one hundred people who exist in P, and the one hundred different people. Transitivity then implies that P is not worse than Q.[6] But note that Q is obtained from P by adding additional people with good lives, while at the same time making existing people better off. So, coupled with a principle of impartiality between people, the Evaluative Principle of Neutrality implies what we can call the "greediness phenomenon": additions of good lives can be

[6]Suppose P were worse than Q. Since Q and R are equally good, it would then follow that P is worse than R, contradicting the Evaluative Principle of Neutrality.

"greedy" in the sense that they cancel out improvements elsewhere.[7] Because the greediness phenomenon is intuitively unacceptable, we should reject the Evaluative Principle of Neutrality – or so goes the greediness argument.

The final step of this argument is controversial. Some philosophers grant that the Evaluative Principle of Neutrality gives rise to the greediness phenomenon, but deny that we should therefore reject it. Frick (2017) thinks that the greediness argument merely demonstrates that the neutrality of creating happy people *is* greedy. Rabinowicz (2009) thinks that the greediness argument establishes that additions of good lives must be able to count for or against other values, but does not establish that additions of good lives must sometimes make an outcome better.

The greediness argument, as it stands, thus leaves us at an impasse. Some philosophers find the greediness phenomenon intuitively unacceptable, and others do not. However, this impasse can be broken. Even if we grant, for the sake of argument, that the greediness phenomenon is not intrinsically very counter-intuitive, we can argue that it forces us to violate principles for comparing different-number populations which are more compelling than the Evaluative Principle of Neutrality itself.[8] One such principle is

[7]Broome's original examples focus on a different greediness phenomenon: the tendency of additions of good lives to swallow up *badness*. That said, Broome is aware that his Intuition of Neutrality implies that both badness and goodness can be swallowed up, and considers both to be ways in which neutrality is "greedy" (Broome, 2004: 170). The Evaluative Principle of Neutrality, as formulated here, only implies that goodness can be swallowed up.

[8]Broome (2004: 202–206) also supplies a few practical arguments against greedy neu-

Different-Number Egalitarian Dominance Let X and Y be any populations. If

(i) X is a perfectly equal non-empty population of good lives;

(ii) each person in X is better off than each person in Y;

(iii) each person in Y exists in X (and is therefore worse off in Y than in X);[9]

(iv) X has greater total wellbeing than Y,

then X is better than Y.

The judgements delivered by Different-Number Egalitarian Dominance are unimpeachable on utilitarian grounds (due to ii and iv), on egalitarian grounds (due to i), and on person-affecting grounds (due to iii).[10] It is hard to find principles for comparing different-number populations that are acceptable to

trality, which appeal to judgements about particular cases (especially cases involving global warming). However, there is a difference between appealing to intuitive judgements about particular cases, and appealing to compelling general principles. Broome gives one argument, which he credits to Erik Carlson, which seems to tacitly appeal to a version of the Absolute Value Principle we shall see later (2004: 205). But this argument only applies against the Strong Evaluative Principle of Neutrality, and it assumes Utilitarianism for fixed-population comparisons.

[9]Notice that this condition is slightly different from condition (iii) of the version of Different-Number Egalitarian Dominance discussed in Chapter 2; both, however, serve the same function: they ensure that the claim that X is better than Y is compatible with the narrow person-affecting principle.

[10]There are some proposed population axiologies on which Different-Number Egalitarian Dominance is false, such as Holtug's (2010) version of Total Prioritarianism, as well as some versions of critical level or critical range utilitarianism (Blackorby and Donaldson, 1984; Blackorby et al., 1995, 2005). Total Prioritarianism does not validate the Evaluative Principle of Neutrality, but critical range and critical level utilitarianism can (if the critical range or level extends into the good wellbeing levels).

all, but Different-Number Egalitarian Dominance comes pretty close.[11] However plausible we find the Evaluative Principle of Neutrality, it seems clear enough that Different-Number Egalitarian Dominance is more plausible still.

To see the incompatibility between the Evaluative Principle of Neutrality (via the greediness phenomenon) and Different-Number Egalitarian Dominance, consider the populations P, Q and R illustrated by the table below.

	One hundred people	Ten billion different people
P	40	Ω
Q	40.01	40.01
R	40	60

Different-Number Egalitarian Dominance implies that Q is better than P. R is better than Q (as I shall soon argue). Transitivity Across Binary Option Sets then implies that R is better than P, contradicting the Evaluative Principle of Neutrality.

Why is R better than Q? Because even if equality or priority matter, it is not plausible that small gains with respect to equality or priority outweigh very large differences in total wellbeing in same-person cases.[12] On our chosen scale of wellbeing, the numbers correspond to years of good life. So a choice between Q and R is a choice between an outcome in which ten billion people

[11] Some might doubt Different-Number Egalitarian Dominance on the grounds that wellbeing, on the scale I have defined, has diminishing marginal value. But those who believe this presumably have a different scale of wellbeing in mind, and should therefore accept the version of Different-Number Egalitarian Dominance where sums are calculated relative to that scale.

[12] See also Parfit 1997: 205 and Crisp 2003: 752.

get nearly twenty extra years of good life, and an outcome in which one hundred different people get about three or four extra days of good life. I don't have much in the way of further argument here, but I do have some forceful language: the former outcome is *obviously* better than the latter.[13]

The Evaluative Principle of Neutrality is therefore inconsistent with Different-Number Egalitarian Dominance on any axiology which satisfies Transitivity Across Binary Option Sets and some very minimal judgements in same-person cases. Since Different-Number Egalitarian Dominance is more compelling than the Evaluative Principle of Neutrality, we should reject the latter in favour of the former.

One might deny that Different-Number Egalitarian Dominance is significantly more compelling than the Evaluative Principle of Neutrality. I think that it is, but at any rate, the Evaluative Principle of Neutrality turns out to also conflict with another principle which is weaker and even more compelling than Different-Number Egalitarian Dominance. This is the

[13]Note that the argument from Different-Number Egalitarian Dominance would work even if the potential benefit for the one hundred people were arbitrarily small – perhaps a few seconds of extra good life. The claim that R is better than Q would not follow on an extreme egalitarian view, on which more equal populations are always better, provided they do not involve levelling down. But Barrett (2020a,b) has argued persuasively that such views are either cyclic, or they cannot avoid the Levelling Down Objection after all. It is also false that R is better than Q on Rawls's Difference Principle (1999). But it is in precisely these sorts of extreme cases that the Difference Principle seems most implausible.

Absolute Value Principle If X is a perfectly equal population consisting solely of good lives, and Y is a population consisting solely of bad lives, then X is better than Y.[14]

The Absolute Value Principle is about as plausible as you can get when it comes to principles for comparing different-number principles: it just says that it would be better for there to be *only* good lives than for there to be *only* bad lives. In particular, the Absolute Value Principle seems to me much more compelling than the Evaluative Principle of Neutrality.

Consider now the populations P, Q and R illustrated by the table below:[15]

	One hundred people	Ten billion different people
P	-0.01	Ω
Q	0.01	0.01
R	-0.01	60

The argument here has exactly the same form as the argument from Different-Number Egalitarian Dominance. Q is better than P by the Absolute Value Principle. R is better than Q, because the small gain to the one hundred

[14]This principle is often called "Priority for Lives Worth Living" in the economics literature (see Blackorby et al., 2005: 135). I avoid using this name because it is misleadingly suggestive of prioritarianism. The Absolute Value Principle is sometimes confused with the negation of Gustaf Arrhenius's "Sadistic Conclusion" (2000: 251) and with his "Non-Sadism" condition (nd). The difference between these non-sadism conditions and the Absolute Value Principle is that the non-sadism conditions involve comparisons of populations which differ in that either good or bad lives are added, without any restrictions on the unaffected part of the population, whereas the Absolute Value Principle involves comparisons of populations which contain only good lives, or contain only bad lives. (To see the difference, it might help to note that Average Utilitarianism violates Arrhenius's Non-Sadism conditions, but satisfies the Absolute Value Principle.)

[15]Negative wellbeing levels may be defined as in Chapter 2.

people in Q is less important than the large gain to the ten billion people in R.[16] Transitivity then implies that R is better than P, contradicting the Evaluative Principle of Neutrality.

The same-person claim that R is better than Q is slightly more controversial than the same-person claim appearing in the argument from Different-Number Egalitarian Dominance, because the move from R to Q lifts one hundred people up from a bad wellbeing level to a good one. But despite this, the same-person claim remains very secure. The size of the benefit to the one hundred people is sufficiently small that even if a very healthy degree of priority should be given to these people in virtue of the fact that they are slightly badly off, it would still be better for the ten billion people to be given much larger benefits.[17]

If we accept the claim that R is better than Q, together with Transitivity Across Binary Option Sets, we must choose between the Principle of Neutrality and the Absolute Value Principle. In such a contest, the Absolute Value Principle should win out.[18] So we should reject the Evaluative Prin-

[16]Gustafsson (2020) has suggested that there is a range of wellbeing levels which are neither good nor bad nor neutral, and that there is no neutral level. If he is right, then it may be that there is a fairly large gap between any good level and any bad level. This would make the judgement that R is better than Q less plausible, but I think it would still be compelling.

[17]There may be some who are persuaded by the view that outcomes in which some have bad lives must be worse than outcomes in which all have good lives. For instance, Kolodny (forthcoming) seems to think that this view is plausible. I disagree, but in any case, the argument from Different-Number Egalitarian Dominance is unaffected.

[18]The Absolute Value Principle is false on some versions of critical range and critical level utilitarianism, and has also recently been denied by Bader (2022a,b). But as far as I can see, these views are untenable precisely because they involve denying the Absolute

ciple of Neutrality if we are happy with the structural assumptions involved in orthodox population axiology.

3.3 Neutrality and Option-Set-Dependent Betterness

We have so far assumed that (i) there is a binary at-least-as-good-as relation on populations, and that (ii) this relation is transitive. These assumptions jointly imply Transitivity Across Binary Option Sets. However, (i) in particular might seem to assume important points at issue. Recall that, according to (i), the at-least-as-good-as relation must be option-set-independent in the sense that whether one outcome is better than another cannot depend on the sets of available alternatives associated with the two outcomes. So far, we have simply assumed that this is true. But it is *not* obviously true in the context of variable population cases, because it is plausible that in such cases, betterness can be option-set-dependent.[19] We might, for instance, accept

> *Comparative Harm Aversion* Suppose that outcomes X and Y have the same anonymous distribution of wellbeing, that some person in X is worse off than she is in an available alternative, and that no person in Y is worse off than she is in an available alternative. Then Y is

Value Principle.

[19]Frick (2022) argues for this point at length.

better than X.[20]

Comparative Harm Aversion is not *clearly* false, and it is incompatible with (i). Suppose X contains Person 1 (only) at wellbeing level 10, while Y contains Person 2 (only) at wellbeing level 10. How do these two outcomes compare? Although we have fully specified the populations instantiated by each outcome, we cannot say. Suppose X and Y have the same set of relevant alternatives, $\{X, Y, Z\}$. If Z contains Person 1 (only) at level 20, then Y will be better than X. But if Z contains Person 2 (only) at level 20, then X will be better than Y. So if Comparative Harm Aversion is true, we cannot derive a unique at-least-as-good-as relation on populations from the at-least-as-good-as relation on outcomes, because the latter is option-set-dependent. "All else equal" comparisons of goodness depend on *how* all else is equal: on which alternatives are available.

It might be replied that there is a natural and non-arbitrary way to make all-else-equal comparisons of populations in light of potential option-set-dependent betterness, which is to always compare populations with respect to the option set containing those two alternatives, and no others. Call this the Pairwise Interpretation. While the Pairwise Interpretation gives us a way of accepting (i), it gives us no grounds to accept (ii), the transitivity of the resulting relation. Suppose that population X is better than Y relative to $\{X, Y\}$, and Y is better than Z relative to $\{Y, Z\}$. Since the outcomes

[20] A deontic version of this claim is endorsed by Otsuka (2018).

involved must be different each time, the transitivity of the at-least-as-good-as relation on outcomes does not imply that X is better than Z relative to $\{X, Z\}$.[21]

It might thus be our assumption of the conjunction of (i) and (ii) in §3.2 was too quick. The Evaluative Principle of Neutrality lends itself naturally to option-set-dependent betterness because it is suggestive of the claim that facts about which people can potentially be created by the agent, and which people exist regardless of the agent's actions, could be evaluatively significant. By assuming (i) and (ii), which are jointly at odds with option-set-dependent betterness, we might have assumed important points at issue. While (i) and (ii) are theoretically attractive, it might seem that we need to reject at least one of them in order to have an intuitively satisfactory population axiology.[22]

I shall therefore leave open that betterness might be option-set-dependent. Consequently, we shall no longer be able to take Transitivity Across Binary Option Sets for granted. However, I shall continue to assume

> *Outcome Transitivity* Let X, Y and Z be any outcomes. If X is at least as good as Y, and Y is at least as good as Z, then X is at least as good as Z.[23]

Provided that it is possible to make all-else-equal comparisons of populations

[21] Voorhoeve (2013) and Broome (1991: 100-104) make similar points.

[22] See especially Frick (2022), who argues that abandoning option-set-independence can help us solve the Mere Addition Paradox.

[23] Some philosophers reject this assumption, such as Temkin (1987, 1996) and Rachels (1998, 2001, 2004). But I shall maintain it nonetheless.

with associated sets of relevant alternatives, Outcome Transitivity implies

> *Transitivity Within Option Sets* For any populations X, Y and Z, and any option set $\mathcal{O}$, if X is at least as good as Y, relative to $\mathcal{O}$, and Y is at least as good as Z, relative to $\mathcal{O}$, then X is at least as good as Z, relative to $\mathcal{O}$.[24]

There is a perfectly general method of transforming arguments which are valid in a transitive option-set-independent framework (as in §3.2) into arguments which are valid in any framework satisfying Transitivity Within Option Sets. We merely need to replace each option-set-independent principle or judgement involved in the initial argument with a principle which has it that the relevant comparison holds relative to all option sets. This method works because if a given set of principles together imply intransitivity in an option-set-independent framework, the transformed versions of these principles will imply the same instance of intransitivity in the option set consisting of all populations involved in the original argument.

Does this mean that it does not matter whether or not our framework is option-set-dependent? No, because the transformed versions of the option-

[24]Frick (2022) could be interpreted as suggesting that the betterness relation itself is a three-place relation whose relata are two populations (or outcomes) and a set of alternatives, rather than a two-place relation on outcome pairs, as assumed here. If this is the way that option-set-dependent betterness is to be understood, then Outcome Transitivity, as stated here, is not well-formed. But Transitivity Within Option Sets is, and this is the only transitivity assumption we shall need for the remainder of the chapter.

set-independent principles may be less plausible than the originals.[25] We can determine whether or not this is true only by examining the transformed principles directly. I shall focus on the transformed version of the argument from the Absolute Value Principle, but most of what I shall say also goes for the argument from Different-Number Egalitarian Dominance.

The argument from the Absolute Value Principle appealed to the populations P, Q and R in the table reprinted below.

	One hundred people	Ten billion different people
P	-0.01	Ω
Q	0.01	0.01
R	-0.01	60

In §3.2 I claimed that Q is better than P and R is better than Q; therefore, by transitivity, R is better than P, contradicting the Evaluative Principle of Neutrality. In order for the transformed version of this argument to work in an option-set-dependent setting, these claims need to be true in all option sets (or at least in $\{P, Q, R\}$).

According to the Generalised Absolute Value Principle (the version applicable to all option sets), Q is better than P in {P,Q, R}. We might object to this claim if we are inclined to give great weight to minimising comparative harms, such that we believe it is a very bad thing for the ten billion different

[25] Frick (2022) makes this point in his discussion of the Mere Addition Paradox. He argues that a version of the Mere Addition Principle is very plausible when there are only two alternatives, but can be rejected when the additional people are unjustifiably worse off than they might have been.

people to exist in Q. On this view, Q might be worse than P in $\{P,\mathrm{Q}, R\}$ because it is bad for the additional people to exist in Q, given that they could have been better off in R. But comparative harm-minimisation of this sort is perverse. While the ten billion different people do suffer comparative losses in Q, it is not better that these comparative losses be avoided by preventing these people from coming into existence at all.[26] Rather than promoting the wellbeing of each person (subject to the equal consideration of others), these sorts of harm-minimisation views instead lead us to ensure that those who might have been better off are not around to complain about it. More generally, if we say that P is not worse than Q due to the additional option of R, we are saying that the existence of people with bad lives is not worse than the existence of people with good lives, just because some of those good lives might have been better. But to say this is to fail to respond appropriately to the interests of the affected people.

A more general case can be made for the Generalised Absolute Value Principle. We can appeal to the following argument from the absolute values of populations:

(P1) If X is a population of good lives, then X is (non-comparatively) good, relative to any option set $\mathcal{O}$.

(P2) If Y is a population of bad lives, then Y is (non-comparatively) bad,

[26] Ross (2015: 443-446) calls this the "Problem of Improvable Life Avoidance".

relative to any option set $\mathcal{O}$.[27]

(P3) For any populations X and Y, and any option set $\mathcal{O}$, if X is good in $\mathcal{O}$ and Y is bad in $\mathcal{O}$, X is better than Y in $\mathcal{O}$.

(C1) So if X is a population of good lives, and Y is a population of bad lives, X is better than Y, relative to any option set.

(P1) can be defended as follows: outcomes instantiating populations consisting solely of good lives are outcomes which are good for everyone, and outcomes that are good for everyone are good. Theories which deny this last claim exhibit an implausible disconnect between what is valuable for people and what is valuable about outcomes.[28]

I thus think that we should accept (P1). The same obviously goes also for (P2), changing what needs to be changed. Finally, (P3) seems platitudinous.[29] So I think that we should take the above argument to be sound, and accept its conclusion, which is the Generalised Absolute Value Principle.

What about the same-person claim that R is better than Q? Assuming that this comparison holds in the option set $\{Q, R\}$, R could only fail to be better than Q in the three-option case if, due to the presence of P, the wellbeing of the ten billion different people does not matter at all. But this

[27]For brevity, I shall drop "non-comparatively" when discussing the absolute values of populations from now on.

[28]This claim should be understood to be restricted to outcomes which involve nothing of value other than people and their wellbeing.

[29]See also Nebel's (2018) defence of the "goodness/badness principles", which are similar to (P3).

claim is not at all plausible. How could large losses to such a large number of people not matter at all? Imagine that we really faced a choice between P, Q and R, and chose to bring about Q. We could not successfully justify our choice to bring about Q rather than R to the ten billion people by saying that, since the existence of the ten billion people was not settled at the time of the decision, their wellbeing did not matter. We could point to the interests of the one hundred people who are better off in Q. But since there is so little at stake for these people, it is not reasonable to claim that the minor interests of these one hundred people could outweigh the much larger interests of the ten billion people in R being brought about rather than Q.

On examination then, I think we should believe that R is better than Q in $\{P, Q, R\}$, and that Q is better than P in $\{P, Q, R\}$. Transitivity Within Option Sets then implies that R is better than P in $\{P, Q, R\}$. This violates the

> *Generalised Evaluative Principle of Neutrality* If Y is a population of lives in the neutral range, and X is any population which does not include the Y-people, then for any option set $\mathcal{O}$, $X + Y$ is not better than X, relative to $\mathcal{O}$.

This sort of evaluative neutrality, then, has to go. But we might think that the Generalised Evaluative Principle of Neutrality is stronger than is warranted by our intuitions regarding the evaluative significance of creating

happy people. We might think that our intuitions only support neutrality in two-option cases, yielding the

> *Pairwise Evaluative Principle of Neutrality* If Y is a population of lives in the neutral range and X is any population which does not include the Y-people, $X + Y$ is not better than X, relative to $\{X, X + Y\}$.

The Pairwise Evaluative Principle of Neutrality is consistent with the premises of the preceding argument. But we can argue against it in another way. Define a population to be *equivalently good* if and only if it is at least as good as some population of good lives, relative to every option set. Now consider the

> *Expanded Absolute Value Principle* If X is an equivalently good population and Y is a population consisting solely of bad lives, X is better than Y, relative to $\{X, Y\}$.

Suppose we accept this principle. Returning to the populations from the Absolute Value Argument, if we believe that R is better than Q in every choice set (as I previously argued), R must be an equivalently good population. Meanwhile, P is a population consisting solely of bad lives. The Expanded Absolute Value Principle therefore implies that R is better than P in $\{P, R\}$, violating the Pairwise Evaluative Principle of Neutrality.

Why accept the Expanded Absolute Value Principle? We can appeal again to an argument from the absolute values of populations, which recycles premises (P2) and (P3):

(P2) If Y is a population of bad lives, then Y is bad, relative to any option set $\mathcal{O}$.

(P3) For any populations X and Y, and any option set $\mathcal{O}$, if X is good in $\mathcal{O}$ and Y is bad in $\mathcal{O}$, X is better than Y in $\mathcal{O}$.

(P4) If X is an equivalently good population, X is good relative to any option set $\mathcal{O}$.

(C2) So if X is equivalently good, and Y is a population of bad lives, X is better than Y, relative to $\{X, Y\}$.

The question is whether we should accept the new premise (P4). I think on balance we should, but as far as I know, no entirely decisive argument can be made for it. We can get part of the way there by assuming an option-set-dependent version of Nebel's (2018: 878) "Goodness Principle":

> *Goodness Principle* For any outcomes X and Y and any option set $\mathcal{O}$, if X is good relative to $\mathcal{O}$ and Y is at least as good as X relative to $\mathcal{O}$, Y must be good, relative to $\mathcal{O}$.

The Goodness Principle implies that if X is equivalently good in virtue of being at least as good as some population of good lives Y, X must be

good relative to any option set *which includes* Y. This is why the Goodness Principle does not get us all the way to (P4): it does not tell us that an equivalently good population must be good relative to *all* option sets.

However, for our purposes we do not need (P4) in its full generality. We can make do with the more limited claim that population R in particular is good relative to any option set. This claim is plausible. The only bad thing to be said for R is that it contains one hundred people who are slightly badly-off; the good thing to be said for R is that it contains ten billion high-quality lives. Since the good thing to be said for R is very good, while the bad thing to be said for R is only slightly bad, R does seem to be non-comparatively good all things considered; moreover, this seems to be true relative to all option sets (including $\{P, R\}$). If so, given that P is bad in $\{P, R\}$ and that the Goodness Principle is true, the Pairwise Evaluative Principle of Neutrality must be false.

Let me add one important caveat. The Evaluative Principle of Neutrality, as I have stated it, says that additions of good lives in a certain range *never* make the world better, regardless of the population we start with. If this principle is false, it does not follow that additions of good lives *always* make the world better, regardless of the population we start with. Consider Average Utilitarianism. This view validates the Absolute Value Principle and Different-Number Egalitarian Dominance, but does not imply that additions of good lives *always* make the world better, just that they *sometimes* do. So

my arguments do not establish that additions of good lives *always* make an outcome better, just that they *sometimes* do. I believe we should accept the further claim that additions of good lives *always* make an outcome better, but I shall not argue for this claim here. At any rate, this qualification may make little difference in practice. The arguments I have given in this section support the claim that additions of good lives make an outcome better when the initial population is non-comparatively bad. And it seems at least an open possibility that the population of lives existing up to the present day *is* non-comparatively bad, given the long history of suffering on earth (both of historical human beings and of historical wild animals). So even if additions of good lives do not *always* make an outcome better, there is some reason to think that in the situation we in fact face, additions of good lives do make the world better, at least compared to the outcome in which there are no future lives at all.

3.4 The Normative and Deontic Principles of Neutrality

We have so far confined our discussion of neutrality to the evaluative case. I mentioned in the introduction two other Principles of Neutrality. These are the

> *Normative Principle of Neutrality* If all else is equal, we never have pro tanto requiring moral reason to create lives within the neutral range.

As well as the

> *Deontic Principle of Neutrality* If all else is equal, it is never the case that we ought to create lives within the neutral range.

Both of these principles are arguably more compelling than the evaluative Principle of Neutrality. They are also, perhaps, better expressions of the basic intuition most of us share regarding the morality of having happy children, namely that "people should have them if they want them" (Narveson, 1973: 70), and that when people who do not want them do not have them, they do nothing that is in any way wrong or objectionable. Now in fact, as I shall explain later, the Principles of Neutrality are less related to the morality of having children than it would be natural to assume. But as a first pass, the Normative and Deontic Principles of Neutrality do seem more tightly bound up with our ordinary moral intuitions than the Evaluative Principle of Neutrality.

A natural question thus arises: if we reject the Evaluative Principle of Neutrality, must we also reject the Normative and Deontic Principles of Neutrality? It might seem that we must. Consider first the Normative Principle

of Neutrality. The negation of this principle follows from the negation of the Evaluative Principle of Neutrality, together with the

> *Goodness-Reasons Bridge Principle* If outcome X is better than outcome Y, we have pro tanto requiring moral reason to bring about X rather than Y.

The Goodness-Reasons Bridge Principle is, at the very least, intuitively plausible. Let me clarify it in a way that defuses two potential worries.

First, it is important that the moral reasons posited by the Goodness-Reasons Bridge Principle are only pro tanto. The Goodness-Reasons Bridge Principle does not imply Consequentialism. It is, for example, compatible with the view that in Thomson's (1976) well-known footbridge case, one has all things considered moral reason not to kill one to save five. The Goodness-Reasons Bridge Principle only requires that there is at least some pro tanto moral reason to save the five, which is true.

Second, it might be argued that the Goodness-Reasons Bridge Principle requires us to be value-fetishists, in that it demands that we have moral reasons to bring about better outcomes because they are better. But this is not true. The Goodness-Reasons Bridge Principle only makes an extensional claim about when we have pro tanto moral reasons. It says nothing about the grounds of these reasons. We can perfectly well accept the Goodness-Reasons Bridge Principle while also accepting (for instance) Scanlon's plausible view

that "being good, or valuable, is not a property that itself provides a reason to respond to a thing in certain ways. Rather, to be good or valuable is to have other properties that constitute such reasons" (1998: 97). In particular, we might believe that we have reasons to create happy people not because doing so would lead to overall better outcomes, but because doing so would provide these people with good lives.

Let us then assume the Goodness-Reasons Bridge Principle for the time being, though we shall come back later to the possibility of rejecting it. The negation of the Evaluative Principle of Neutrality then implies the negation of the Normative Principle of Neutrality. Another bridge principle suffices to take us to the negation of the Deontic Principle of Neutrality. This is the

> *Reasons-Obligations Bridge Principle* If we have pro tanto requiring moral reason to bring about X rather than Y, and no other pro tanto reasons of any sort, then we ought to bring about X rather than Y.

We should accept the Reasons-Obligations Bridge Principle because requiring moral reasons, when unopposed, are decisive. Since requiring moral reasons are the sorts of things capable of delivering obligations, they will do exactly that unless there are other reasons there to stop them.

The Reasons-Obligations Bridge Principle suffices to take us from the negation of the Normative Principle of Neutrality to the negation of the Deontic Principle of neutrality. However, there is a snag. The Reasons-

Obligations Bridge Principle is compelling precisely because it involves *requiring* moral reasons. On a simple picture of how moral reasons work, these are all the moral reasons that there are. But on more complicated pictures, there are non-requiring reasons which are incapable of generating moral obligations but which may have other effects, such as generating permissions or rendering an agent's action praiseworthy.[30] If our reasons really do look like this, it might seem that we were too quick in rejecting the Normative Principle of Neutrality. Our argument assumed that if X is better than Y, then there is pro tanto *requiring* moral reason to bring about X rather than Y. But we might think that only a weaker principle is warranted, namely the

> *Goodness-Reasons Bridge Principle** If outcome X is better than outcome Y, we have pro tanto moral reason *of some sort* to bring about X rather than Y.

This weakened bridge principle would only be enough to force us to reject the

> *Normative Principle of Neutrality** If all else is equal, we never have pro tanto moral reason *of any sort* to create lives within the neutral range.

But there is no compelling argument from the negation of the Normative Principle of Neutrality* to the negation of the Deontic Principle of Neutrality.

[30] See for instance Gert (2003) or Portmore (2021).

If we only have non-requiring pro tanto moral reason to bring about X rather than Y, then (platitudinously) we are not required to bring about X rather than Y.

It seems, then, that much hangs on whether we accept the stronger Goodness-Reasons Bridge Principle, or whether we instead accept only the weaker Goodness-Reasons Bridge Principle*. If we accept only the weaker principle, we will believe that there must be *some* moral reasons to create happy people – but it is an open possibility that such reasons are non-requiring (most likely, they are merely justifying).[31]

Is it plausible to accept the weaker but not the stronger Goodness-Reasons Bridge Principle? That is, is it plausible that while we have moral reasons to bring about better outcomes, such reasons might be non-requiring? I find it hard to judge. When we can bring about a better outcome by providing ordinary benefits to people who already exist, it seems plausible that we have requiring moral reasons to do so: when we are permitted not to benefit people, this is generally explained (assuming such cases arise) by the existence of agent-centred prerogatives or other morally significant factors.[32] This seems to at least partially support the view that there are requiring reasons to create happy people: it seems ad hoc to claim that while we have requiring

[31]McMahan (2013: 20–23) and Thomas (2019: 15) suggest, on independent grounds, that advocates of the asymmetry might accept the existence of justifying or "cancelling" moral reasons to create happy people.

[32]See for instance Scheffler (1994). I am imagining that prerogatives count as reasons, but some philosophers dispute this; see for instance Muñoz (2021: 702).

moral reasons to produce better outcomes by benefiting existing people, we do *not* have requiring moral reasons to produce better outcomes by creating additional happy people. But this combination of claims is not ad hoc to the point of unbelievability. Making existing people better off and creating additional happy people are, after all, very different things, and it is not crazy to think that we have different sorts of reasons in either case. To be satisfactory, a proponent of the Normative Principle of Neutrality would of course need to provide an explanation for *why* we have different sorts of reasons in the two sorts of cases, but this task seems like it could be manageable.

As it stands, then, there seems to be no clearly decisive reason to accept the stronger Goodness-Reasons Bridge Principle over the weaker Goodness-Reasons Bridge Principle*. Yet it still seems to me that there are powerful reasons to reject the Normative Principle of Neutrality. Even if we cannot argue from the negation of the Evaluative Principle of Neutrality, we can still marshal the same arguments that led us to reject the Evaluative Principle of Neutrality directly against the Normative Principle of Neutrality. By allowing that value might be option-set-dependent, we have removed an important structural disanalogy between the normative and the evaluative case. This makes translations of our evaluative arguments into normative terms more plausible. More precisely, we can re-run the argument from the Absolute Value Principle found in §3.3, replacing all instances of "at-least-as-good-

as" with "at-least-as-much-reason-to-choose". Recall the three populations involved in that argument, reprinted below:

	One hundred people	Ten billion different people
P	-0.01	Ω
Q	0.01	0.01
R	-0.01	60

What are our reasons like in the option set $\{P, Q, R\}$? Plausibly, the following two claims are true:

(i) There is more (and requiring) moral reason to choose Q than P.

(ii) There is more (and requiring) moral reason to choose R than Q.

These two claims imply that there is more moral reason to choose R than P, assuming the principle of

> *Normative Transitivity Within Option Sets* For any populations X, Y and Z, and any option set $\mathcal{O}$, if there is at least as much reason to bring about X rather than Y, relative to $\mathcal{O}$, and there is at least as much reason to bring about Y rather than Z, relative to $\mathcal{O}$, then there is at least as much reason to bring about X rather than Z, relative to $\mathcal{O}$.

This normative principle of transitivity is plausible. (Note that it says nothing about transitivity *across* option sets.) If we accept it, we thus seem to

have a violation of the version of the Normative Principle of Neutrality which applies across all option sets. Similar adaptations of other arguments in §3.3 can be used to put pressure on the version of the Normative Principle of Neutrality which applies only to pairwise option sets: provided we always have more moral reason to bring about good outcomes than bad ones, it is plausible that we will end up with a violation of the Pairwise Normative Principle of Neutrality.

These arguments against the Normative and Deontic Principles of Neutrality seem to me somewhat less secure than the arguments against the Evaluative Principle of Neutrality developed in §3.2 and §3.3. But they are not without force. Unless we are prepared to deny Normative Transitivity Within Option Sets (which is admittedly less secure than the transitivity of the at-least-as-good-as relation), it seems to me that there is significant pressure to reject the Normative and Deontic Principles of Neutrality.

Suppose we do reject these principles. One might worry that we will then be forced to conclude that we ordinarily have moral obligations to have children, provided our potential children would have lives worth living. I believe that this worry is misguided.

First, it is plausible that we have strong personal prerogatives as to whether or not to have children. These prerogatives are intuitively strong enough to explain, for instance, why we would not be morally obligated to have children even if we learned that doing so will somehow make a neighbour

very happy (for innocent reasons, let us suppose), raising her from the level of a life barely worth living to that of a life well worth living. So they seem strong enough to explain why we would be permitted not to have children, even if we had a pro tanto moral reason in favour of doing so which is about as strong as our pro tanto moral reasons to greatly benefit existing people in a way that does not involve saving them from great harms. My point is not that these common-sense moral judgements are correct, it is that the existence of a strong pro tanto moral reason to have children could not make it obligatory to have children unless we have already jettisoned important parts of the common-sense morality of ordinary procreation.

Second (and more importantly), it is unclear whether having children is a way of increasing the expected number of happy people in the first place. While having a child certainly makes it the case that there is *one* individual who would not have existed otherwise, it does not do only that. The child will go on to live a life of their own, be a part of the global economy, and affect the future in unpredictable ways. It could be, for instance, that having an additional child will add to overpopulation pressures, that people will respond to these overpopulation pressures by having fewer children themselves, and that the net effect will be a *decrease* in population size. I am not saying that we know that this will happen. I am saying that we are radically uncertain about the future (especially the further future), so much so that we have almost *no idea* whether having children would increase or

decrease the population size in the long run.[33] What we do know is that having children will have morally significant effects – and large ones – on the future. This is true regardless of whether or not we accept the various Principles of Neutrality.

There is, however, one case in which we can be fairly sure that our actions will increase or decrease the expected (human) population size. This is the case in which we can decrease or increase the risk of near-term human extinction. We can, fairly predictably, decrease the expected size of the human population by increasing extinction risk, and reduce the expected size by decreasing extinction risk. Denying the various Principles of Neutrality thus suggests something like

> *Ex Ante Anti-Extinction* It would be ex ante better to/we would have moral reason to/we have a moral obligation, all else equal, to: reduce the risk of near-term human extinction, provided that future human lives would otherwise be good.

Ex Ante Anti-Extinction is intuitively plausible. But one might think that the way in which we endorse Ex Ante Anti-Extinction, if we reject the Principles of Neutrality, is problematic. Frick (2017) has pointed out that if our reasons to prevent extinction are just reasons to create happy people, we should be indifferent between an increase in the population size at a particu-

[33] See Greaves 2016.

lar time and an increase in the population size across time (by extending the lifetime of humanity as a whole). But most of us care more about the latter than about the former.

I do not think that these considerations reveal any major problem with rejecting the Principles of Neutrality. Assuming our intuitions about these cases are not misguided (though I am sceptical of this), the proper conclusion seems to be that we should care about the future of humanity over and above the extent to which we care about increasing the number of happy people. But there is no contradiction in caring about the future of humanity in this way and *also* caring about increasing the number of happy people. It is plausible that here we have two sorts of reasons, and it is an open question how they are to be weighed up.

There is, however, one way in which it is of practical importance whether we come to Ex Ante Anti-Extinction via negated Principles of Neutrality, or via some other way. If we accept Ex Ante Anti-Extinction because we believe there are reasons to increase the (expected) number of happy people, then it matters whether, as some have argued, there are an enormously large number of people in humanity's future, conditional on our near-term survival (Greaves and MacAskill, 2021). If there are, then a reasonable first pass has it that if we have reasons to create happy people, then our reasons to bring about the existence of *enormously many* happy people by preventing extinction must be overwhelmingly strong: so strong that they will override

practically all of our reasons associated with near-term considerations.

This is a sensible thing to think, but it is perhaps a little too quick. It is not immediately obvious that our moral reasons must be aggregative in the sense that the strengths of our reasons to provide benefits or avert harms grow proportionally to the number of people susceptible to these benefits or harms. Many philosophers hold a different view, on which sometimes our reasons *don't* aggregate in this straightforward way: paradigmatically, our reasons to save a single person from death are stronger than our reasons to save any number of people from suffering mild headaches.[34] The question is: on a less-than-fully aggregative view like this, how strong are our reasons to *slightly* reduce the risk of *enormously many* people being prevented from coming into existence by near-term human extinction, assuming the Normative Principle of Neutrality is false? If the Normative Principle of Neutrality *is* false, as I have argued, this question is of great practical importance. But regrettably, I cannot answer it here.

3.5 Conclusion

Many of us are inclined to accept the Evaluative Principle of Neutrality. However, the arguments of §3.2 show that the Evaluative Principle of Neutrality is inconsistent with compelling principles for comparing different-number

[34]See Scanlon (1998: 235), Voorhoeve (2014, 2017), Frick (2015), Lazar (2018) and Rüger (2020), among many others.

populations which are more compelling than the Evaluative Principle of Neutrality itself. In §3.3, I showed how to modify these arguments so that they apply even if we allow that betterness may be option-set-dependent. In light of these arguments, we should reject the Evaluative Principle of Neutrality.

In §3.4, I showed that we can apply various plausible bridge principles to argue from the negation of the Evaluative Principle of Neutrality to the negations of the Normative and Deontic Principles of Neutrality. It is unclear whether such arguments are successful. We can also modify the arguments of §3.3 so that they apply directly to our moral reasons. These arguments are more powerful, though perhaps not completely decisive. So while there is significant pressure to reject the Normative and Deontic Principles of Neutrality, it is not clear whether we ought to do so, all things considered. If we do reject these principles, the main practical upshot is not that we might be morally obligated to have children. The main practical upshot is that preventing human extinction may be much more morally important than most of us are antecedently inclined to believe.

Chapter 4

Repugnance Without Mere Addition

Abstract

This chapter presents the Additive Repugnance Theorem, an impossibility theorem regarding the difficulty of avoiding the Repugnant Conclusion. It differs from other results primarily in that it replaces conditions such as the "Mere Addition Principle", or various "Non-Sadism" conditions, with the significantly more plausible "Minimal Absolute Value Principle", on which populations which contain only very bad lives are worse than populations which contain only good lives. It also replaces the popular assumption of transitivity with the logically weaker and more compelling assumption of acyclicity. I argue that the Additive Repugnance Theorem shows that the Repugnant Conclusion cannot reasonably be avoided by population-ethical means alone; it is to be avoided, if at all, by modifying our commitments in fixed population cases.

4.1 Introduction

Most people are inclined to deny Derek Parfit's

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

However, successfully doing so has turned out to be extraordinarily difficult. Along with the statement of the Repugnant Conclusion, Parfit provided his Mere Addition Paradox: a compelling argument for the Repugnant Conclusion. Since its publication, philosophers and economists have contributed yet stronger arguments and impossibility theorems, revealing the problem of avoiding the Repugnant Conclusion to be even harder than Parfit first thought.[1] In this chapter, I provide another argument in this vein: the Additive Repugnance Theorem.

The most controversial premise of Parfit's Mere Addition Paradox is (arguably) the Mere Addition Principle, according to which additions of good lives to the world cannot result in a worse outcome. The main point of the Additive Repugnance Theorem is to do without the Mere Addition Principle,

[1]See Arrhenius 2000, 2003, 2009, 2011, nd, Ng 1989, Carlson 1998, Blackorby et al. 2003, 2005, Nebel 2019 and Spears and Budolfson 2021.

and to do without anything which looks like the Mere Addition Principle. Instead, the Additive Repugnance Theorem appeals to the *Minimal Absolute Value Principle* (a weakening of Blackorby et al.'s (2003: 351–2) "Priority for Lives Worth Living"), which says that there is a population of very bad lives which is worse than any population consisting solely of good lives. I present the Additive Repugnance Theorem in §4.3.

The Minimal Absolute Value Principle is closely related to the "Weak Non-Sadism" condition, which is the replacement for Mere Addition used in Gustaf Arrhenius's favoured Sixth Impossibility Theorem (Arrhenius, 2009, 2011, nd). The Weak Non-Sadism condition is an additive version of the Minimal Absolute Value Principle, which says that some population of very bad lives makes a worse addition to any unaffected background population than any population of good lives. Arrhenius's Sixth Theorem uses this additive version of the Minimal Absolute Value Principle to derive a non-additive version of the Repugnant Conclusion.[2] The Additive Repugnance Theorem works the other way round: it uses the non-additive Minimal Absolute Value Principle in order to derive an additive version of the Repugnant Conclusion.[3] This switch in direction has two important payoffs. First, as I shall

[2]Strictly speaking, Arrhenius's avoidance condition for the Repugnant Conclusion is additive, but in a weaker sense: the size of the added population of excellent lives can depend on the choice of unaffected background population.

[3]Spears and Budolfson (2021) have also provided an impossibility theorem which, like the Additive Repugnance Theorem, derives an additive version of the Repugnant Conclusion from a principle like the Minimal Absolute Value Principle (and from other premises). However, Spears and Budolfson make a number of assumptions which are controversial

argue in §4.2, Weak Non-Sadism is not very compelling on the assumption that the Mere Addition Principle is false, and is therefore not much more compelling than the Mere Addition Principle itself. In contrast, the Minimal Absolute Value Principle is extremely compelling, regardless of whether the Mere Addition Principle is true.

Second, while the Weak Non-Sadism condition is very plausible, it is not obviously *more* plausible than the negation of the (Very) Repugnant Conclusion.[4] It is therefore unclear, in light of the Sixth Theorem, whether we need to reject Non-Sadism or accept the Repugnant Conclusion, if it comes down to a choice between the two. In contrast, the negation of the Minimal Absolute Value Principle is clearly more implausible than the additive version of the Repugnant Conclusion; yet the additive version of the Repugnant Conclusion is still "repugnant" in the same sort of way as the non-additive version.[5] The

among philosophers (though admittedly less so among economists), including the completeness of the at-least-as-good-as relation on populations, the transitivity of the same relation, choice-set-independence, anonymity, and the representability of wellbeing levels by real numbers. The Additive Repugnance Theorem makes none of these assumptions (although in the cases of transitivity and choice-set-independent betterness, it makes only slightly weaker assumptions).

Incompleteness in the context of population ethics is defended by many authors, including Blackorby et al. 1996, Qizilbash 2007a,b, 2018, Rabinowicz 2009, Chang 2016, Parfit 2016, Gustafsson 2020 and Nebel 2022. Anonymity is less controversial, though some authors reject it, including Roberts 2011 and Temkin 1987, 2012. Intransitivity is defended by Temkin 1987, 1996, 2012 and Rachels 1998, 2001, 2004. Choice-set-dependence is defended by Frick 2014, 2022 and, on some readings, by Temkin 2012 (according to Cusbert 2017).

[4]The Very Repugnant Conclusion differs from the Repugnant Conclusion in that it compares a population of excellent lives to an extremely large number of lives barely worth living, *plus* a smaller number of very bad lives. See Arrhenius 2003, 2009, 2011.

[5]Consider a future for humanity in which many people live excellent lives, and contrast a drab future in which an enormously large number of people live lives barely worth living.

Additive Repugnance Theorem therefore closes a question which Arrhenius's Sixth Theorem arguably leaves open: is the least implausible population axiology one which avoids the Repugnant Conclusion and all its variants by rejecting compelling principles for comparing different-number populations, such as the Mere Addition Principle and Weak Non-Sadism? The Additive Repugnance Theorem shows that the answer to this question is: No.

The Additive Repugnance Theorem improves on the Sixth Impossibility Theorem in two other minor respects. First, rather than assuming the transitivity of the at-least-as-good-as relation, the Additive Repugnance Theorem instead assumes the logically weaker and more compelling assumption of acyclicity.[6] Second, the Additive Repugnance Theorem allows for the possibility of choice-set-dependent betterness: whether one population is better than another might depend on the set of available alternatives.[7] The validity of the argument is preserved, despite this structural weakening, by using premises which apply with respect to all choice-sets. It is of course not surprising that this can be done; still, there is a minor payoff in that it makes

It is "repugnant" for the drab future to be better. Suppose now we are informed that regardless of which future we choose, some (perhaps very large) number of alien persons will also come to exist in a causally isolated part of the universe, and we are further informed as to numbers and wellbeing levels of these aliens (which could be anything). No matter the wellbeing distribution associated with the aliens, it is still "repugnant", in the same way, for the drab future to be better.

[6]Recall that the at-least-as-good-as relation $\succeq$ over populations is transitive just in case whenever $A \succeq B$ and $B \succeq C$, we must have $A \succeq C$. Acyclicity, a logically weaker notion, rules out all cycles $A_1 \succ \cdots \succ A_n \succ A_1$.

[7]See Cusbert 2017: §4.2 for a more detailed summary of what I mean by choice-set-dependence and independence.

clear that choice-set-dependent betterness is no panacea for dealing with impossibility theorems: one must reject at least one premise of the Additive Repugnance Theorem *with respect to some choice-set*.

The Additive Repugnance Theorem shows that we must either admit cyclicity, adopt a view of the structure of wellbeing on which there are large gaps which are unbridgeable by finitely many small steps, reject one of two compelling fixed-population principles, reject the Minimal Separable Quality Condition, or accept the additive version of the Repugnant Conclusion. I argue in §4.4 that the last option is the least implausible.

4.2 Mere Addition and Non-Sadism

Parfit's Mere Addition Paradox (1984: ch.19) shows that the Repugnant Conclusion can be derived from plausible principles for comparing same-person populations (which will be left unspecified here, but will be specified in §4.3), the transitivity of the at-least-as-good-as relation, and a single compelling principle for comparing different-number populations, namely the

> *Mere Addition Principle* If X is a population, and Y is a population consisting solely of good lives, then $X + Y$ is at least as good as X.

Recall that the Mere Addition Paradox begins with an arbitrary population A, consisting of many excellent lives. We then consider an addition of

an extremely large number of lives barely worth living to A; call the resulting population A^+. By the Mere Addition Principle, A^+ is not worse than A. By repeatedly applying plausible fixed population principles, we find that A^+ is worse than some population Z, consisting of the same people as in A^+, where everyone has a life which is just a little better than the additional lives in A^+. Applying transitivity, we conclude that Z is not worse than A, which is a Repugnant Conclusion.

How should we respond to the Mere Addition Paradox? One (relatively) attractive option is to simply deny the Mere Addition Principle. It is implausible on its face to think that additions of good lives could make the world worse, but the Repugnant Conclusion may be more implausible still. There is a big (apparent) advantage to taking this path out of the Mere Addition Paradox: we can thereby confine our population-ethical difficulties to variable-population cases.[8] To be sure, it is hard to see *why* the Mere Addition Principle should be false. But perhaps some explanation could be offered.[9]

Later arguments and impossibility theorems are widely thought to close the door on this intuitively appealing exit from the Mere Addition Paradox. In particular, Gustaf Arrhenius has shown that the Mere Addition Principle

[8]We obviously cannot say the same about denying plausible same-person principles, and it also seems unlikely on its face that intransitivity could be confined to variable-population cases.

[9]See for instance Frick 2022.

can be replaced by other different-number principles which are supposed to be more compelling. These are

Dominance Addition If A is a population, and B is a population consisting only of good lives, and A^+ consists of the A-people with higher wellbeing levels than they enjoy in A, then

$$A \preceq A^+ + B$$

Non-Sadism If A is a population consisting of good lives, and B is a population consisting of bad lives, then for any unaffected background population I,

$$I + B \prec I + A$$

Weak Non-Sadism There is some bad wellbeing level b, and some number of lives at this level, such that if B is a population consisting of at least this number of lives at level b, or at some worse level, and A is any population consisting of good lives, and I is any unaffected background population, then

$$I + B \prec I + A$$

Yet it seems to me that if we take seriously the idea that the Mere Addition Principle is false, we should reject all three of these principles. Consider first the principle of Dominance Addition. This says that whenever we add good lives to a population, and additionally make existing people better off, the result is at least as good. But on the assumption that Weak Mere Addition is false, it is sometimes a bad thing that a population contains additional good lives. Although it is clearly a good thing for existing people to be better off, there is no obvious reason to expect that the good thing must always outweigh the bad thing. There is thus no real reason to expect that it cannot be worse for existing people to be made better off, and additional good lives to be added at the same time. This claim is plausible, but only because the Mere Addition Principle is plausible.

Much the same can be said about Non-Sadism. Although it would clearly be a bad thing for an additional bad life to be added to a population, there is no reason to expect that this must always be worse than an addition of good lives, if we think that an addition of good lives can also be a bad thing.[10] Non-Sadism is plausible, but only because the Mere Addition Principle is plausible.

Consider finally the Weak Non-Sadism condition. The sense in which this

[10]Some population axiologies, such as Average Utilitarianism, say that it can be better to add bad lives than not to do so. This kind of claim is very implausible, but axiologies which violate Non-Sadism need not imply this implausible claim. This is noted by Carlson (1998: 302–304).

principle is more compelling than regular Non-Sadism is that intuitively, an addition of good lives could not be *so bad* as to be worse than an addition of a large number of very bad lives. But those who deny the Mere Addition Principle in order to avoid the Repugnant Conclusion need not accept this claim, because they must think that an addition of good lives can be very bad indeed, in certain circumstances.

To see this, let B be some population which is so bad as to make a worse addition than any number of good lives, according to Weak Non-Sadism. We can choose, on the basis of B, some far larger population A in which every person is many times better off than the people in B are badly off, and which contains vastly more people than B. Those who wish to avoid the Mere Addition Paradox by denying the Weak Mere Addition Principle will think that A can, by adding a suitably large population of barely good lives Z^-, be made worse than Z, a population consisting only of barely good lives. But is it obvious that $A + B$ is likewise worse than Z? It is not. While it would be a tragedy for the B-lives to exist, there could be trillions of excellent A-lives for every B-life. Such a population might plausibly be better (or at least not worse) than any population consisting only of lives barely worth living. But if that is right, then (given transitivity) $A + B$ must not be worse than $A + Z^-$. This is exactly what it takes for Weak Non-Sadism to be false.

It is thus unclear at best that the replacements for the Mere Addition Principle typically proposed in the literature really are much more compelling

than the Mere Addition Principle itself. (This is not to say that they are not compelling in the absolute sense.) It would be better to replace the Mere Addition Principle with a principle which is compelling even on the assumption that the Mere Addition Principle is false. Ideally, this replacement should also be clearly more plausible than avoidance of the Repugnant Conclusion. In the next section, I show that we can replace the Mere Addition Principle in a way that meets these desiderata.

4.3 The Additive Repugnance Theorem

4.3.1 Framework: Wellbeing and Populations

I begin by setting out the framework in which we shall operate. I assume that there are infinitely many possible people, and that we have at our disposal some set of wellbeing levels. A population is an assignment of finitely many of these people to these wellbeing levels; any logically possible assignment constitutes a population. If X is a set of possible persons (which I shall usually refer to as a "group" of people) $X[w]$ denotes the population in which the X people exist at wellbeing level w, and nobody else exists. When populations (or groups) X and Y are disjoint, we write $X + Y$ to denote the set-theoretic union of X and Y; that is, the population (group) consisting of the X people and Y people, who are (in the case of populations) at their

respective wellbeing levels in populations X and Y.[11] A choice-set is any finite non-empty set of populations; let $\mathcal{C}$ be the set of all choice-sets. A *population axiology* $\succeq$ is a three-place relation on $P \times P \times \mathcal{C}$, where P is the set of all possible populations; $\succeq$ is reflexive in the sense that for any population X and any choice set C, we have $\succeq (X, X, C)$. Other relations, such as $\succ$, $\preceq$ and $\prec$, are defined from $\succeq$ in the standard way. I shall often use "at-least-as-good-as", "better", and so on to refer to a population axiology and its derived relations.

We will need to make some fairly minimal assumptions about the structure of wellbeing. We will need to assume the existence of a prudential at-least-as-good-as relation on wellbeing levels. I shall denote this relation by $\geq$, with other symbols standing for derived relations in the obvious way, and "betterness" talk referring to the relevant derived relations. The relation $\geq$ is assumed to be transitive and reflexive.[12] For technical reasons, we need to assume that the set of all wellbeing levels is both *upwards directed* and *downwards directed*: for any wellbeing levels w_1, w_2, there exists some w_3 such that $w_3 \geq w_1, w_2$, and some w_4 such that $w_4 \leq w_1, w_2$.[13] This is logically

[11] Any principles which contain this notation should be taken to implicitly quantify only over disjoint populations when the unions of such populations are taken; I shall generally omit such qualifications in order to improve readability.

[12] One might worry that the assumption of transitivity here undermines the motivation to drop it as a requirement for population axiology, since authors who reject transitivity as a requirement on population axiology usually reject it as a requirement on prudential betterness (e.g. Temkin, 2012; Rachels, 1998). We can get around this worry by restricting our attention to a transitive subset of the set of all wellbeing levels.

[13] Alternatively, as in the case of transitivity, it is sufficient to restrict our attention to an upwards- and downwards-directed subset of the set of all wellbeing levels.

weaker (and more compelling) than the completeness requirement, according to which for any wellbeing levels w and w', either $w \geq w'$ or $w \leq w'$.

We shall assume that wellbeing levels can be categorised as being *good* or as *bad*. These correspond to the levels of a life worth living and of a life worth not living respectively. Good lives are better than bad ones. A *neutral* life (wellbeing level) is defined to be a life (wellbeing level) which is neither good nor bad. A *barely good* life is one at a wellbeing level which is good, and which is close to being neutral.[14]

I take "closeness" to be a primitive binary relation on wellbeing levels. Its intended interpretation is exactly what it sounds like: two wellbeing levels are "close" if they are close together. For instance, w and w' are close if the addition or omission of a few pinpricks of pain can make the difference between w being better or worse than w'. The notion of closeness is crucial to our first comparatively controversial (though still not very controversial) assumption about the structure of wellbeing, namely

> *Finite Fine-Grainedness* For any wellbeing levels $w > u$, there exists a finite chain of wellbeing levels $w = w_0 > w_1 > \ldots > w_n = u$ such that each w_i is close to w_{i+1}.

[14]The category of neutral lives, as I have defined it, includes what Gustafsson (2020) calls "undistinguished" lives, which are neither good not bad, yet may not be close to being bad. Gustafsson thinks that repugnant conclusions are not repugnant when they involve lives that are "barely good" in my sense, but are not close to being bad (instead being nearly undistinguished). I disagree, but it would take us too far afield to treat this issue thoroughly.

Thomas (2018) has convincingly argued that Finite Fine-Grainedness is not a principle which is *automatically* true of the structure of wellbeing. *Pace* Arrhenius (2009, 2011), it is possible to produce mathematical models purporting to represent the structure of wellbeing which fail to satisfy Finite Fine-Grainedness. However, no such model succeeds in faithfully representing the structure of wellbeing.

Here's why.[15] Suppose that we take any life, and either make an existing second of that life slightly more painful, or extend the life by a second. Both ways of modifying a life clearly result in at most a small difference in wellbeing. Yet by applying a sufficiently large (but finite) number of such modifications, we can turn any arbitrarily good life into an arbitrarily long life of constant, agonising torture. Since any life is worse than some other life which lasts sufficiently long, and involves sufficiently painful torture at every point, this means that for any lives $w > u$, we can find a finite consecutively close chain from w towards some life w_n which is worse than u. We thus have that any large difference in wellbeing can be bridged by finitely many small steps, and we are therefore justified in assuming Finite Fine-Grainedness.[16]

[15] A similar argument is offered by Arrhenius 2016: 171–172. Another similar argument is suggested by Thomas 2018.

[16] Strictly speaking, this argument supports only the weaker principle of

> *Weak Finite Fine-Grainedness* For any wellbeing levels $w > u$, there exists a finite chain of wellbeing levels $w = w_0 > w_1 > \ldots > w_n$ such that each w_i is close to w_{i+1}, and w_n is worse than u.

But it is hard to see how Finite Fine-Grainedness could be false if Weak Finite Fine-Grainedness is true. In any case, if the objection that Finite Fine-Grainedness is logically stronger were to be pressed, the problem could be avoided in another way: we could

4.3.2 Choice-Set-Dependence and Acyclicity

Say that $A \lhd B$ if and only if there exists some sequence $A_1, A_2, ..., A_n$, with $A_1 = A$ and $A_n = B$, such that for all $i < n$, $A_i \prec A_{i+1}$ in every choice-set containing A_i and A_{i+1}. Given transitivity, this relation is equivalent to strict worseness in all choice sets. Since we are not assuming transitivity, the $\lhd$ relation is weaker than strict worseness in all choice sets. We shall use it as a substitute for strict worseness. There are two reasons to appeal to $\lhd$. First, comparisons in terms of $\lhd$ can be made without the need to worry about choice-set-dependence. Second, and more importantly, $\lhd$ is automatically transitive, even if $\prec$ is not. More precisely, it is easy to prove the

> *Transitivity Lemma* Let X, Y and Z be any populations. Suppose $X \lhd Y$ and $Y \lhd Z$. Then $X \lhd Z$.

The premises of the Additive Repugnance Theorem (except for Acyclicity) will be stated in terms of the $\lhd$ relation. Since this relation is rather artificial, this could make it difficult to determine how plausible these premises are. A small trick can help us to get around this problem. We can imagine that the premises are stated in terms of the "worseness in all choice-sets" relation, rather than $\lhd$. Because "worseness in all choice-sets" is a logically stronger

simply restrict our attention to a chain of wellbeing levels whose consecutive members differ only by pinpricks. Given Weak Finite Fine-Grainedness, such a set can span between any arbitrarily good wellbeing level and any arbitrarily bad wellbeing level, and so the premises of the Additive Repugnance Theorem will remain compelling when our attention is restricted in this way.

relation than $\lhd$, reinterpreting the premises in this way will not make them seem more plausible than they are. (Alternatively, one could literally replace the premises in the way just sketched, but to do so would further complicate the proof.)

Our acyclicity condition will be as follows:

> *Acyclicity* For every population A, it is not the case that $X \lhd X$.

4.3.3 Premises

The Additive Repugnance Theorem has four non-structural premises. Our avoidance condition for the Repugnant Conclusion shall be the

> *Minimal Separable Quality Condition*[17] For any barely good wellbeing level z, there exists some good wellbeing level $a > z$ and bad wellbeing level $b' < z$, and numbers of lives n and m, such that for any groups A and B consisting of n and at least m lives respectively, any group Z, any unaffected background population I, and any bad level $b \leq b'$,
>
> $$I + B[b] + Z[z] \lhd I + A[a]$$

The Minimal Separable Quality Condition requires that an A-population must always make a better addition than a Z-population, but only when some

[17] This principle is similar to "Priority for Lives Worth Living" (Blackorby et al., 2005: 135)

arbitrarily bad population B is also added along with the Z-population. In other words, it is an avoidance condition for an additive version of the Very Repugnant Conclusion.

Figure 4.1: The Minimal Separable Quality Condition

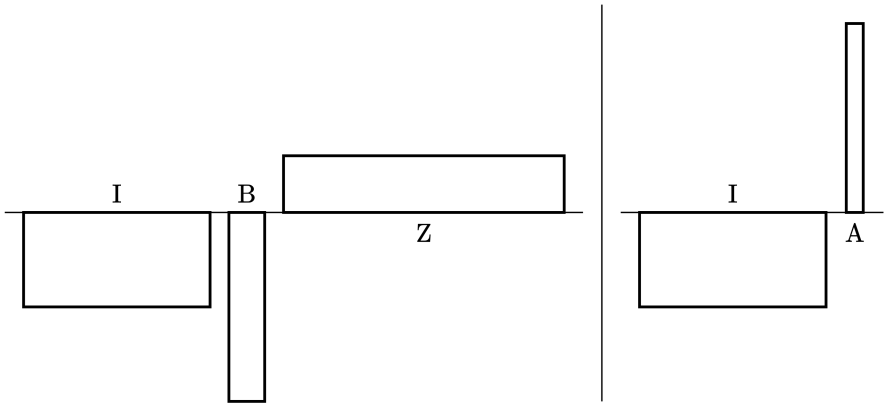


Figure 4.2: *
$I + B + Z \lhd I + A$

Our other different-number condition shall be the

> *Minimal Absolute Value Principle* There exists a bad wellbeing level b which is not minimal,[18] and some number n, such that for any group B of size at least n, any $b^- \leq b$, and any population A containing only good lives,
>
> $$B[b^-] \lhd A$$

The Minimal Absolute Value Principle requires that there is *some* population which is so bad that it is worse than any population consisting solely of

[18] A wellbeing level w is minimal just in case there does not exist any $w^- < w$.

Figure 4.3: The Minimal Absolute Value Principle

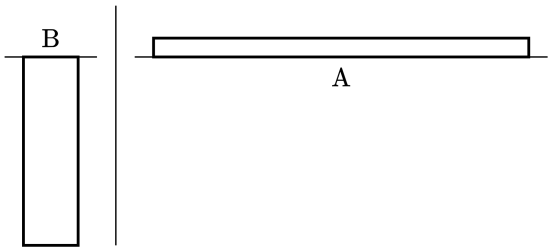


Figure 4.4: *
$B \lhd A$

good lives. While some of the premises of our impossibility theorem might be up for debate, I believe that this one is not, *provided* one countenances any different-number comparisons whatsoever.[19] The Minimal Absolute Value Principle is not satisfied by all proposed population axiologies; it is not, for instance, satisfied by positive critical level views, such as those discussed by Blackorby et al. (2005: ch.5). But such views are to be rejected precisely because they do not satisfy the Minimal Absolute Value Principle.[20]

Our fixed population principles shall be analogues of the ones appealed to by Arrhenius (2009, 2011, nd). They are more or less the same as the same-named principles in these works, with the main difference being that their

[19]Bader (2022a,b) notably demurs from this assumption, claiming instead that all different-number populations are incomparable. I am inclined to reject his view precisely on the grounds that it does not validate the Minimal Absolute Value Principle. In any case, Bader's view fails to satisfy the Minimal Separable Quality Condition.

[20]The usual rationale offered for critical level views is precisely that they allow for avoidance of the Repugnant Conclusion. But since avoidance of the (additive or non-additive) Repugnant Conclusion is less intrinsically plausible than the Minimal Absolute Value Principle, this kind of rationale is insufficient to justify these views.

claims are in terms of the $\lhd$ relation, rather than the worseness relation.[21] Apart from this, they differ only notationally, and in one or two unimportant respects.[22] They are as follows.

General Non-Elitism

For any wellbeing levels $a > e > c$, where a is close to e, there exists a number n such that for any single-person group A, any group C of at least n people, and any unaffected background population I,

$$I + A[a] + C[c] \lhd I + A[e] + C[e]$$

General Non-Extreme Priority For any wellbeing level w, and any barely good level z, there is a good wellbeing level $a' > z$, and a number n, such that for any wellbeing levels $a \geq a'$ and $w^+ > w$, where w^+ is

[21] Since Arrhenius (2009, 2011) assumes transitivity and implicitly assumes choice-set-independence, the two relations are equivalent in his framework.

[22] My version of General Non-Extreme Priority differs in that because of a reordering of some of the quantifiers it is compatible with the view that priority given to the worse-off increases without a bound as people move down the wellbeing scale. Additionally, my versions of General Non-Elitism and General Non-Extreme Priority are formulated as fixed-population principles; in contrast, Arrhenius (2009, 2011) states them as same-number principles. The final difference is that many variables in my statements of the conditions are formulated as inequalities: such-and-such holds for any population of at least this number of lives, at at least/at most this level, rather than for any population of exactly this many lives at exactly this level. The strengthenings in this respect are harmless in terms of their impact on the plausibility of the premises.

Figure 4.5: General Non-Elitism

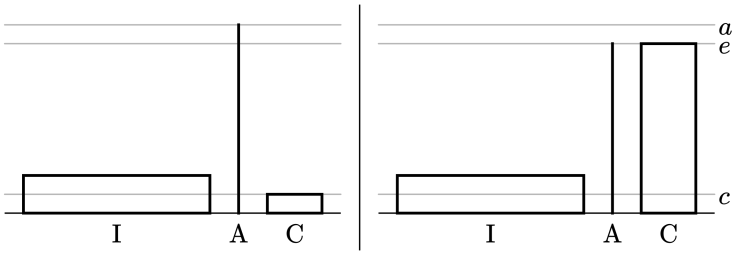


Figure 4.6: *
$I + A[a] + C[c] \lhd I + A[e] + C[e]$

close to w, any group A of size at least n, any single-person group C, and any unaffected background population I,

$$I + A[z] + C[w^+] \lhd I + A[a] + C[w]$$

Figure 4.7: General Non-Extreme Priority

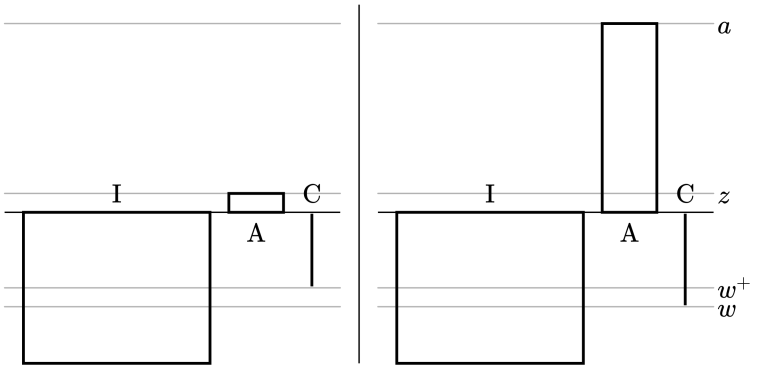


Figure 4.8: *
$I + A[z] + C[w^+] \lhd I + A[a] + C[w]$

General Non-Elitism says that rather than providing a small benefit to

a single better off person, it would be better to instead provide benefits of a fixed size to a sufficiently large number of worse-off people. General Non-Extreme Priority says that rather than benefiting a single person who is perhaps badly off to some small extent, it would be better to instead lift some sufficiently large number of people up from a barely good wellbeing level to some sufficiently good wellbeing level. Both principles are very plausible.

4.3.4 The Additive Repugnance Theorem

This is enough groundwork to state the theorem:

> *The Additive Repugnance Theorem* There is no population axiology which satisfies all of the following conditions:
>
> (1) Finite Fine-Grainedness
>
> (2) Acyclicity
>
> (3) The Minimal Separable Quality Condition
>
> (4) The Minimal Absolute Value Principle
>
> (5) General Non-Elitism
>
> (6) General Non-Extreme Priority

4.3.5 Lemmas

To prove the Additive Repugnance Theorem, we shall need to appeal to three lemmas. We shall need to show that our premises imply "Inequality-Averse Addition", "Sufficient Trade-Offs", and "Axiological Aggregation". Roughly, according to Inequality-Averse Addition, large improvements to a first group of people are outweighed by small improvements to a sufficiently large second group of people who are worse off than the first group. According to Sufficient Trade-Offs, rather than having a first and second group of people, all at some barely good wellbeing level, it would be better if instead the first group were at some very good wellbeing level, and the second group were at some very bad wellbeing level, provided the first group is sufficiently larger than the second. According to Axiological Aggregation, rather than having a first and second group of people, all at some barely good wellbeing level, it would be better if the first group were at some slightly better level, and the second group were at some bad level, provided the first group is sufficiently larger than the second. Each of these principles applies in the presence of any unaffected background population. More precisely, the three lemmas we need are as follows:

> *The Inequality Aversion Lemma* The General Non-Elitism Condition and Finite Fine-Grainedness imply

Inequality-Averse Addition For any wellbeing levels $a > e > c$, and any number n, there is a number m such that if A and C are groups containing n and at least m lives respectively, and I is any unaffected background population,

$$I + A[a] + C[c] \lhd I + A[e] + C[e]$$

The Sufficient Trade-Offs Lemma The General Non-Extreme Priority condition and Finite Fine-Grainedness imply

Sufficient Trade-Offs For any barely good z, any wellbeing level $b < z$, and any number n, there is a good wellbeing level $a' > z$ and a number m such that for any $a \geq a'$, any group A of size at least m, any group C of size n, and any unaffected background population I,

$$I + A[z] + C[z] \lhd I + A[a] + C[b]$$

The Axiological Aggregation Lemma The General Non-Elitism Condition, General Non-Extreme Priority, and Finite Fine-Grainedness imply

Axiological Aggregation For any barely good z, any $z^+ > z$ which is not maximal, any $b < z$, and any number of lives n, there

is a number of lives m such that if B and Z are groups of n and m people respectively, and I is any unaffected background population, then

$$I + B[z] + Z[z] \lhd I + B[b] + Z[z^+]$$

The ideas of the proofs of the first two lemmas are fairly simple, and are exactly analogous to the proofs of Arrhenius's (2011) Lemmas 1.1 and 1.2 respectively. Both Inequality-Averse Addition and Sufficient Trade-Offs result from applying General Non-Elitism and General Non-Extreme Priority finitely many times. If a trade-off can be made between one person and m people, then, by applying such a trade-off n times, one can show that the same kind of trade-off can be made between n people and $n \cdot m$ people. Similarly, we can drop the restriction that only small differences in wellbeing can be traded-off by repeatedly applying such many-person trade-offs along a finite chain of consecutively close wellbeing levels between any two wellbeing levels.[23] Such a chain always exists by the Finite Fine-Grainedness condition. Essentially, the proofs of the first two lemmas just consist in showing that both things can be done. Axiological Aggregation may then be proved by applying both lemmas finitely many times (that is, by induction).

[23] A chain w_i of wellbeing levels is consecutively close just in case each w_j is close to w_{j+1}.

4.3.6 Proof of the Additive Repugnance Theorem

The theorem is proved by constructing groups B, B', Z and A, and wellbeing levels $a > z > z^- > b > b^-$, in such a way that the populations illustrated in Figure 4.10 below bear the $\lhd$ relations as stated in the caption of the same Figure. In this construction, a is a good wellbeing level, z and z^- are barely good levels, and b and b^- are bad wellbeing levels.

Figure 4.9: The Proof of the Impossibility Theorem

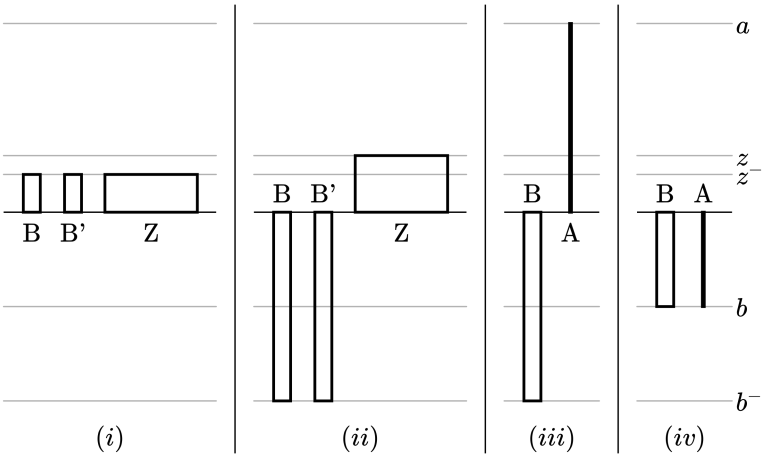


Figure 4.10: *
$(i) \lhd (ii) \lhd (iii) \lhd (iv) \lhd (i)$

We choose the sizes of groups A and B', and wellbeing levels a and b^-, in such a way that the claim that $(ii) \lhd (iii)$ is an instance of the Minimal Separable Quality Condition, with B the unaffected background population. We additionally stipulate that the size of group B is sufficiently large for us to obtain that $(iii) \lhd (iv)$, by Inequality-Averse Addition. We obtain $(iv) \lhd (i)$

by the Minimal Absolute Value Principle.[24] Since nothing said so far turns on the size of group Z, we can ensure that Z is chosen to be large enough that Axiological Aggregation implies that $(i) \lhd (ii)$. Finally, applying the Transitivity Lemma, we obtain that $(i) \lhd (i)$, contradicting Acyclicity.

4.4 Options

Since the premises of the Additive Repugnance Theorem are mutually inconsistent, at least one of them has to go. But which? For reasons stated earlier, I believe that we should not reject either Finite Fine-Grainedness or the Minimal Absolute Value Principle.[25] The remaining premises are Acyclicity, General Non-Elitism and General Non-Extreme Priority. Alternatively, we can reject the Minimal Absolute Value Principle, thereby accepting an additive version of the Repugnant Conclusion. I shall consider each option in turn.

[24]That is, in constructing B, we ensure that it is large enough to the appropriate role in the application of both Inequality-Averse Addition and the Minimal Absolute Value Principle.

[25]Rejecting Finite Fine-Grainedness seems to be a non-starter in any case: Thornley (2021) has shown that we can make do without this principle in Arrhenius's Sixth Theorem if we use probabilistic versions of General Non-Elitism and General Non-Extreme Priority. This move could be adapted for the Additive Repugnance Theorem. (I have avoided doing so in order to keep the proof relatively simple.)

4.4.1 Acyclicity

It has been suggested, most prominently by Larry Temkin (1987; 1996; 2012) and Stuart Rachels (1998; 2001; 2004), that we might avoid the Repugnant Conclusion by denying the transitivity of the at-least-as-good-as relation. While both authors also deny Acyclicity, it seems to me that the two moves are not on a par: it is less plausible to deny Acyclicity than it is to deny transitivity. To see why, let us first consider the case for (in)transitivity, and next consider the case for Acyclicity.

Why might we antecedently want to accept transitivity, beyond the mere intuitive plausibility of that principle? One standard argument holds that we must accept transitivity because it is an analytic feature of comparatives: as a matter of logic, whenever A is F-er than B, and B is F-er than C, A must be F-er than C (Broome, 2004). It might alternatively be claimed that transitivity is otherwise central to the concept of value, or that value is inherently quantitative (and therefore transitive).[26] I'm not sure whether these arguments succeed, but even if we assume that they do, sceptics of transitivity still have the nuclear option: they can say that talking about value is a mistake, and that we should instead theorise in terms of some other non-transitive normative relation.[27] If some such normative relation

[26] See Klocksiem 2016.

[27] Rachels (2001: 218–219) suggests that even if transitivity is held to be central to value, one might nevertheless understand intransitivity in terms of some other normative relation.

fulfils whatever role we wanted goodness to play in our moral theory, it is unclear that much is lost by abandoning value-talk in favour of theorising in terms of this new relation.

As an example, say that one population is *impartially preferable* to another just in case there is all-things-considered reason to hope, from an impartial perspective, that the first population would come about rather than the second.[28] The impartial preferability relation would appear to be able to do everything we want a value relation to do, but as far as I can see, there is no obvious argument to the effect that the impartial preferability relation must be transitive.[29]

There are, however, strong pragmatic arguments for the irrationality of cyclic preferences. The most prominent of these are money pump (or value pump) arguments, some versions of which even apply to agents who take into account their expected future choices in their present decision-making.[30] If these arguments successfully show that cyclic preferences are irrational, the

[28]Some authors, such as Parfit (2011: 41–2), understand *value* in a similar "impartial-reason-implying sense".

[29]There are some pragmatic arguments against acyclic intransitivity; see for example Gustafsson 2010, nd. The problem is that these arguments assume completeness, which on its face is less secure than transitivity itself. Gustafsson (nd) provides pragmatic arguments for completeness, but these arguments are weaker than the pragmatic arguments for acyclicity.

[30]A recent version of this argument, provided by Gustafsson and Rabinowicz (2020), shows that any agent with cyclic preferences is susceptible to exploitation, provided only that she obeys a minimal principle of backwards induction. Moreover, the pump they provide can be iterated, showing that any such agent can be forced to give up arbitrarily much of something they care about (2020: 586). In the present context, the relevant quantity would be people's wellbeing.

betterness relation must be acyclic (given that choosing in accordance with the betterness relation is not irrational). Importantly, pragmatic arguments apply just as well to any relation which might be offered up by a transitivity sceptic taking the nuclear option. If a normative relation R is to play the role of traditional value relations in our conceptual theorising, one feature it must have is to be morally decisive in cases in which only R-relevant factors are at stake. As a result, if R is cyclic, abiding by morality will sometimes require one to act on cyclic preferences. Therefore, given that successfully abiding by morality is not irrational, that rational agents are not susceptible to exploitation by money pump, and that agents with cyclic preferences are susceptible to exploitation by money pump, any normative relation R purporting to stand in for traditional value relations must satisfy Acyclicity.

Note also that the only way to deny that agents with cyclic preferences are susceptible to exploitation by money pump is to claim that an agent's preferences at some point in a decision tree can turn on more than just the options achievable for her going forward from that point.[31] But if that is true, then even if a theory says (for instance) that populations of excellent lives are better than very large populations of lives barely worth living, the same theory may nevertheless instruct an agent to bring about the population of lives barely worth living, if the agent is at some suitable point in a decision

[31]See McClennen 1985, or more recently, Ahmed 2017.

tree. Any non-exploitable theory must avoid actually issuing its usual judgement for at least some pair of populations involved in a betterness cycle, for at least some points in a decision tree. Denial of Acyclicity is therefore not a clean way of avoiding the Repugnant Conclusion without incurring other problematic commitments. Even if Acyclicity is false, we must still deny at least one of the premises of the Additive Repugnance Theorem in at least some sequential decision contexts (on pain of having to say, absurdly, that rational moral agents sometimes *should* get money pumped, even when they see it coming).

4.4.2 General Non-Elitism and General Non-Extreme Priority

Recall that according to General Non-Elitism, rather than slightly benefiting one person who is comparatively well off, it would be better to instead benefit some sufficiently large number of people who are worse off. According to General Non-Extreme Priority, rather than slightly benefiting one person, who may initially be very badly off, it would be better to instead provide large benefits to a sufficiently large number of people whose lives are barely worth living.

For reasons mentioned in §4.3.2, we shall understand these claims as applying in every choice set. Such claims to the effect that some population

A is worse than B in every choice set can be denied in a number of ways, some of which are more plausible than others. First, one might reject the claim outright, insisting instead that there are cases where A is better than B in every choice-set. Second, one could adopt the weaker position that A and B could be incomparable in every choice-set.[32] Third, one could admit that it is determinately true that there is some case in which A is not worse than B (in every choice-set), but deny, for each pair of populations A and B falling under the principle, that it is determinately true that A is not worse than B (in every choice-set).[33] Fourth, one could accept that A is worse than B in the pairwise choice set consisting of just A and B, while allowing that this may fail to be the case in some larger choice sets.[34]

There is little to be said for the first option. It would be very implausible to claim that it would be better to benefit a single better off person, rather than arbitrarily many worse-off people. It would be only slightly less implausible to claim that it would be better to benefit a single badly off person by a very small amount, rather than benefiting a large number of people with lives barely worth living by a great amount. In fact, in taking the first option, one would have have to accept something even harder to believe than these

[32]I define two populations to be incomparable just in case they are unrelated to each other by $\preceq$. It does not matter for the purposes of this chapter whether this means that no positive evaluative relation at all holds between the populations, or whether some positive evaluative relation not definable in terms of $\preceq$, such as parity (Chang, 2002) or imprecise equality (Parfit, 2016), might nevertheless hold.

[33]See Thomas nd,f.

[34]This option is advocated by Frick 2022 and, for a non-evaluative interpretation of $\succ$, by Boonin-Vail 1996.

implausible claims. Thornley (2021) has shown that the two fixed-population principles can be weakened in the following way: the potential losses to one person in each principle can be replaced by arbitrarily small probabilities of a loss to one person. As far as I can see, this sinks the first option.

The remaining three options represent different paths for mitigating the intuitive costs of denying General Non-Elitism or General Non-Extreme Priority, but it is unclear how much they really help. Consider, for instance, General Non-Elitism. It seems clearly false that benefiting a single person by a small amount could be better than benefiting an arbitrarily large number of less well-off people by larger amounts. It also seems clearly false that benefiting the better-off person could be not worse than benefiting the less well-off people, or that benefiting the better-off person could fail to be determinately worse than benefiting the less well-off people. It is hard to see how even the last, weakest judgement could fail in *any* choice set. Again, these claims become even more secure if we shift to probabilistic versions of our fixed-population principles. The prospects for rejecting General Non-Elitism or General Non-Extreme Priority, even when mitigated by taking one or more of options two to four, thus look dim to me.

4.4.3 Accepting the Repugnant Conclusion

To reject the Minimal Separable Quality Condition is to accept that a population of excellent lives may fail to make a better addition than a population consisting of some very bad lives, together with arbitrarily many lives barely worth living. If this is accepted, it would seem natural to also accept a stronger, non-additive version of this claim: a population of excellent lives can be worse, by itself, than a combination of many bad lives with arbitrarily many lives which are barely worth living. That is, if we reject the Minimal Separable Quality Condition, we should accept the Very Repugnant Conclusion.

One might deny this last claim. Avoidance of the Very Repugnant Conclusion is *consistent* with the negation of the Minimal Separable Quality Condition, just as the Minimal Absolute Value Principle is consistent with the negation of Weak Non-Sadism. (To see that the first of these claims is true, note that Average Utilitarianism avoids the (Very) Repugnant Conclusion, but does not satisfy the Minimal Separable Quality Condition.)[35] One might thus worry that the Additive Repugnance Theorem tells us nothing new that is important: Weak Non-Sadism is arguably more plausible than the Minimal Separable Quality Condition, and we already know from the Sixth Impossibility Theorem that (assuming the other premises) we must choose

[35]This was first noticed by Bill Anglin (1977), and more recently by Gustaf Arrhenius (nd).

between Weak Non-Sadism and avoidance of the Repugnant Conclusion. So why worry? If we deny the Minimal Separable Quality Condition, we are not much worse off than before with respect to avoiding the Very Repugnant Conclusion, because we have only rejected a principle which is less plausible than Weak Non-Sadism.

Put this way, I think the worry is overstated, because it seems to me that it matters if we are forced to accept the Minimal Separable Quality Condition, even if this condition is intrinsically less plausible than Weak Non-Sadism. We are inclined to think that the Repugnant Conclusion is false because of its particular character: it *could not be*, we think, that a population of lives barely worth living is better than a population of excellent lives, just because the former contains a very large number of people. But if the Minimal Separable Quality Condition is false, this intuition is undermined: it *can* be that a population of lives barely worth living is better as an addition than a population of excellent lives, just because the former contains a very large number of people. In contrast, if we deny Weak Non-Sadism, this does not by itself undermine the intuition of repugnance. Rejecting the Weak Non-Sadism condition can be a way of robustly avoiding the Repugnant Conclusion; rejecting the Minimal Separable Quality Condition cannot.

Accepting the Repugnant Conclusion is a radical step. Most authors who have thought seriously about the Repugnant Conclusion take it to be deeply counterintuitive, and many of these authors take there to be decisive reason

to reject any population axiology which implies the Repugnant Conclusion.[36] Even so, it seems to me that accepting the (Very) Repugnant Conclusion is the least implausible response to the Additive Repugnance Theorem.

4.5 Conclusion

I have presented the Additive Repugnance Theorem, which demonstrates that an additive version of the Repugnant Conclusion is extremely difficult to avoid. Unlike most other impossibility theorems, the Additive Repugnance Theorem does not include any version of the Mere Addition Principle, or any different-number principle which could feasibly be denied on the grounds that Mere Addition is false. Instead, it assumes the compelling Minimal Absolute Value Principle. The main practical upshot, I think, is this. In a choice between Weak Non-Sadism and avoidance of the Repugnant Conclusion, it may be unclear which should stay and which should go. Arguably, both are similarly compelling. In contrast, in a choice between the Minimal Absolute Value Principle and the Minimal Separable Quality Condition, it is clear that the former is by some margin the more plausible principle. Since these are the only two different-number principles involved in the Additive Repugnance Theorem, we may conclude that the Minimal Separable Quality Condition cannot reasonably be maintained by abandoning the Minimal

[36]That said, the traditional view that the Repugnant Conclusion is to be avoided at all costs no longer enjoys the near-unanimity it once had; see Zuber et al. 2021.

Absolute Value Principle. Neither do denial of Acyclicity or Finite Fine-Grainedness offer a straightforward path to retaining the Minimal Separable Quality Condition. Instead, the most plausible response to the Additive Repugnance Theorem, for opponents of the Repugnant Conclusion, may be to reject the fixed-population assumptions: either General Non-Elitism or General Non-Extreme Priority (or both). I have suggested that the intuitive costs of doing so might be mitigated by appealing to incompleteness, choice-set-dependent betterness, indeterminacy, or some combination of the three. Yet even with these mitigating factors in play, I suspect that these principles are simply too compelling to deny.

Let me make one final point. While the relation $\succeq$ was interpreted throughout this chapter as the moral at-least-as-good-as relation, its structure as a three-place relation incorporating choice-set-dependence means that it can be given other interpretations. In particular, it can be interpreted as the at-least-as-much-reason-to-bring-about relation. The Additive Repugnance Theorem on this normative interpretation of $\succeq$ might be more troubling than the axiological interpretation of the theorem. On my favoured response to the Additive Repugnance Theorem, we would sometimes have more reason to bring about a future consisting of many tortured lives, and many more barely good lives, than we would have to bring about a future consisting of some smaller (but still perhaps large) number of excellent lives. This might be even more difficult to believe than the axiological Very Re-

pugnant Conclusion.

Chapter 5

Intrapersonal Arguments for the Repugnant Conclusion

Abstract

In "An Intrapersonal Addition Paradox" (2019), Jacob Nebel provides a novel "intrapersonal" argument for the Repugnant Conclusion. The most controversial premise of Nebel's Intrapersonal Addition Paradox is the "Probable Addition Principle", on which (roughly) it is better for individuals to receive additional chances of existence with a life worth living. This chapter argues that those who believe that existence cannot be better or worse for an individual than non-existence need not accept the Probable Addition Principle. However, an alternative intrapersonal argument for the Repugnant Conclusion is provided which does not assume the Probable Addition Principle. Another argument is used to show that "Weak Pareto for Equal Risk" (another premise of the Intrapersonal Addition Paradox) implies an intuitively "repugnant" conclusion, given just one very minimal intrapersonal principle. It is argued that in light of these arguments, those who wish to avoid the Repugnant Conclusion must deny Weak Pareto for Equal Risk.

5.1 Introduction

Most people find the following proposition very difficult to believe:

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

However, there are powerful arguments for the Repugnant Conclusion; consequently, avoiding it may be even more difficult than believing it. The most influential such argument is Parfit's Mere Addition Paradox, but there are many others.[1] Most of these arguments appeal solely to intuitively plausible principles for comparing populations involving different numbers of people, and to principles for making trade-offs between people who exist in both populations under consideration. Let's call these "interpersonal arguments" for the Repugnant Conclusion.

In "An Intrapersonal Addition Paradox", Jacob Nebel (2019) pioneers another kind of argument for the Repugnant Conclusion: an *intrapersonal*

[1] See Blackorby et al. 2003, Carlson 1998, Arrhenius 2000, 2003, 2011 and Spears and Budolfson 2021, among many others.

argument.[2] Rather than appealing directly to plausible principles for comparing populations, intrapersonal arguments appeal to intrapersonal principles (that is, principles governing which prospects are better for *individuals*), and then derive principles about which prospects are better *overall* by applying certain bridge principles, including some version of the ex ante Pareto principle, on which if prospect $\mathcal{X}$ is better for everyone than $\mathcal{Y}$, $\mathcal{X}$ must be better overall than $\mathcal{Y}$.

This chapter is about intrapersonal arguments for the Repugnant Conclusion, with particular attention to Nebel's Intrapersonal Addition Paradox, which I recap in §5.2. As Nebel recognises, the most controversial premise of the Intrapersonal Addition Paradox is the "Probable Addition Principle", on which (roughly) it is better for a person to receive additional chances of existence with a good life. This principle is controversial in large part because it seems hard to square with

> *Non-Comparativism* An outcome in which $\mathcal{S}$ exists cannot be better or worse for $\mathcal{S}$ than an outcome in which $\mathcal{S}$ does not exist.

In §5.3, I make more precise the sense in which Probable Addition is incompatible with Non-Comparativism: the two principles are inconsistent if we assume the Sure Thing Principle. Although it is not completely obvious that Non-Comparativists should accept the Sure Thing Principle, I show

[2]Intrapersonal arguments in population ethics are also discussed by Thomas (2016) and McCarthy et al. (2020).

that without it, Nebel's positive argument for the Probable Addition Principle does not succeed. Either way, Non-Comparativists have little reason to accept the Probable Addition Principle, and thus little reason to worry about the Intrapersonal Addition Paradox.

In §5.4, however, I provide a new intrapersonal argument for the Repugnant Conclusion which replaces the Probable Addition Principle with the "Conditional Value Principle", which says that if $\mathcal{X}$ gives $\mathcal{S}$ a non-zero probability of existence with a good life, otherwise non-existence, and $\mathcal{Y}$ gives $\mathcal{S}$ a non-zero probability of existence with a bad life, otherwise non-existence, then $\mathcal{X}$ is better for $\mathcal{S}$ than $\mathcal{Y}$. The Conditional Value Principle is intuitively much more compelling than the Probable Addition Principle, and does not tacitly assume Comparativism. I argue that in light of this new intrapersonal argument, the only plausible response for opponents of the Repugnant Conclusion is to reject "Weak Pareto for Equal Risk", the ex ante Pareto principle appealed to by the Intrapersonal Addition Paradox and by the argument of §5.4. I further substantiate this last claim by showing in §5.5 that Weak Pareto for Equal Risk implies an intuitively "repugnant" conclusion when conjoined with just one compelling intrapersonal principle. Notably, unlike most arguments for "repugnant" conclusions, the argument of §5.5 does not assume the transitivity of the at-least-as-good-as relation. I conclude in §5.6 with some thoughts on the choice between Weak Pareto for Equal Risk and avoidance of the Repugnant Conclusion.

5.2 The Intrapersonal Addition Paradox

The Intrapersonal Addition Paradox can be broken down into two stages. The first stage, which Nebel calls the "probable addition argument", aims to establish the

> *Intrapersonal Repugnant Conclusion* For any person $\mathcal{S}$, there is some probability p such that any prospect in which $\mathcal{S}$ would have a wonderful life with probability p or less, and would otherwise never exist, is worse for $\mathcal{S}$ than a certainly mediocre life.
>
> Nebel 2019: 314

By a "mediocre" life (or wellbeing level), I mean one that is barely worth living. By a "wonderful" or "excellent" life (or wellbeing level), I mean one that contains sufficiently large quantities of whatever makes life worth living.

The second "interpersonal" stage of the argument derives the Repugnant Conclusion from the Intrapersonal Repugnant Conclusion, via four bridge principles. We now turn to the Probable Addition Argument.

5.2.1 The Probable Addition Argument

The probable addition argument has two premises. First, the

> *Probable Addition Principle* For any prospects $\mathcal{X}$ and $\mathcal{Y}$ and any person $\mathcal{S}$ who might exist in those prospects: if, in every state of nature in

> which $\mathcal{S}$ would exist in Y, $\mathcal{S}$ would be better off in X, and if, in every other state of nature, $\mathcal{S}$'s life would be worth living in X, X is better for $\mathcal{S}$ than Y.
>
> Nebel 2019: 315

The Probable Addition Principle says that it is better for an individual to get a greater level of wellbeing in the states of nature in which she would have existed anyway, plus existence with a life worth living in some additional states of nature in which she otherwise would not have existed, even if she is significantly worse off in these additional states. We shall discuss the Probable Addition Principle at length in §5.3, but it is worth noting at the outset that the Probable Addition Principle, while not implausible, is not all that compelling considered by itself.

The second premise is

> *Minimal Prudence* For any individual $\mathcal{S}$ and very high welfare level a, there are some mediocre welfare levels z and z^- (where $z > z^-$) and some probability p such that any prospect in which $\mathcal{S}$ is certain to exist at level z is better for $\mathcal{S}$ than any prospect in which $\mathcal{S}$ might, with any probability less than or equal to p, exist at level a, and would otherwise exist at level z^-.
>
> Nebel 2019: 316

Minimal Prudence looks a little complicated when set out like this, but what it says is very plausible: it is better for an individual to receive certainty of a slightly better wellbeing level, even if this comes at the cost of foregoing a sufficiently small probability of a much better wellbeing level. I share Nebel's opinion that judgements like this are beyond serious doubt (2019: 317), and so I shall generally assume that rejecting Minimal Prudence is off the table.

The probable addition argument proceeds as follows. Let $a^+ > a$ be excellent wellbeing levels, and let $z > z^-$ be mediocre wellbeing levels. By minimal prudence, there is some probability p such that certainty of existence at z is better than existence a probability p of existence at a^+, or existence with z^- otherwise. Consider now the following three prospects, where the wellbeing levels specified are the wellbeing levels of some person $\mathcal{S}$, and Ω represents the non-existence of $\mathcal{S}$:

	State 1 (probability p)	State 2 (probability $1-p$)
$\mathcal{A}$	a	Ω
$\mathcal{A}^+$	a^+	z^-
$\mathcal{Z}$	z	z

The Probable Addition Principle implies that $\mathcal{A}^+$ is better for $\mathcal{S}$ than $\mathcal{A}$. Minimal Prudence implies that $\mathcal{Z}$ is better for $\mathcal{S}$ than $\mathcal{A}^+$. Transitivity of the at-least-as-good-as relation then implies that $\mathcal{Z}$ is better for $\mathcal{S}$ than $\mathcal{A}$, which is the Intrapersonal Repugnant Conclusion.[3]

[3]Like Nebel, I assume the transitivity of the moral and prudential at-least-as-good-as relations throughout this chapter. Some philosophers deny (or at least doubt) transitivity

Next, we shall use the Intrapersonal Repugnant Conclusion to derive the (interpersonal) Repugnant Conclusion.

5.2.2 The Interpersonal Stage

The interpersonal stage of the Intrapersonal Addition Paradox has four more premises. First, some terminology: a prospect is *egalitarian* if and only if each person who might exist has the same chances of existing at each wellbeing level as everyone else (so that there is perfect ex ante equality), and each possible outcome is a perfectly equal population (so that there is perfect ex post equality). We can now state the four premises:

> *Stochastic Indifference for Equal Risk* For any egalitarian prospects $\mathcal{X}$ and $\mathcal{Y}$, if every possible outcome of $\mathcal{X}$ and every possible outcome of $\mathcal{Y}$ are equally good, then $\mathcal{X}$ and $\mathcal{Y}$ are equally good.
>
> Nebel 2019: 319

> *Weak Pareto for Equal Risk* For any egalitarian prospects $\mathcal{X}$ and $\mathcal{Y}$, if $\mathcal{X}$ is better than $\mathcal{Y}$ for each person who might exist in either prospect, then $\mathcal{X}$ is better than $\mathcal{Y}$.
>
> Nebel 2019: 320

due to its role in arguments for the Repugnant Conclusion, notably Temkin 1987, 2012 and Rachels 1998, 2001, 2004. In §5.5, we shall see an intrapersonal argument for a "repugnant" conclusion which does not assume transitivity.

> *Same-Number Quality Claim* Any two populations consisting of the same number of people, all at the same levels of well-being, are equally good.
>
> Nebel 2019: 318

> *Certainty Equivalence* For any riskless prospects $\mathcal{X}$ and $\mathcal{Y}$, which guarantee populations X and Y respectively, $\mathcal{X}$ is better than $\mathcal{Y}$ just in case X is better than Y.
>
> Nebel 2019: 322

The interpersonal stage of the argument goes as follows. Let A be any population of m lives at excellent wellbeing level a. Let p be a sufficiently small probability that certainty of life at the mediocre level z is better than a p chance of life at level a and non-existence otherwise (by the Intrapersonal Repugnant Conclusion). Given this small probability p, we can construct a set of states of nature $S_1, \ldots S_n$, where each state has probability $1/n < p$. We can also construct n disjoint sets of possible people $G_1, G_2, \ldots, G_n$, each containing m people, and with G_1 being the set of A-people; write G to denote the set of all people who belong to some G_i. Some notation: if X is any set of people, we write $X[w]$ to denote the outcome in which the X people exist at level w, and nobody else exists. Now consider the following three prospects:

	S_1	S_2	...	S_n
$\mathcal{A}$	$G_1[a]$	$G_1[a]$	...	$G_1[a]$
$\mathcal{A}'$	$G_1[a]$	$G_2[a]$	...	$G_n[a]$
$\mathcal{Z}$	$G[z]$	$G[z]$	...	$G[z]$

Notice that all three prospects are egalitarian. The Same-Number Quality Claim and Stochastic Indifference for Equal Risk thus imply that $\mathcal{A}$ and $\mathcal{A}'$ are equally good. Since certainty of z is better for each G-person than a tiny probability of a and non-existence otherwise (by the Intrapersonal Repugnant Conclusion), Weak Pareto for Equal Risk implies that $\mathcal{Z}$ is better than $\mathcal{A}'$. Transitivity then implies that $\mathcal{Z}$ is better than $\mathcal{A}$. Since $\mathcal{A}$ and $\mathcal{Z}$ guarantee populations A and $G[z]$ respectively, Certainty Equivalence implies that $G[z]$ is better than A, which is the Repugnant Conclusion.

5.3 The Probable Addition Principle

For most of us, accepting the Repugnant Conclusion will be a last resort.[4] To avoid this last resort, we need to deny at least one of the six premises of the Intrapersonal Addition Paradox. Some of these premises, I think, are not up for rejection. Minimal Prudence is one of them, for reasons mentioned earlier. It is also not promising to reject Certainty Equivalence, Stochastic Indifference for Equal Risk or the Same-Number Quality Claim. Even if we

[4]Not everyone takes acceptance of the Repugnant Conclusion to be a last resort. See Ng 1989 and Huemer 2008.

deny all three of these principles, we would still be left with the conclusion that $\mathcal{A}'$ is better than $\mathcal{Z}$. This seems to me just as "repugnant" as the Repugnant Conclusion.

Those who wish to avoid the Repugnant Conclusion thus need to deny either the Probable Addition Principle or Weak Pareto for Equal Risk. Of the two, the Probable Addition Principle is significantly less intuitively compelling. So the most obvious way to avoid the Intrapersonal Addition Paradox is to deny the Probable Addition Principle. It might seem that there are good grounds to do so. Recall that according to *Non-Comparativism*, an outcome in which $\mathcal{S}$ exists cannot be better or worse for $\mathcal{S}$ than an outcome in which $\mathcal{S}$ does not exist. Many people find Non-Comparativism independently compelling.[5] But technically, Non-Comparativism is outright inconsistent with the Probable Addition Principle.[6] To see this, note that if $\mathcal{S}$ certainly exists with a good life in $\mathcal{X}$, and certainly does not exist in $\mathcal{Y}$, then $\mathcal{S}$ is better off in $\mathcal{X}$ for all states of nature in which she would exist in $\mathcal{Y}$ (because there are none), and has a life worth living in $\mathcal{X}$ for all other states of nature. The Probable Addition Principle therefore implies that $\mathcal{X}$ is better for $\mathcal{S}$ than $\mathcal{Y}$. This amounts to a violation of Non-Comparativism, given Certainty Equivalence.

[5]See Broome 1999: 168, Bykvist 2007, McMahan 2013: 7 and (seemingly) Parfit 1984: 489.

[6]This point is not my own; [Redacted] credit it to an anonymous referee (private correspondence).

This argument might seem a little cheap, because the Probable Addition Principle might be intended to apply only when $\mathcal{S}$ is better off in $\mathcal{X}$ than in $\mathcal{Y}$ for some state of nature (so that she must have a non-zero probability of existence in $\mathcal{X}$, assuming Non-Comparativism). But even this weaker version of the Probable Addition Principle is incompatible with Non-Comparativism, provided we assume the Sure Thing Principle, which says that when comparing two prospects, we can ignore states in which the two prospects yield the same outcome. Or, more precisely:

> *Sure Thing Principle* If prospects $\mathcal{X}$ and $\mathcal{Y}$ yield the same outcome in state S_i, then $\mathcal{X}$ is at least as good for $\mathcal{S}$ as $\mathcal{Y}$ if and only if $\mathcal{X}$ is at least as good for $\mathcal{S}$ as $\mathcal{Y}$ conditional on the non-occurrence of state S_i.

To see that the Probable Addition Principle is incompatible with Non-Comparativism (assuming the Sure Thing Principle), let $a > b$ be wellbeing levels corresponding to lives worth living, and consider the following three prospects for person $\mathcal{S}$ over two equi-probable states of nature:

	State 1	State 2
$\mathcal{P}_1$	b	Ω
$\mathcal{P}_2$	a	b
$\mathcal{P}_3$	b	a

Intuitively, $\mathcal{P}_2$ and $\mathcal{P}_3$ are equally good, since they guarantee existence and give the same chances of the same wellbeing levels. The Probable Addition Principle implies that $\mathcal{P}_2$ is better than $\mathcal{P}_1$. Transitivity then implies

that $\mathcal{P}_3$ is better than $\mathcal{P}_1$. Applying the Sure Thing Principle to ignore State 1 for this last comparison, we find that certain existence at wellbeing level a is better for $\mathcal{S}$ than certain non-existence, which violates Non-Comparativism (given Certainty Equivalence).

Although the Sure Thing Principle is intuitively compelling, it is not completely obvious that Non-Comparativists should accept it.[7] Without it, however, the positive argument Nebel (2019: 340–341) provides for the Probable Addition Principle is not persuasive. This argument appeals to the following prospects for $\mathcal{S}$, where a is an excellent level of wellbeing, d is some small additional quantity of wellbeing, $-z$ is the level of a life barely worth not living, y is the level of a life barely worth living, and p and q are probabilities:

	State 1 $(1-p)(1-q)$	State 2 $(p(1-q))$	State 3 (q)
$\mathcal{A}$	a	a	Ω
$\mathcal{A}'$	$a+d$	$-z$	Ω
$\mathcal{A}^+$	$a+d$	$a+d$	y

Nebel notes that if we ignore State 3 (i.e., set $q = 0$), $\mathcal{A}'$ is intuitively better than $\mathcal{A}$ if p is sufficiently small. Since these prospects have the same outcome in State 3, he claims that $\mathcal{A}'$ must be better than $\mathcal{A}$ even if $q \neq 0$. Similarly, $\mathcal{A}^+$ is clearly better than $\mathcal{A}'$ if we ignore State 1; since these

[7] Consider prospect $\mathcal{P}_1$ from the earlier table. Clearly, $\mathcal{P}_1$ is equally as good as itself. Applying the Sure Thing Principle to ignore State 1, we can conclude that a prospect guaranteeing non-existence is equally as good as itself. This conclusion may be difficult for Non-Comparativists to accept.

prospects give the same result in State 1, it seems we can conclude that $\mathcal{A}^+$ is better than $\mathcal{A}'$. Transitivity then implies that $\mathcal{A}^+$ is better than $\mathcal{A}$. Since d, q and y were chosen arbitrarily, this is sufficient to establish the Probable Addition Principle.

In asking us to ignore states in which two prospects yield the same outcome, the preceding argument tacitly assumes the Sure Thing Principle. Are its claims still compelling if we do not assume the Sure Thing Principle? They are not. To see this, consider the following prospects:

	State 1 (0.01)	State 2 (0.00001)	State 3 (0.98999)
$\mathcal{A}'$	101	-1	Ω
$\mathcal{A}^+$	101	101	1

$\mathcal{A}^+$ is not *clearly* better than $\mathcal{A}'$, because it involves moving from a situation in which one is almost certain that one will enjoy an excellent life if one exists to a situation in which one is almost certain to only have a mediocre life if one exists. It's not crazy to think that this could make $\mathcal{A}^+$ worse than $\mathcal{A}'$.

The Probable Addition Principle is therefore a shaky foundation for the Intrapersonal Addition Paradox. However, it turns out that we can do without it. In the next section, I provide an intrapersonal argument for the Repugnant Conclusion which does not assume the Probable Addition Principle.

5.4 Repugnance Without Probable Addition

For this argument, we shall need all of the premises of the Intrapersonal Addition Paradox apart from the Probable Addition Principle, which we shall replace with a more compelling principle. To help state it, let me introduce some notation. Let us say that a prospect is *conditionally good (bad)* for $\mathcal{S}$ if it guarantees $\mathcal{S}$ a good (bad) life, provided she exists, and gives her a non-zero probability of existence. The principle we shall use in place of the Probable Addition Principle is the

> *Conditional Value Principle* If $\mathcal{X}$ is conditionally good for $\mathcal{S}$ and $\mathcal{Y}$ is conditionally bad for $\mathcal{S}$, then $\mathcal{X}$ is better for $\mathcal{S}$ than $\mathcal{Y}$.

Unlike the Probable Addition Principle, the Conditional Value Principle is intuitively compelling. It can also be supported by the following brief argument.[8] A conditionally good prospect is good for the person receiving it. A conditionally bad prospect, on the other hand, is bad for the person receiving it. If something is good for a person, and another thing is bad for the same person, the first thing is better for her than the second. The Conditional Value Principle follows from these three claims.

[8]This argument extends Parfit's notion of an outcome being non-comparatively "good for" a person, even when the alternative is non-existence, to the case of prospects (1984: 489). Parfit believed that this notion is compatible with Non-Comparativism. The argument is addressed to those who agree with Parfit on this matter.

We shall need to strengthen two of the other premises, but in ways which do not make much difference to their intuitive plausibility. First, we shall need to move to a stronger version of Minimal Prudence, which applies in cases in which a person is not guaranteed to exist. To state it, it will be helpful to introduce some more notation. If $w_1, w_2, \ldots, w_n$ are wellbeing levels and $p_1, \ldots p_n$ are probabilities adding up to 1, we write $(w_1[p_1], \ldots, w_n[p_n])$ to denote any prospect giving $\mathcal{S}$ w_1 with probability p_1, $\ldots$, and w_n with probability p_n. For this notation, we shall denote non-existence by Ω. We can now state

> *Prudence* Let $a \succ b$ and $c \succ d$ be any wellbeing levels. There exists some sufficiently small probability p' such that for any $p \leq p'$ and any probability of non-existence q,
>
> $$(a[p(1-q)], d[(1-p)(1-q)], \Omega[q])$$
>
> is better for $\mathcal{S}$ than
>
> $$(b[p(1-q)], c[(1-p)(1-q)], \Omega[q])$$
>
> provided $\mathcal{S}$ exists in exactly the same states of nature in each case.

Like Minimal Prudence, Prudence looks more complicated than it is. It just says that for a sufficiently small probability of getting a rather than b, it

would be better to instead get c rather than d with a much larger probability (proportionally speaking), no matter one's probability of non-existence. It is supported by the same sorts of considerations in favour of Minimal Prudence: decision theories which do not satisfy Prudence are reckless in the sense that they sometimes prioritise what happens in arbitrarily small probability states over what happens in states with proportionally much larger probabilities.

The second principle we shall need to strengthen is Stochastic Indifference for Equal Risk. We shall assume the slightly stronger

> *Statewise Dominance for Equal Risk* For any egalitarian prospects $\mathcal{X}$ and $\mathcal{Y}$ over the same states of nature, if each possible outcome of $\mathcal{X}$ is at least as good as its same-state counterpart in $\mathcal{Y}$, then $\mathcal{X}$ is at least as good as $\mathcal{Y}$. If, additionally, some outcome of $\mathcal{X}$ is better than the corresponding outcome of $\mathcal{Y}$, $\mathcal{X}$ is better than $\mathcal{Y}$.

Statewise Dominance for Equal Risk is supported by the same considerations which lend credence to Stochastic Indifference for Equal Risk: if a prospect has some chance of being better, and is certain to be at least as good, the prospect is better (at least when perfect ex post and ex ante equality is guaranteed).

We now have all the premises we need. Before beginning the argument proper, we first need to derive the

Absolute Value Principle[9] Let a be any wellbeing level corresponding to a life worth living, and b any level of a life worth not living. If population X consists solely of lives at a, and Y consists solely of lives at b, X is better than Y.

Let n and m be any numbers of people, a any good wellbeing level, and b any bad wellbeing level. Let $G_1, \ldots G_m$ be disjoint sets of n people each, and let $H_1, \ldots, H_n$ be disjoint sets of m people each, such that the G_i and the H_j sets contain the same $n \cdot m$ people. (That is, $\bigcup_{i=1}^{m} G_i = \bigcup_{i=1}^{n} H_i$.) Consider $n \cdot m$ equi-probable states of nature, which can be partitioned into events in two ways: (i) $S_1, \ldots, S_m$, where each S_i is the disjunction of n states of nature; (ii) $T_1, \ldots, T_n$, where each T_k is the disjunction of m states of nature. Now consider the following two prospects:

	S_1	S_2	$\ldots$	S_m
$\mathcal{A}$	$G_1[a]$	$G_2[a]$	$\ldots$	$G_m[a]$

	T_1	T_2	$\ldots$	T_n
$\mathcal{B}$	$H_1[b]$	$H_2[b]$	$\ldots$	$H_n[b]$

The Conditional Value Principle implies that $\mathcal{A}$ is better than $\mathcal{B}$ for each person who might exist; Weak Pareto for Equal Risk then implies that $\mathcal{A}$ is better than $\mathcal{B}$. Note that $\mathcal{A}$ guarantees the existence of n people at level

[9]This principle is often called "Priority for Lives Worth Living" in the economics literature (see for instance Blackorby et al., 2005: 135).

a, while $\mathcal{B}$ guarantees the existence of m people at level B. The Same-Number Quality Claim, Stochastic Indifference for Equal Risk and Certainty Equivalence therefore imply that an arbitrary population of n people at level a must be better than an arbitrary population of m people at level b; since a, b, n and m were chosen arbitrarily, we can conclude that the Absolute Value Principle is true.

We will not need the Conditional Value Principle from this point on: its only purpose was to provide an intrapersonal justification for the Absolute Value Principle. (That said, the Absolute Value Principle is intuitively compelling in its own right.)

We shall now derive the Repugnant Conclusion from our premises. Let A be any population of m lives at some excellent level of wellbeing a. Let $a^+ \succ a$ be a slightly higher level of wellbeing. Let $z \succ z^-$ be mediocre wellbeing levels, and let b be the level of a life barely worth not living. Naturally, we have $b \prec z^-$ and $a \succ z$.

Applying Prudence twice and taking the minimum of the two probabilities, we find that there is some small probability p' such that for any $p < p'$ and any probability of non-existence q:

(i)

$$(b[p(1-q)], a^+[(1-p)(1-q)], q[\Omega])$$

is better for $\mathcal{S}$ than

$$(a[1-q], \Omega[q])$$

(ii)

$$(b[p(1-q)], z[(1-p)(1-q)], \Omega[q])$$

is better for $\mathcal{S}$ than

$$(a^+[p(1-q)], z^-[(1-p)(1-q)], \Omega[q])$$

That is, a sufficiently small chance of getting b but larger chance of getting a^+ is better than getting a for sure (conditional on existence in the same states of nature), and a sufficiently small chance of getting b but larger chance of getting z is better than the same small chance of getting a^+ and larger chance of getting z^- (conditional on existence in the same states of nature). Now define n to be the smallest positive integer such that $1/n < p'/2$, and write $p = 1/n$.

We will consider prospects over states of nature $S_1, \ldots, S_{n^2}, T_1, \ldots, T_{n^2}$, where each S_i has probability $p^2(1-p)$ and each T_i has probability p^3. (These probabilities add up to 1.) We can construct disjoint sets $G_1, \ldots, G_{n^2}$, each containing m possible people, and where G_1 contains the A-people. Now consider the following two prospects:

	S_1	...	S_{n^2}	T_1	...	T_{n^2}
$\mathcal{R}_1$	$G_1[a]$	...	$G_1[a]$	$G_1[a]$	...	$G_1[a]$
$\mathcal{R}_2$	$G_1[a]$	...	$G_{n^2}[a]$	$G_1[a]$	...	$G_{n^2}[a]$

Since $\mathcal{R}_1$ and $\mathcal{R}_2$ are both egalitarian prospects, and both guarantee the existence of m people at wellbeing level a, the Same-Number Quality Claim and Statewise Dominance for Equal Risk imply that these prospects are equally good. Next, consider

	S_1	...	S_{n^2}	T_1	...	T_{n^2}
$\mathcal{R}_2$	$G_1[a]$	...	$G_{n^2}[a]$	$G_1[a]$	...	$G_{n^2}[a]$
$\mathcal{R}_3$	$G_1[a^+]$	...	$G_{n^2}[a^+]$	$G_1[b]$	...	$G_{n^2}[b]$

$\mathcal{R}_3$ is an egalitarian prospect, and each person exists in precisely the same states of nature in each prospect. Prudence therefore implies that $\mathcal{R}_3$ is better than $\mathcal{R}_2$ for every person who might exist, since each person is less than $2p$ times as likely to receive b as they are to receive a^+. Weak Pareto for Equal Risk then implies that $\mathcal{R}_3$ is better than $\mathcal{R}_2$. Next, we have

	S_1	...	S_{n^2}	T_1	...	T_{n^2}
$\mathcal{R}_3$	$G_1[a^+]$	...	$G_{n^2}[a^+]$	$G_1[b]$	...	$G_{n^2}[b]$
$\mathcal{R}_4$	$G_1[a^+]$	...	$G_{n^2}[a^+]$	$G[z^-]$	...	$G[z^-]$

Once again, $\mathcal{R}_4$ is egalitarian. The two prospects are equally good in states S_1 to S_{n^2} (because they yield the same outcomes), and the Absolute Value Principle implies that $\mathcal{R}_4$ is better than $\mathcal{R}_3$ in states T_1 to T_{n^2}. Statewise Dominance for Equal Risk therefore implies that $\mathcal{R}_4$ is better than $\mathcal{R}_3$. Next,

	S_1	...	S_{n^2}	T_1	...	T_{n^2}
$\mathcal{R}_4$	$G_1[a^+]$	...	$G_{n^2}[a^+]$	$G[z^-]$	...	$G[z^-]$
$\mathcal{R}_5$	$G_1[b]$	...	$G_{n^2}[b]$	$G[z]$	...	$G[z]$

$\mathcal{R}_5$ is egalitarian, and each person exists in precisely the same states of nature in each prospect. $\mathcal{R}_4$ gives each person probability $p^2(1-p)$ of getting a^+ and probability p of getting z^-. Thus, the probability of getting a^+ rather than b is less than $2p$ times as much as the probability of getting z^- rather than z. Prudence therefore implies that $\mathcal{R}_5$ is better than $\mathcal{R}_4$ for each person who might exist; Weak Pareto for Equal Risk then implies that $\mathcal{R}_5$ is better than $\mathcal{R}_4$. Finally, consider

	S_1	...	S_{n^2}	T_1	...	T_{n^2}
$\mathcal{R}_5$	$G_1[b]$	...	$G_{n^2}[b]$	$G[z]$	...	$G[z]$
$\mathcal{R}_6$	$G[z]$	...	$G[z]$	$G[z]$	...	$G[z]$

$\mathcal{R}_6$ is egalitarian. As in the case of $\mathcal{R}_4$ and $\mathcal{R}_3$, $\mathcal{R}_5$ and $\mathcal{R}_6$ have the same (and therefore equally good) outcomes in each state T_i, and the Absolute Value Principle implies that $\mathcal{R}_6$ is better than $\mathcal{R}_5$ in each state S_i. Statewise Dominance for Equal Risk therefore implies that $\mathcal{R}_6$ is better than $\mathcal{R}_5$. Putting all of our claims together with transitivity, we have that $\mathcal{R}_6$ is better than $\mathcal{R}_1$; Certainty Equivalence then implies that $G[z]$ is better than A, which is the Repugnant Conclusion.

The point of this argument is to show that intrapersonal arguments for the Repugnant Conclusion can go through without the Probable Addition Principle. Prudence is only slightly less compelling than Minimal Prudence,

while Statewise Dominance for Equal Risk is just as compelling as Stochastic Indifference for Equal Risk. If we want to block intrapersonal arguments for the Repugnant Conclusion, our only serious option is therefore to reject Weak Pareto for Equal Risk. To further substantiate this claim, I shall next show that with one very minimal further assumption, Weak Pareto for Equal Risk implies intuitively "repugnant" conclusions.

5.5 Pareto Principles and Repugnant Conclusions

Weak Pareto for Equal Risk implies nothing at all if we make no intrapersonal assumptions. But we only need one weak and compelling assumption in order for Weak Pareto for Equal Risk to make trouble for us. Our assumption shall be

> *Stochastic Dominance for Personal Prospects* Suppose that $\mathcal{S}$ exists in precisely the same states of nature in prospects $\mathcal{X}$ and $\mathcal{Y}$. If, for each wellbeing level w, $\mathcal{Z}$ gives $\mathcal{S}$ at least as great a probability of getting at least w as $\mathcal{Y}$ does, then $\mathcal{X}$ is better than $\mathcal{Y}$. If, additionally, there is some wellbeing level w for which $\mathcal{X}$ gives $\mathcal{S}$ a greater chance of getting at least w than $\mathcal{Y}$ does, then $\mathcal{X}$ is better for $\mathcal{S}$ than $\mathcal{Y}$.

Stochastic Dominance for Personal Prospects just says that if $\mathcal{S}$ has the

same chances of getting each wellbeing level between two prospects, these prospects are equally good for her; if she has better chances of being better off in one prospect, then that prospect is better for her. Because it only applies when $\mathcal{S}$ exists in the same states of nature in both prospects, it is compatible with hefty Non-Comparativist restrictions on the ranking of prospects involving risks of non-existence.[10]

To see how Stochastic Dominance for Personal Prospects spells trouble when conjoined with Weak Pareto for Equal Risk, consider the following two prospects over equi-probable states of nature S_1 to S_{n+1}, where G_1 to G_n are disjoint sets of m people each (where m is any number), G is the set of all people in some G_i, a is an excellent wellbeing level, and $z^+ \succ z$ are mediocre wellbeing levels:

	S_1	...	S_n	S_{n+1}
$\mathcal{M}_1$	$G_1[a]$	...	$G_n[a]$	$G[z]$
$\mathcal{M}_2$	$G_1[z^+]$	...	$G_n[z^+]$	$G[a]$

Each person in G exists in the same states of nature in $\mathcal{M}_1$ and $\mathcal{M}_2$. $\mathcal{M}_2$ gives each person a better chance of getting at least z^+, and at least as good a chance of getting any other wellbeing level. Stochastic Dominance for Personal Prospects therefore implies that $\mathcal{M}_2$ is better for each person than $\mathcal{M}_1$. Since both prospects are egalitarian, Weak Pareto for Equal Risk

[10] In particular, it is compatible with Non-Comparativism, plus "deference" type responses to Hare's (2010) Opaque Sweetening Problem. (Deference type responses are endorsed by Bales et al. 2014 and Schoenfield 2014.)

then implies that $\mathcal{M}_2$ is better than $\mathcal{M}_1$.

This is pretty "repugnant". The G_i sets could each have at least ten billion people, and n could be arbitrarily large. Under these conditions, to say that $\mathcal{M}_2$ is better than $\mathcal{M}_1$ is to say that a near-certainty of having ten billion people with excellent lives (otherwise a very large number of lives barely worth living), is worse than the same near-certainty of there being an enormous number of people with lives barely worth living (otherwise a very large number of excellent lives). Apart from being intuitively "repugnant", this conclusion is hard to square with a rejection of the Repugnant Conclusion proper. If we believe that the Repugnant Conclusion is false, we believe that the mere fact that an outcome contains *very many* people is not enough to make that outcome *very good*. Yet in the case of $\mathcal{M}_1$ and $\mathcal{M}_2$, we have to believe that the tiny chance of getting a better outcome in state S_{n+1} is more important than the much larger chance of getting a better outcome in all other states of nature. The only important difference between the comparison between $\mathcal{M}_1$ and $\mathcal{M}_2$ in state S_{n+1} and the comparisons in the other states is that in S_{n+1}, vastly more people exist.[11] The claim that this can make all the difference is precisely the claim we deny when we reject the Repugnant Conclusion.

[11] Another difference is that the mediocre wellbeing levels in S_1 to S_n are z^+, rather than the slightly worse level z in S_{n+1}. But this difference is not important enough to do the work required. (The mediocre wellbeing levels in S_1 to S_n are only set at z^+ in order for us to be able to say that $\mathcal{M}_2$ is *better* than $\mathcal{M}_1$, rather than the two merely being equally good.)

5.6 Conclusion

Nebel's Intrapersonal Addition Paradox brings yet more trouble for those who wish to avoid the Repugnant Conclusion. In §5.2, we saw that provided one accepts the transitivity of the prudential and moral at-least-as-good-as relations, the Repugnant Conclusion can only be robustly avoided by denying the Probable Addition Principle or Weak Pareto for Equal Risk. Because the Probable Addition Principle is the less compelling of the two principles by a significant margin, denying Probable Addition seemed to be the best bet. However, we found in §5.3 that Probable Addition can be replaced by the much more plausible Conditional Value Principle. The argument demonstrating this required slightly stronger auxiliary assumptions than the Intrapersonal Addition Paradox, but these stronger assumptions are not significantly less plausible than the originals.

Since denying the Probable Addition Principle is not enough to block the Repugnant Conclusion, the only remaining option is to deny Weak Pareto for Equal Risk. This was confirmed in §5.5, where we saw that an intuitively "repugnant" conclusion can be derived from Weak Pareto for Equal Risk, together with the compelling principle of Stochastic Dominance for Personal Prospects. Notably, unlike most arguments for versions of the Repugnant Conclusion, this argument did not assume transitivity or any similar principle.

Given that we must choose between avoiding the Repugnant Conclusion and accepting Weak Pareto for Equal Risk, which option is best? I am inclined to favour the second option, mostly because accepting the Repugnant Conclusion provides a response to interpersonal and intrapersonal arguments with a single stroke, whereas if we deny Weak Pareto for Equal Risk, we are still left with the considerable task of responding to interpersonal arguments which do not appeal to Pareto principles.[12]

That said, denying Weak Pareto for Equal Risk remains a serious option. Although it is a deeply compelling principle, it is false on some proposed axiologies, notably on standard ex post prioritarianism.[13] We might deny Weak Pareto for Equal Risk on the basis that ex ante Pareto principles mistakenly require moral and prudential importance to march in lock-step. We might instead think that moral and prudential importance sometimes come apart, either because some things are morally but not prudentially valuable, or because the extent to which lives are morally valuable can depend on the number and nature of the other lives that exist. For instance, we might believe that it is more important for there to be ten billion rather than zero people with excellent lives than it is for there to be twenty billion rather than ten billion people with excellent lives. If so, we sometimes value the existence of excellent lives in a way that is not entirely explained by the prudential

[12]The most concerning interpersonal arguments seem to me those given by Arrhenius (2000, 2003, 2011) and by Spears and Budolfson (2021).

[13]See Ord 2015: 301.

value of these lives for the people who enjoy them.

In his closing remarks, Nebel suggests something similar:

> We care very strongly about the existence of the things in wonderful lives—things like loving relationships, creative activities, and sophisticated pleasures. But perhaps we do not value these things—primarily, at least—because they are good for the people whose lives contain them. Perhaps we value these things primarily as impersonal goods.
>
> Nebel 2019: 342

It seems to me that this is the right lesson to draw from the various intrapersonal arguments discussed in this chapter. If the existence of ten billion excellent lives would be better than the existence of any number of lives barely worth living, then excellent lives must have impersonal as well as personal value.

Chapter 6

Prudence in Different-Number Fission Cases

Abstract

This chapter argues for *Fission Totalism*, which is a version of Totalism applicable to prudence in fission cases: cases where a fission parent "splits" into two or more fission offspring, while retaining what matters in survival. Sections 6.3 and 6.4 jointly provide an argument for Fission Totalism. §6.3 argues for the Neutral Addition Principle, which says that splitting into additional fission offspring with neutral lives is prudentially neutral. §6.4 uses an analogue of Harsanyi's (1955) Aggregation Theorem to derive the restriction of Fission Totalism restricted to cases involving the same number of people. These two principles together imply Fission Totalism, where the total amount of wellbeing is measured on the scale generated by Expected Utility Theory: that is, where the scale is chosen such that by definition, it is best, when facing a case of risk, to maximise the expectation of this function. Next, §6.5 provides an arugment for Fission Totalism on a conceptually distinct scale of wellbeing: the life-years scale, where amounts of wellbeing are directly proportional to years of good life. Since these arguments are both sound only if the two conceptually distinct scales of wellbeing are co-extensive, §6.6 provides a direct argument for the coincidence of these two scales of wellbeing.

6.1 Introduction

Usually, we can survive into the future as at most one person. If I survive through tonight, then tomorrow there will be one of me; if I die tonight, then tomorrow there will be none of me. These are the only two options in ordinary cases. When we think about what it would take for a choice to be prudent, we usually focus on ordinary cases like these. That is only natural: ordinary cases are the only ones we actually face. But perhaps other sorts of cases are also possible. In these "fission cases", we can, roughly speaking, survive into the future as multiple people: one individual, the fission parent, splits into multiple fission offspring.

Derek Parfit (1984: 253–261) argued that fission cases are possible in humans, at least in principle. Human brains can be separated into two halves, each of which by itself seems to be sufficient for ordinary survival. So, if we imagine slitting a human brain into its separate halves, and then successfully transplanting these halves into two separate bodies, the apparent result would be that two people would wake up, both would think themselves to be the person who owned the original brain, and each *would* have been the owner of the original brain, if their competitor had not existed.

Even if fission cases like these are *not* possible in humans, we can imagine creatures for whom fission cases are perfectly ordinary things to face. Such creatures might reproduce by splitting their brains and bodies into pieces,

each of which then regrows the missing parts, just as many plants are able to reproduce when we take cuttings. There seems to be nothing which rules out that creatures like these could also be persons. If fission-prone persons are metaphysically possible, as they seem to be, then fission cases are metaphysically possible, and we can ask what prudence demands of agents facing these cases.

This chapter aims to provide part of the answer to that question. More specifically, it is about what prudence requires of us when, depending on what we do, we will have different numbers of fission offspring. Is it better, for instance, to split into three fission offspring with good lives than to split into two fission offspring with equally good lives? (I shall argue: Yes.) Could it be better to split into three fission offspring with good but slightly worse lives, rather than splitting into two fission offspring (again, I shall argue: Yes.) More generally, we can ask for a general theory of the comparison of *fission populations*: sets of fission offspring with associated wellbeing levels. Call this the

> *Fission Question* Under what conditions is one fission population prudentially better than another, when the two fission populations may contain different numbers of people?[1]

[1]By one fission population being "prudentially better" than another, I mean that one would have more prudential reason – the sort of reason we usually have to do things for our own sake – to make it the case that one would undergo a fission event which results in the first fission population, rather than one which results in the second fission population.

Two obvious answers to the Fission Question are

Fission Averagism: Fission population X is prudentially better than fission population Y if and only if X has greater average wellbeing than Y.

Fission Totalism: Fission population X is prudentially better than fission population Y if and only if X has greater total wellbeing than Y.

(I'll say more about what I mean by total/average wellbeing in the next section.)

My task in this chapter shall be to provide an answer to the Fission Question. In particular, I shall argue that Fission Totalism is correct.

6.2 Preliminaries

6.2.1 The Identity Requirement

The Fission Question is only going to have an interesting answer if we have *prudential concern* for our fission offspring, by which I mean that the well-being of our fission offspring is a source of prudential reasons for us, in the same way that our future wellbeing is a source of prudential reasons. This might be doubted if we accept the antecedently compelling

> *Identity Requirement* Person-stage p_i has prudential concern for person-stage p_j only if p_i and p_j are personally identical.

The Identity Requirement rules out that we have prudential concern for our fission offspring if the fission parent is a different person to each of the fission offspring. The standard argument from the Identity Requirement to the lack of prudential concern of a fission parent for her fission offspring goes as follows. Consider a simple fission case, in which a fission parent, "Wholly", splits into two fission offspring, "Lefty" and "Righty". Since Lefty and Righty are related to Wholly in the same way in all relevant respects, Wholly is the same person as Lefty if and only if Wholly is the same person as Righty. Suppose, for contradiction, that Wholly *is* the same person as Lefty. In that case, Wholly is also the same person as Righty, and then by transitivity, Lefty and Righty must also be the same person. But this is false: Lefty and Righty are different people. Hence Wholly is *not* the same person as Lefty, and it follows from this that Wholly is not the same person as Righty, either.

However, the claim that fission parents have no prudential concern for their fission offspring is difficult to accept. Consider a case in which both halves of one's brain are successfully transplanted into new bodies to a case in which only one half is successfully transplanted, and the other is destroyed. Parfit writes about this case:

> I would survive if my brain was successfully transplanted. And people have in fact survived with half their brains destroyed. Given these facts, it seems clear that I would survive if half my

> brain was successfully transplanted, and the other half was destroyed. So how could I fail to survive if the other half was also successfully transplanted? How could a double success be a failure?
>
> Parfit (1984: 255)

Parfit thus concludes not that fission parents should have no prudential concern for their fission offspring because the Identity Requirement is true, but that the Identity Requirement is false because fission parents should have prudential concern for their fission offspring. This leaves us with

> *No Identity Requirement* Person-stage p_i can have prudential concern for person-stage p_j, even if p_i and p_j are not personally identical.

Additionally, by symmetry, fission parents must have prudential concern for *all* their fission offspring. And since fission-into-Lefty seems not to be worse in any respect than destruction of the right side of the brain followed by ordinary survival, it seems we should have *full* prudential concern for our fission offspring, in the sense that the fact that we are not personally identical with our fission offspring does not, by itself, give us any less prudential reason to care about their wellbeing.

Let's put all these claims together under a simple heading: Parfit's slogan that *Identity Does Not Matter*. While I have briefly sketched a standard argument for this position, the main point of this chapter is not to argue that

Identity Does Not Matter. It is to answer the following question: *provided* that Identity Does Not Matter, what does prudence require when choosing between fission cases, if different numbers of fission offspring will result, depending on what we do?

I shall also assume throughout this chapter that fission offspring do not bear the relation of prudential concern, in any degree, to each other. (Note that if fission offspring *did* have full prudential concern for each other, then the "aggregate life version" of Fission Totalism I shall discuss in §6.5.5 would seem to follow more or less automatically.)

6.2.2 'All Else Equal'

The Fission Question calls for a ranking of all possible fission operations, including ones which result in different numbers of fission offspring, in terms of how prudentially good they are. For the most part, I shall consider only "simple fission operations" in which a fission parent has a choice to split into varying numbers of fission offspring at various wellbeing levels, has full prudential concern for all of these fission offspring, and in which there are no further fission or fusion operations down the line. (Fusion operations, in which multiple individuals merge into one, will be important in §6.7.1, but not before then.) I shall generally rank the results of fission operations, which I call "fission populations", by looking only at which fission offspring

exist in them, and how well off these fission offspring are. If it is possible to make all-else-equal comparisons of fission populations in terms of prudential value, then we can understand $\succeq$ to be exactly this *ceteris paribus* prudential at-least-as-good relation.

This assumption might be questioned, for two reasons. First, it is at least a logical possibility that that how two fission populations compare depends on the past history of the fission parent. For example, it could be that if you have lived a drab life thus far, then it would be better to split into two fission offspring who would have long futures at a decent but not amazing quality, while if you have lived an amazing life thus far, it would be better for you to split into two fission offspring with shorter but equally amazing futures. (I use "quality" to refer to how well a life goes at each time, independently of how long the life lasts.)

Second, and relatedly, it might be that the contributive prudential value of a fission offspring's life to the fission parent is not fully determined by the wellbeing level of the fission offspring in question. That is, one might deny

> *Prudential Ex Post Pareto* If, no matter how things turn out (i.e., in every state of nature) fission operations O_1 and O_2 would produce exactly the same fission offspring at exactly the same wellbeing levels, then O_1 and O_2 are equally good for the fission parent.

For example, a mediocre life which lasts for a hundred years might be

equally as good as an exciting life which lasts for ten years. But you might perhaps think that it would be better to split into one hundred fission offspring with the second sort of life than it would be to split into one hundred fission offspring with the first sort of life.

For the purposes of this chapter, I shall simply assume that despite worries like this which one might have, we can indeed make all-else-equal comparisons of fission populations, and the principle of Prudential Ex Post Pareto is true. But it would be misleading not to admit clearly that this is a *substantive* assumption made in order to simplify a complex topic. It is not a harmless stipulation, nor is it an unimpeachable axiom.

6.2.3 Incomplete Holistic Goods

I also need to make an important restriction on the sorts of goods which will be available for fission offspring in the cases considered in this chapter. This is that the lives of the fission parent or fission offspring should never contain any incomplete *holistic goods*. By holistic goods, I mean composite goods (or bads) which contribute to lifetime wellbeing in a way which is not explicable in terms of their component parts. *Incomplete* holistic goods do not contain all of their component parts in the life of a single fission offspring or fission parent. An (admittedly tired) example should make this plain. Suppose it would be good for me to write a paper, which would take two months. I might

be better off relaxing in the first month, rather than writing. I might also be better off relaxing in the second month, rather than writing. But I might yet be better off in terms of lifetime wellbeing if I write both months and finish my paper, rather than relaxing both months and being embarrassed when my more productive colleagues ask me what I've been working on.

If we allowed for incomplete holistic goods, then principles like Prudential Ex Post Pareto Indifference would probably fail. Suppose I will split into two fission offspring, Lefty and Righty. I can either choose to ensure that both relax next month, or instead ensure that each will write one half of my planned paper next month. Imagine that while being the sole author of a paper is very good for you, being one of two co-authors is barely worth mentioning. In that case, if I force Lefty and Righty to write, I will plausibly be better off, because getting my fission offspring to finish a co-authored paper looks, from my perspective, just like finishing a sole-authored paper myself. On the other hand, for both Lefty and Righty, co-authorship isn't worth the effort, and each would be better off relaxing instead. So the option that would be better for me is the one that would be worse for both of my fission offspring.

Be that as it may, I shall focus my attention on cases where there are no incomplete holistic goods within lives, so that the contributive prudential value of a fission offspring's life does not depend on whether the fission parent, or other fission offspring, are around to complete otherwise unfinished

projects.

6.2.4 Wellbeing Scales

Speaking of wellbeing, we shall need a way of denoting the wellbeing levels experienced by fission offspring. I shall generally use numbers to denote these wellbeing levels, and I shall do it in different ways in different parts of the chapter. In all cases, I shall assign the level 0 to the neutral level of wellbeing: the level at which one would, for one's own sake, be indifferent between ordinary survival at that wellbeing level and immediate ordinary death.[2] For now we shall assume that there is a wellbeing level which is neutral regardless of one's prior life, i.e., one would be indifferent between ordinary survival at this level and immediate death, regardless of what one's previous life was like. But we shall see in §6.6.4 that, given the assumption of Time-Separability defended in §6.6, it is enough to assume the weaker

[2]I'm assuming that *there is* such a level of wellbeing. In particular, I am assuming that there is a wellbeing level which is *exactly* equally as good as ordinary death. There are clearly "good" levels of wellbeing such that you'd have an interest in survival at these levels, and "bad" levels of wellbeing such that you'd have an interest in receiving ordinary death rather than survival at these levels. So it might seem obvious that in between, there must be a "neutral" level of the sort I'd like. But actually this is a little too quick: it's possible that instead there is a range of wellbeing levels in between, each of which is neither better, nor worse, nor equally as good as ordinary survival. For a similar view in the context of population axiology, see Gustafsson (2020).

While such views are logically possible, they should be rejected in the case of prudence. As the length of time for which one ordinarily survives at some constant level of quality approaches zero, it seems clear that the wellbeing level associated with survival for that length of time must get arbitrarily close to the level associated with ordinary death; and it would be very strange if there were wellbeing levels which got arbitrarily close to prudential neutrality, but no completely neutral wellbeing level.

claim that there is a wellbeing level which is neutral for *some* prior life; Time-Separability implies that this wellbeing is neutral for *all* prior lives.

To attach numbers to wellbeing levels uniquely (up to positive scalar multiplication) I shall use two methods, which produce two different scales. The first is the life-years scale, on which x units represents x years of good life at some constant good quality, while $-x$ represents the level at which one would be indifferent between ordinary death and 50% chance of x, 50% chance of $-x$. If there are any wellbeing levels which cannot be represented in this way, perhaps because they are better than lives of any length at constant quality q, then I shall restricted my attention to the set of wellbeing levels which *can* be so represented. I shall also assume that this representation function is monotonic. That is, lives which get a higher life-years wellbeing level are genuinely better than those which get lower levels: it's better to live longer at a good quality of life than to live for a shorter amount of time, all else equal.

The second is the scale generated by Expected Utility Theory, or the EUT scale for short. To generate this scale, we need a substantive assumption, namely that the prudential betterness relation over *prospects* – probability distributions over outcomes – in ordinary cases satisfies the axioms of Expected Utility Theory.[3] It need not concern us here exactly what these

[3]Here, and henceforth, I use "ordinary cases" in a technical sense to mean cases in which there is no fission or fusion at issue.

axioms are: what's important for us is that provided the prudential betterness relation satisfies them, which is quite plausible, the central theorem of Expected Utility Theory implies that this relation can be represented by a real-valued utility function in the following sense: the utility function assigns real numbers to outcomes in such a way that prospect X is prudentially at least as good as prospect Y if and only if the expected value of X under this utility function is at least as great as the expected value of Y under the same function.[4] This utility function is unique up to positive affine transformation (addition followed by positive scaling), so the utility function mapping 0 to the neutral level in the way defined above (that is, the EUT scale) is unique up to positive scaling.

So, we have two wellbeing scales we can talk about: the life-years scale and the EUT scale. And I shall discuss two main arguments for Fission Totalism. The first argument, which I give in §6.4, is for Fission Totalism on the EUT scale, while the second argument, given in 6.5, is for Fission Totalism on the life-years scale. Obviously, it cannot be that both arguments are sound unless the two scales coincide up to positive scaling; otherwise the two versions of Fission Totalism would yield contradictory judgements. As it happens, I think the two scales do coincide, and I shall argue for this view in 6.3.3.

[4]See Morgenstern and Von Neumann (1944), Herstein and Milnor (1953), Savage (1954) or Fishburn (1982) on the central theorem of Expected Utility Theory.

6.2.5 Fission Populations and Prospects

We shall sometimes need to distinguish between different fission offspring. We shall therefore need a way of identifying fission offspring as being *the same individual* across outcomes and states of nature. Formally, we might as well identify possible fission offspring with natural numbers, giving us an unlimited supply of them. A fission population contains information about which fission offspring come into existence, and how well-off each one is. Formally, fission populations can be thought of as sets of ordered pairs (n_i, x_i), where the n_i are all positive integers and the x_i are all real numbers. But I shall generally represent them as tables, and I shall give the fission offspring names rather than numbers. Another bit of terminology: we shall say that fission populations X and Y are *disjoint* if and only if they have no fission offspring in common, and whenever X and Y are disjoint, we may write $X + Y$ to denote the fission population consisting of the X-offspring at their respective levels, the Y-offspring at their respective levels, and nobody else. (Formally, $X + Y = X \cup Y$.)

I shall assume that $\succeq$ is a reflexive and transitive relation.[5] I shall not rule out that $\succeq$ is incomplete in the sense that perhaps neither $X \succeq Y$ nor $Y \succeq X$. But I shall not much discuss this possibility either.[6]

[5] $\succeq$ is reflexive if and only if, for any fission population X, we have $X \succeq X$. $\succeq$ is transitive if and only if, for any fission populations X, Y and Z, if $X \succeq Y$ and $Y \succeq Z$ then $X \succeq Z$.

[6] That said, I shall later assume that $\succeq$ is complete in ordinary cases, at least for a

In addition to there being a prudential betterness relation over fission populations, I shall assume that there is also a prudential betterness relation over fission population *prospects*. By fission population prospects, I mean functions which take *states of nature* with associated probabilities adding up to 1, and assign them fission populations.[7] Put less formally, a fission population prospect just takes a set of ways the world could be – perhaps the possible rolls of a die, or possible results of a sequence of coin flips – and spits out the fission populations which will result if the world turns out this way, given that this prospect is chosen. I shall use the same symbol, $\succeq$, for this relation; in any case, the $\succeq$ relation on fission populations can be thought of as a special case of the $\succeq$ relation on fission population prospects: the case where one fission population is certain to come about and all others are certain not to come about.[8]

As an aside, much of this framework will look familiar to those who are familiar with population axiology, the study of what we might call the "moral"

restricted range of wellbeing levels.

[7]By *states of nature*, I mean ways the world could be (or, if you like, propositions about the world) which are statistically independent of one's actions and which, together with the prospect chosen, fully determine the fission population which will result. For example, if one is making a choice between fission population prospects which yield different fission populations depending on the result of a die roll, and the results of the die roll are statistically independent of which fission population prospect is chosen, then the relevant states of nature will be all the possible results of the die roll.

[8]This is a little underspecified: for any fission populations X and Y, there could be many formally distinct fission populations prospects yielding certainty of X and Y, namely ones over different sets of states of nature. I assume this doesn't matter: if X', X'' and Y', Y'' are certain to result in fission populations X and Y respectively, then $X' \succeq Y'$ iff $X'' \succeq Y''$ iff $X \succeq Y$.

betterness relation on populations, or on population prospects. Indeed, the two subject matters are exactly structurally analogous: "fission populations" in the axiology of prudence in fission cases act just like "populations" in population axiology, while the prudential at-least-as-good-as relation and the moral at-least-as-good-as relation seem to have exactly the same structural features: reflexivity, transitivity and option-set-independence.[9] This will be important later because it means that many arguments in population axiology, including complex theorems, can be reinterpreted so that they continue to hold for the axiology of prudence in different-number fission cases.

If X is a fission population, we can write $To(X)$ for the total wellbeing in X, $Av(X)$ for the average wellbeing in X, and $|X|$ for the size of X: the number of fission offspring contained in X. Naturally, for any fission population X, we have $Av(X) = \frac{To(X)}{|X|}$.

6.2.6 Fission Totalism and Averagism

The total and average views mentioned in §6.1 can now be made more precise:

Fission Totalism For any fission populations X and Y, $X \succeq Y$ if and

[9]The relation $\succeq$ is option-set-independent if and only if, for any two populations X and Y and option sets $\mathcal{O}$ and $\mathcal{O}'$ containing X and Y, we have $X \succeq Y$ relative to $\mathcal{O}$ if and only if $X \succeq Y$ relative to $\mathcal{O}'$. Strictly speaking, since the prudential at-least-as-good-as relation $\succeq$, as we have specified it, is a binary relation on fission populations, this condition is not even well-formed; speaking more loosely, we can say that since $\succeq$ is a binary relation on fission populations, whether $X \succeq Y$ holds does not change depending on which set of alternatives is under consideration, and therefore $\succeq$ is automatically option-set-independent.

only if $To(X) \geq To(Y)$.

Fission Averagism For any fission populations X and Y, $X \succeq Y$ if and only if $Av(X) \geq Av(Y)$.

Fission Totalism and Fission Averagism have in common a simple view on how to weigh up the wellbeing of fission offspring in cases where the number of fission offspring is fixed. We can call this view

Same-Number Fission Totalism For any fission populations X and Y containing the same number of fission offspring, $X \succeq Y$ if and only if $To(X) \geq To(Y)$.

Fission Totalism and Fission Averagism disagree, however, on the matter of adding fission offspring to a fission population. According to Fission Totalism, the valence of adding a fission offspring always matches the valence of the life for that fission offspring. More formally, Fission Totalists accept

Prudential Neutral Addition If X and Y are disjoint fission populations and Y contains only neutral lives, then $X + Y \sim X$.

Fission Totalism follows from Same-Number Fission Totalism and Prudential Neutral Addition.[10] Same-Number Fission Totalism can thus be

[10]Let X and Y be arbitrary non-empty fission populations, and let x and y be individuals in X and Y respectively. Let X' be the fission population consisting of just x at wellbeing

thought of as capturing the "same-number part" of Fission Totalism, while Prudential Neutral Addition captures the "different-number part" of Fission Totalism.

Fission Totalists and Fission Averagists both accept Same-Number Fission Totalism: they agree on same-number cases. But they disagree on Prudential Neutral Addition: they disagree on different-number cases.

The aim of this chapter is to argue for Fission Totalism. As we have just seen, Fission Totalism can be thought of as having two components: Same-Number Fission Totalism and Prudential Neutral Addition. In the next section, I shall argue for Prudential Neutral Addition. Afterwards, in §6.4, I shall argue for Same-Number Fission Totalism on the EUT scale of wellbeing, thereby completing the argument for the EUT version of Fission Totalism. In §6.5, I shall give a different argument for Fission Totalism on the life-years scale of wellbeing. Unless the life-years and EUT scales of wellbeing coincide, it cannot be that both arguments are sound. However, I shall argue in §6.6 that the two scales *do* coincide. If this claim is correct, then we only need *one* of the arguments of §6.3-§6.4 and §6.5 to succeed in order to get Fission Totalism on *both* scales of wellbeing. Lastly, I shall

level $To(X)$; let Y' consist of y at level $To(Y)$. Let 0_X contain everyone in X, except x, at the neutral level of wellbeing 0, and let O_Y contain everyone in Y, except y, also at level 0. Same-Number Fission Totalism implies $X \sim X' + 0_X$ and $Y \sim Y' + 0_Y$. The neutral addition principle implies $X' + 0_X = X'$ and $Y' + 0_Y = Y'$. Transitivity then implies that $X \sim X'$ and $Y \sim Y'$, and that therefore $X \succeq Y$ if and only if $X' \succeq Y'$. Same-Number Fission Totalism now implies that $X \succeq Y$ if and only if $To(X) \succeq To(Y)$, which is the statement of Fission Totalism.

consider two objections to Fission Totalism in §6.7.

6.3 Arguments for Neutral Addition

6.3.1 Positive and Negative Addition

Prudential Neutral Addition is supported by the following two principles:

> *Prudential Positive Addition* If X and Y are disjoint fission populations and everyone in Y has a positive wellbeing level, then $X+Y \succ X$.

> *Prudential Negative Addition* If X and Y are disjoint fission populations and everyone in Y has a negative wellbeing level, then $X+Y \succ X$.

Given these two principles, if a fission offspring is at a slightly positive level, no matter how small, adding them to a fission population always makes that population better. If the fission offspring is at a slightly negative level, no matter how small, adding them is always worse. Any wellbeing level even slightly lower than the neutral level is negative, and any wellbeing level even slightly higher than the neutral level is positive. Therefore, if a fission offspring is at a neutral level, then adding them at a slightly worse level is worse, and adding them at a slightly better level is better. This is the mark of equal goodness: if two fission populations are such that slight improvements

to one result it in being better, while slight worsenings result in it being worse, then the two fission populations must be exactly equally good.[11]

Conversely, if Prudential Neutral Addition is true, then Prudential Positive and Negative Addition must be true as well: if adding a fission offspring at a neutral level results in a fission population that is equally good, and moving the fission offspring from the neutral level to a positive level makes things prudentially better (as it clearly does), then adding a fission offspring at a positive wellbeing level makes things better. Changing what needs to be changed, the same is true of adding fission offspring with negative wellbeing levels.

The conjunction of Prudential Positive and Negative Addition on the one hand, and Prudential Neutral Addition on the other hand, are thus practically equivalent to each other. They are not *logically* equivalent (unless we stack the deck by defining strict betterness *à la Broome*), but any sensible theory of prudence will satisfy Prudential Neutral Addition if and only if it satisfies both Prudential Positive and Negative Addition.

Fission Averagism violates both Prudential Positive Addition and Prudential Negative Addition. The latter violation is intuitively more troubling than the former. Fission Averagism implies that we should be indifferent between having two fission offspring at wellbeing level 100, and having three

[11]This point is not limited to equal goodness as it pertains to fission populations: it applies to equal goodness in general. See for instance Broome (2004: 21).

fission offspring at wellbeing level 100. At first sight, there is nothing particularly troubling about this claim. Intuitively, fission into two people looks like "ordinary survival as one of two people, each of whom are at wellbeing level 100", while fission into three people looks like "ordinary survival as one of three people, each of whom are at wellbeing level 100"; on the face of it, both look equally as good as ordinary survival at level 100, and hence equally as good as each other.

In contrast, common-sense intuitions seem to cut against Fission Averagism when it comes to cases involving negative wellbeing. Suppose that you could either undergo fission into two tortured lives, or the same two tortured lives plus one slightly less tortured life. Having an additional tortured life as a fission offspring doesn't seem to do anything at all to make the second scenario more appealing; intuitively, fission-into-two is prudentially at least as good as fission-into-two-plus-one.[12] But fission into three fully tortured people is clearly worse than fission-into-two-plus-one, and hence transitivity implies fission-into-three is worse than fission-into-two.

Notice the similarity to *non-sadism* principles in population axiology. Averagism violates the Mere Addition Principle, but most people find it worse that it says that adding negative lives to a population can make things better. Similarly, Fission Averagism violates Prudential Positive Addition, but most people will find it worse that it says that adding tortured fission

[12] A similar claim is made by Ross (2014: 227, 257)

offspring can make things prudentially better.

We thus have reason to accept Prudential Negative Addition, but Prudential Negative Addition alone does not imply Prudential Neutral Addition. We need other arguments to support Prudential Positive Addition, and thereby support Prudential Neutral Addition. We can take one such argument directly from population axiology. The "Mere Addition Principle", on which adding good lives to a population results in an outcome which is at least as good, can be supported by the more intuitively compelling "Dominance Addition Principle", on which adding good lives to a population while at the same time making every existing person better off results in an outcome which is at least as good. Similarly, Prudential Positive Addition can be supported by the more compelling principle of

> *Prudential Dominance Addition* If X and Y are disjoint fission populations, everyone in Y has a positive wellbeing level, and X^+ is a fission population containing the same fission offspring as X, each of whom is better off in X^+, then $X^+ + Y \succ X$.

In fact, given Same-Number Fission Totalism, Prudential Dominance Addition straightforwardly implies Prudential Positive Addition.[13] And Pru-

[13]Let X and Y be disjoint fission populations, where Y consists of good lives. Let X'_Y be the fission population containing the X and Y fission offspring, in which the Y people have half the levels of wellbeing they have in Y, and this excess wellbeing is distributed evenly to all of the X people, giving them slightly higher wellbeing than they have in X. Same-Number Fission Totalism implies that $X'_Y \sim X + Y$, and Prudential Dominance Addition implies that $X'_Y \succ X$; transitivity then implies that $X + Y \succ X$, as required for Prudential Positive Addition.

dential Dominance Addition is significantly more compelling than Prudential Mere Addition, because in order for it to fail, adding another fission offspring, who perhaps is less well-off than one's other fission offspring, would need to be able to somehow cancel out the added prudential value of making one's other fission offspring better off. And it's hard to see how that could be the case.

At the root of this intuition, I think, is Parfit's rhetorical question: how could an additional success – an additional fission offspring with a good life – constitute a failure? This is a powerful intuition, and I think it provides reasonable grounds to accept Prudential Dominance Addition (and, consequently, Prudential Positive Addition).

6.3.2 Neutral Addition and Separability

In recent work, Gustafsson and Kosonen (forthcoming) have provided another argument for the sorts of Addition principles we have considered. They write:

> Since the relation [of prudential concern] is plausibly intrinsic, the value of standing in that relation to a future person should not be diminished if you also stand in [the relation of prudential concern] to some other person. Gustafsson and Kosonen (forthcoming: 3)

This argument is intended to support only Prudential Positive Addition

but, more broadly, I think that if this argument succeeds, it supports the claim that the value of standing in the relation of prudential concern should not only be *not diminished*, but should be entirely unchanged by whether or not you also stand in this relation to some other person. Their argument, if successful, thus seems to support the principle of

> *Fission Population Separability* Let X and Y be any fission populations, and let Z be any fission population disjoint from X and Y. Then $X \succeq Y$ if and only if $X + Z \succeq Y + Z$.

Fission Population Separability (henceforth just "Separability" wherever this is unambiguous) is supported because, if the value of standing in relations of prudential concern to the X-people and the Y-people does not change depending on whether or not one also stands in the relation of prudential concern to the Z-people, then the comparison of X and Y (i.e. the case where one does not have prudential concern for the Z-people) should turn out the same way as the comparison of $X + Z$ and $Y + Z$ (i.e. the case where one does have prudential concern for the Z-people). Furthermore, given our background assumptions, Separability implies Prudential Neutral Addition: recall that we *defined* the zero level to be the level of wellbeing such that ordinary survival at that level and ordinary death would be equally prudentially good. If we write 0_1 for a fission population containing one life at this level, this amounts to the claim that $0_1 \sim \emptyset$: continuation as a single

individual at level 0 is equally as good as continuation as no individuals at all, i.e. ordinary death. Given Separability, this implies that $X + 0_1 \sim X$ for any fission population X disjoint from 0_1, and we can apply transitivity to show that addition of *any* finite number of neutral lives would be likewise equally good as having no addition at all.

If Gustafsson and Kosonen's argument succeeds, we can therefore use it to establish Prudential Neutral Addition. But *whether* their argument succeeds depends on how we understand the claim that the relation of prudential concern is intrinsic. Here are two interpretations of this claim.

Intrinsic Instantiation Claim: The matter of whether you stand in the relation of prudential concern to some person stage or stages does not depend on whether or not you also stand in the relation of prudential concern to some further person stage or stages.

Intrinsic Value Claim: The prudential value of standing in the relation of prudential concern to some person stage or stages is unaffected by whether or not you also stand in the relation of prudential concern to some further person stage or stages.

The Intrinsic Instantiation Claim is clearly a presupposition of the Parfitian view outlined in §6.2.1: we need it to rule out the view that in fission cases one lacks what matters in survival because of the mere fact of duplication. However, the Intrinsic Instantiation Claim does not by itself imply

Separability. To see this, simply note that on Fission Averagism, the Intrinsic Instantiation Claim may well be true, but Separability will certainly be false.

Gustafsson and Kosonen seem to have the Intrinsic Value Claim in mind instead. It is clear enough that this principle supports Separability: if the prudential value of standing in the relation of prudential concern to your fission offspring is intrinsic, then the way fission populations are ranked should not depend on what happens to their unaffected parts.[14] But does it follow from the Parfitian view outlined in §6.2.1? Well, perhaps. The idea that an additional success cannot be a failure seems to presuppose that an additional success could not diminish the prudential value of one's initial success. And it's hard to see why we should think this unless we accept the Intrinsic Value Claim. Still, this admittedly powerful intuition seems to me on somewhat shakier ground than the Intrinsic Instantiation Claim.

[14]Suppose that $X \succeq Y$. The value of standing in the relation of prudential concern to the fission offspring in X must therefore be at least as great as the value of standing in the relation of prudential concern to the fission offspring in Y. Given the Intrinsic Value Claim, when we compare $X + Z$ to $Y + Z$, it must still be true that the value of standing in the relation of prudential concern to X is at least as great as the value of standing in this relation to Y. The value of standing in the relation of prudential concern to Z is likewise unaffected by whether X or Y forms the remaining part of the fission population. Putting these claims together, the situation where one stands in the relation of prudential concern to $X + Z$ with the situation in which one stands in the relation of prudential concern to $Y + Z$, the first case involves standing in an equally prudentially good relation, plus another relation which is prudentially at least as good, compared to the second. All things considered, the first case $(X + Z)$ must therefore be at least as good as the second $(Y + Z)$.

6.3.3 Why Accept Fission Population Separability?

Along with appealing to principles like the Intrinsic Value Claim, we can argue for Separability, or against theories which violate Separability, directly. We can do this by simply transplanting the standard arguments for separability in population axiology to the case of fission.

Intrinsic Plausibility

If you are making a decision about the outcome of a fission operation, intuitively it seems quite rational to bracket the facts about what happens to those of your fission offspring whom you cannot affect.

Suppose that you are a person from a non-human species whose brains can be split into three parts, each of which are enough for survival: a left part, a right part, and a centre part. Imagine you are deciding whether to undergo a fission operation which will result in the left-part-person surviving at a higher wellbeing level while the right part of the brain is destroyed, or instead choosing a fission operation which will allow both the left-part-person and the right-part-person to survive, but at lower wellbeing levels. Intuitively, to make your decision, you only need to know about the potential wellbeing levels of the left-part and the right-part persons: you do not also need to know whether the centre-part-person will survive, or what their wellbeing level will be. And that is just what Separability says: if what happens to some of your

fission offspring is the same in two alternatives, then what happens to these fission offspring is irrelevant to the evaluation of these alternatives.

Egyptology: Aliens Edition

The most famous objection to non-separable population axiologies is known as the "Egyptology objection".[15] According to this objection, non-separable population axiologies like the Average view should be rejected because they imply, implausibly, that what we ought to do now might depend on facts about the Ancient Egyptians: how many of them there were, what their wellbeing levels were. With a little effort, something similar to this objection can be adapted for fission cases. Consider the following choice situation:

> *Andromedology* Aliens have visited earth with a new technology allowing humans to undergo fission into any number of offspring. They give you a choice of two fission operations, telling you in detail the numbers and wellbeing levels of your fission offspring, all of whom will remain on Earth. (They won't take 'no' for an answer.) They have also already made their mind up to take a certain number of additional fission offspring back with them to their home in the Andromeda galaxy; the trip will take around three million years, during which your fission offspring will be in stasis. They haven't told you the number or

[15]See McMahan (1981: 115), Parfit (1984: 420) and Thomas (2022).

wellbeing levels of the fission offspring they will be taking back with them.

If you know the numbers and wellbeing levels of your Earth-bound fission offspring in either choice in Future Fission, is this enough to decide which fission operation to go for? On the face of it, the answer is clearly: Yes. While what happens in Andromeda will obviously be of prudential significance to you, you cannot do anything to change it, and by the time your Andromeda-bound fission offspring arrive, anyone who your choice might have affected will be long dead. If this answer is right, then it seems that we should accept Separability.

The Argument from Anteriority

Another argument for separability in population axiology, given by Thomas (2022: 280–282), can be repurposed for fission cases. A little notation will make this argument easier to follow. If $X_1 \ldots X_n$ are any fission populations, we can write $[X_1, \ldots, X_n]$ to denote the fission population prospect over a set $\{s_1, \ldots s_n\}$ of equi-probable states of nature, where X_i comes about on state s_i.

The main premise of Thomas's argument says that the prudential ranking of fission population prospects depends solely on the probability distribution of wellbeing levels for each individual fission offspring. More precisely, we need

Fission Anteriority Let X and Y be fission population prospects in which exactly the same fission offspring have positive probabilities of existence. If, for each fission offspring i and each wellbeing level w, X and Y give i the same probability of existence at w, it holds that $X \sim Y$.[16]

We also need the principle of

Simple Prudential Dominance For any fission populations X, Y and Z,

$$X \succeq Y \iff [X, Z] \succeq [Y, Z]$$

In most respects, Simple Prudential Dominance is a weaker version of the principle of Prudential Statewise Dominance I shall discuss in §6.4.[17] Since I shall defend Prudential Statewise Dominance in §6.4.5, I shall not argue for Simple Prudential Dominance at this point.

[16]Goodsell (2021) has provided an argument against population-axiological Anteriority in cases involving prospects which have infinite support, meaning that they give infinitely many finite outcomes a positive probability. (This is possible if the probabilities of successively less-likely outcomes decrease quickly enough that infinitely many of them sum to 1). Goodsell's argument shows that Anteriority is incompatible with a version of stochastic dominance which applies to infinite gambles, together with some apparently minor auxiliary assumptions. The same argument presumably works against Fission Anteriority, but as far as I can see, it requires Anteriority to apply to prospects of infinite support. Since we are not dealing with any prospects of infinite support, we can appeal to this restricted version of Anteriority; though admittedly, restricting Anteriority in this way may be hard to motivate.

[17]The exception is that Simple Prudential Dominance, but not Prudential Statewise Dominance, implies that if $[X, Z] \succeq [Y, Z]$, then X must be at least as good as Y. (Given reflexivity, Prudential Statewise Dominance rules out that X is worse than Y, but leaves open that the two populations may be incomparable in this case.)

Thomas's argument proceeds as follows. Let X, Y and Z be any fission populations, and assume that Z is disjoint from X and Y. Simple Prudential Dominance implies that

$$X + Z \succeq Y + Z \iff [X + Z, \emptyset] \succeq [Y + Z, \emptyset]$$

Because the X-people, Y-people and Z-people get the same chances of existence at the same wellbeing levels, Anteriority implies both

$$[X + Z, \emptyset] \sim [X, Z]$$

and

$$[X + Y, \emptyset] \sim [X, Y].$$

It follows that

$$[X + Z] \succeq [Y + Z] \iff [X, Z] \succeq [Y, Z] \qquad (*1)$$

Finally, Prudential Statewise Dominance implies that

$$[X, Z] \succeq [Y, Z] \iff X \succeq Y \qquad (*2)$$

From (*1) and (*2), we obtain

$$X + Z \succeq Y + Z \iff X \succeq Y$$

as required.

Given our assumption of transitivity, Separability therefore follows from Fission Anteriority and Simple Prudential Dominance. These principles are, respectively, closely related to the principles of Prudential Ex Ante Pareto and Prudential Statewise Dominance which we shall discuss presently.

6.3.4 Summing Up the Case for the Neutral Addition Principle

To summarise, given our definition of the neutral level the Neutral Addition Principle follows from Fission Population Separability. We have four considerations in favour of Separability. The first is that it follows from the Intrinsic Value Claim, which is a natural principle to accept on the sort of Parfitian view we are assuming. The second is that Separability is compelling in its own right. The third is that if Separability were false, we would have to accept implausible conclusions in cases like *Andromedology*. The fourth is that Separability follows from Fission Anteriority and Simple Prudential Dominance.

The fourth consideration might seem the most promising of all, since both

Fission Anteriority and Simple Prudential Dominance are compelling. However, as we shall see in the next section, the main reason to reject principles like Anteriority (or Prudential Ex Ante Pareto) is via denial of a somewhat stronger version of Separability. In light of this, the argument from Fission Anteriority and Simple Prudential Dominance might be better thought of as showing how Anteriority-like principles and Separability are closer together than they might at first appear, rather than as showing how Separability can be supported by appeal to independently plausible Anteriority-like principles.

In the next section, we shall see an argument for Same-Number Fission Totalism on the EUT scale of wellbeing. The two most important premises of this argument, Prudential Statewise Dominance and Prudential Ex Ante Pareto, are closely related to the premises of the Anteriority argument for Separability discussed in §6.3.3. The former do not imply the latter as a matter of logic, but I think we should certainly accept the latter if we do accept the former.[18] Since I do not think we should reject the other premises of the forthcoming argument for Same-Number Fission Totalism, the lesson I shall draw is that we should accept Fission Totalism on the EUT scale of wellbeing if and only if we accept Prudential Statewise Dominance and Prudential Ex Ante Pareto.

[18] Anteriority follows from Prudential Ex Ante Pareto if we also accept a principle of stochastic indifference for individual fission offspring, on which fission population prospects are equally good *for* individual fission offspring if they grant the same chances of existence at the same wellbeing levels, including in cases where there are risks of non-existence. I believe that we should accept this principle.

6.4 Same-Number Fission Cases

6.4.1 Harsanyi: Fission Edition

We have seen that prudential axiology and moral population axiology are structurally analogous. Because of this structural analogy, we can provide an exact analogue of Harsanyi's Aggregation Theorem, adapted for fission cases, to support Same-Number Fission Totalism on the EUT scale of wellbeing.[19] This argument is powerful and compelling, but I do not claim that it is completely decisive. As we shall see, one of its premises, though rather plausible, might be rejected if we doubt Separability.

6.4.2 A Note On Expected Utility Theory

In order to get this argument going, we need to use the EUT scale of wellbeing, which means we need to assume that, in ordinary cases, the prudential betterness relation on prospects for individuals satisfies the axioms of Expected Utility Theory. As mentioned before, I won't be attempting to justify the claim that these axioms are satisfied by the prudential betterness relation in this chapter. Still, I think I can say something which may mollify those who suspect that the prudential betterness relation does not satisfy these

[19] See Harsanyi (1955), as well as later renditions and related results such as those found in Fishburn (1982), Broome (1991) and especially Fleurbaey (2009). The version I shall use is closest to Fleurbaey's result.

axioms in general. This is that, even if the prudential betterness relation does not satisfy these axioms in general, it may yet satisfy them on some proper subset of the set of all wellbeing levels.

To be more specific, I expect that most of those who are inclined to reject the axioms of *Completeness* and *Transitivity* in their full generality may yet be happy to accept restricted versions of these principles.[20] Recall that, as mentioned in §6.2.4, we are only considering prospects over lives or wellbeing levels which can be represented on the life-years scale of wellbeing. It is very plausible, I think, that Completeness and Transitivity apply to the prudential betterness relation on prospects over such lives, even if they do not apply to prospects over all lives in general.[21]

Consider first Completeness. Philosophers who reject Completeness typically do so because they believe that certain goods do not trade off against each other in a precise way.[22] But this kind of consideration does not appear to make any real trouble for the restriction of Completeness which only applies to wellbeing levels which are representable on the years-of-good-life scale: prospects over such lives merely trade off different quantities of exactly the same type of good.

[20]Recall that Completeness says that, if X and Y are lives, then either $X \succeq Y$, or $Y \succeq X$ (or both, in which case they are equally good). Transitivity says that, if $X \succeq Y$ and $Y \succeq Z$, then $X \succeq Z$.

[21]When it comes to Completeness, Broome (2004: 81) makes a similar move to my own: he ignores the potential incompleteness of the prudential betterness relation as a simplifying assumption, despite admitting that completeness may not be true of the prudential betterness relation in general.

[22]See for example Raz (1986: 326), Chang (2002), or Thornley (forthcoming).

Next, consider Transitivity. Some opponents of Transitivity argue that the reason Transitivity can fail is that different considerations can become prudentially (or morally) relevant as we consider different pairs of alternatives: a consideration can matter for the choice between X and Y, without mattering in the same way for the choice between Y and Z.[23] Others argue that pleasures and pains are goods (or bads) for which differences in intensity can lead to a difference in kind, such that sufficiently intense pleasures or pains can outweigh any quantities of mild pleasures or pains.[24] Once again, these kinds of objections seem to pose little threat to the restricted version of Transitivity. If we compare prospects involving different probabilities of different lengths of life at quality q, it does not at all seem plausible that different considerations could become relevant as we move between pairwise comparisons, because at all stages our only concern is with getting as much of q as we can, taking uncertainty into account. And, since the lives at issue only ever contain more or less time at the same quality level q, there is no question of any differences in intensity arising in the first place.

Thus, even those who reject Completeness and Transitivity in their full generality for the prudential betterness relation may well accept restricted versions of these principles. If these restricted principles are true (along with the other axioms of Expected Utility Theory), then the axioms of Expected

[23]See Temkin (2012: ch. 11-12).
[24]See Rachels (1998) and Temkin (1996).

Utility Theory will be satisfied by the corresponding restriction of the prudential betterness relation on prospects in ordinary cases. The argument of this section will then yield a restricted version of Same-Number Fission Totalism on the EUT scale, which applies only to fission populations whose fission offspring have wellbeing levels within this restricted range.

6.4.3 Premises of the Aggregation Theorem

Let us then proceed to the argument for Same-Number Fission Totalism, taking wellbeing levels to be those generated by Expected Utility Theory in ordinary cases. We may then assume that it is best for each fission offspring to maximise their expected wellbeing:

> *Expected Utility Maximisation for Fission Offspring* Let X and Y be any fission population prospects, and let i be any fission offspring who exists for sure in both prospects. $X \succeq_i Y$ if and only if the expected wellbeing of i in X is at least as great as the expected wellbeing of i in Y.

We also need three further substantive principles.

The first captures the idea that if one fission population is at least as good for every fission offspring as another ex ante, it is at least as good for the fission parent ex ante. To state it, we shall need a bit of notation: if i is a fission offspring, let $\succeq_i$ stand for the *individual* at-least-as-good-as

relation on fission population (prospect)s That is, $\succeq_i$ ranks fission population (prospect)s in terms of how good they are *for* i.

> *Prudential Ex Ante Pareto* Let X and Y be any fission population prospects which result in the same fission offspring in every state of nature. If, for all fission offspring i who have a positive probability of existence in either X or Y, $X \succeq_i Y$, then $X \succeq Y$. If, additionally, $X \succ_i Y$ for some fission offspring i with positive probability of existence, then $X \succ Y$.

The second is

> *Prudential Statewise Dominance* Let X and Y be any fission population prospects over the set S of states of nature. If, for each $s \in S$, the fission population $X(s)$ is at least as good as the fission population $Y(s)$, then $X \succeq Y$. If additionally $X(s) \succ Y(s)$ for some s, then $X \succ Y$.

The third and final principle we need is a principle of anonymity, which says that permuting the identities of one's fission offspring while keeping their wellbeing levels the same makes no all-things-considered prudential difference:

> *Prudential Anonymity* Let X and Y be any fission populations. If there is a permutation f of the possible fission offspring such that all

the fission offspring in X are mapped one-to-one onto all the fission offspring in Y, and this permutation preserves wellbeing levels (so that each fission offspring has the same wellbeing as her counterpart), then $X \sim Y$.

6.4.4 The Argument

For ease of exposition, I shall show here only that having one fission offspring at 100 and another at 0 is equally as good as having both at 50. But the argument generalises.

Consider the following two fission population prospects over equi-probable states of nature s_1 and s_2:

X	s_1	s_2
Lefty	100	100
Righty	0	0

Y	s_1	s_2
Lefty	100	0
Righty	0	100

According to Prudential Anonymity, the fission populations resulting from X and Y on s_2 are equally good, since it does not matter whether Lefty or Righty gets the 100 utiles; in s_1, X and Y are also equally good by reflexivity. Statewise Dominance therefore implies that $X \sim Y$. Now, consider the equalised fission population prospect:

Z	s_1	s_2
Lefty	50	50
Righty	50	50

By Expected Utility Maximisation, Y and Z are equally good for both Lefty and Righty. Ex Ante Pareto therefore implies that $Y \sim Z$. Applying transitivity, we find that $X \sim Z$. More generally, it can be shown that the principles used in this argument imply Same-Number Fission Totalism.

6.4.5 Justifying the Premises

Given the structural similarities between the axiology of prudence in different-number fission cases and population axiology, it should be no surprise that Harsanyi's aggregation theorem can be adapted for fission cases. The question is, how compelling are the three substantive premises of this theorem in the prudential case compared to their better-known moral counterparts? Let us see by examining each principle in turn.

Prudential Anonymity

Next, consider Prudential Anonymity. I defended the population-axiological version of this principle at length in Chapter 1; the same arguments also apply to the case of fission, with the added bonus that the standard person-affecting considerations against population-axiological anonymity simply do not apply to the case of fission if Identity Does Not Matter. If the facts about

personal identity make no difference, then surely changes in the identity of one's fission offspring also make no prudential difference.

Prudential Statewise Dominance

Next, consider Prudential Statewise Dominance. The corresponding moral principle says that

> *Moral Statewise Dominance* Let X and Y be population prospects over the set S of states of nature. If, for each $s \in S$, the population $X(s)$ is at least as good as the population $Y(s)$, then $X \succeq Y$. If additionally for some s, $X(s) \succ Y(s)$, then $X \succ Y$.[25]

The standard argument against Moral Statewise Dominance in the context of the Aggregation Theorem is due to Diamond (1967). The nub of this argument is that, if we compare a gamble which gives a good thing to one person for sure with an alternative gamble which gives each person an equal chance of receiving the same good thing, Moral Statewise Dominance (given Anonymity) says that the two gambles are equally good. But intuitively, that is not the case, because the second gamble gives each person a fair chance at receiving the good thing, while the first gamble unfairly delivers it to one person for sure.

Whether you are persuaded by this objection or not, Prudential Statewise Dominance has a clear advantage over its moral counterpart in that this

[25] $\succ$ here denotes the overall betterness relation on population prospects.

objection does not apply in the prudential case. Considerations of fairness intuitively matter in the moral case, but they do not intuitively matter in the prudential case. Consider, for example, a standard "Separateness of Persons" style objection to moving from the risky-but-fair gamble to the unfair-sure-thing gamble. Those who lose out on their chances to receive the good thing at stake can object that, although by giving up their chance of receiving the good thing, they thereby increase the chance that the lucky singled-out person will receive the good thing, this does nothing to compensate them for *their* loss. But in the prudential case, if one fission offspring gains at the expense of another fission offspring, as far as the fission parent is concerned the loss to one really *is* compensated by the gain to another, because the fission parent has an interest in the wellbeing of *all* her fission offspring. Fission parents are not in the business of weighing up the competing interests of their fission offspring in a fair manner, respecting the fact that they are separate people; they are instead in the business of getting the best overall outcome for *themselves*.

Diamond (1967: 766) is willing to reject Moral Statewise Dominance because, in addition to being interested in final states, "society is [...] interested in the process of choice". But he explicitly contrasts social choice with individual choice, in which we are *not* concerned with the process of choice, only in attaining better final states. Prudence in fission cases just *is* a matter of individual choice, concerning individual betterness. A concern for the pro-

cess of choice is not justifiable in this context, nor is a mandate to respect the separateness of one's fission offspring by taking into account considerations of ex ante fairness.

Ex Ante Pareto for Fission Offspring

The final premise needed to complete the argument for Same-Number Fission Totalism is Prudential Ex Ante Pareto. The moral analogue of this principle is

> *Moral Ex Ante Pareto* Let X and Y be population prospects over the same set of states of nature, and suppose X and Y result in the same people existing in every state of nature. If, for all persons i with a positive probability of existence in X or Y, $X \succeq_i Y$, then $X \succeq Y$. If, additionally, $X \succ_i Y$ for some i, then $X \succ Y$.

It can be shown by an inductive argument that both Prudential Ex Ante Pareto and its moral counterpart follow from the restriction of Ex Ante Pareto to cases where only a single person is affected. That is, Prudential Ex Ante Pareto follows from

> *Restricted Prudential Ex Ante Pareto* Let X and Y be any fission population prospects which result in the same fission offspring in every state of nature. Suppose that these fission population prospects differ for at most one fission offspring: that is, there is some fission offspring i

with a positive probability of existence such that for all fission offspring $j \neq i$, j has the same wellbeing level in X and Y in each state of nature in which j exists. Then $X \geq Y$ if and only if $X \geq_i Y$.

Suppose that the antecedent of Prudential Ex Ante Pareto is satisfied for fission population prospects X and Y, i.e., X and Y result in exactly the same fission offspring in each state of nature, and $X \succeq_i Y$ for each fission offspring i. Suppose there are exactly n fission offspring with a positive probability of existence in X (and therefore also in Y). We can then turn Y into X in n steps: we first replace the Y prospect of the first fission offspring with their X prospect; then we do the same for the second fission offspring, and so on until the prospect of each fission offspring has been replaced and everyone receives their X prospect. The fission population prospect resulting from the completion of all n steps will be X. By Restricted Prudential Ex Ante Pareto, the $k + 1^{th}$ step will always be at least as good as the k^{th} step. Transitivity thus implies that $X \succeq Y$. Of course this argument also applies, changing what needs to be changed, to Moral Ex Ante Pareto.

Since we are assuming Transitivity, it thus makes sense to focus our attention on the logically weaker Restricted versions of Ex Ante Pareto. Both the moral and the prudential versions of these are compelling. Even so, Moral Ex Ante Pareto has its detractors. Rabinowicz (2002) has argued that we should reject Moral Ex Ante Pareto in favour of an ex post version of Priori-

tarianism. Roughly, the view is that we should give priority to the worse off, and that the "worse off" are precisely those who are badly off in the state of nature under consideration, independently of how these people fare in other states of nature.[26] Restricted Moral Ex Ante Pareto will fail on this version of Prioritarianism: giving an individual five units of wellbeing for sure will produce greater expected *priority-weighted* wellbeing than giving her a half chance of 10, otherwise nothing, even though the two alternatives would be prudentially equally good for the individual in question.[27]

Ex Post Prioritarianism of this sort implies that even in the case where only one person exists at all, Moral Ex Ante Pareto can still fail. However, the analogous situation cannot hold in fission cases. If only one fission offspring will exist, then what's best for the fission parent ex ante must be the same as what's best for the fission offspring ex ante.[28] While it is conceptually

[26]In contrast, *ex ante* Prioritarianism, which has been most comprehensively developed by McCarthy (2006), holds that whether someone counts as "worse off" depends on their *expected* wellbeing, which takes into account their wellbeing levels across all states of nature.

[27]Recall that we are presently speaking in terms of the wellbeing scale generated by Expected Utility Theory (assuming that the individual betterness of prospects in ordinary cases satisfies the relevant axioms). This is what guarantees that receiving a sure thing of 5 units of wellbeing must be equally as good as receiving a 50-50 gamble yielding either 10 or 0 units of wellbeing. This does not immediately imply that the same will hold true if we are speaking in terms of another wellbeing scale, such as the life-years scale.

[28]Given that Identity Does Not Matter, any single-person fission population prospect will have the same prudential value as the corresponding prospect in which the fission parent instead gets ordinary survival with the same sort of life as her fission offspring in each state of nature. Imagine that the fission parent has *no* history (or an arbitrarily short, uneventful history with neutral prudential value). A situation in which one has no history at all (or an arbitrarily short and uneventful history), followed by a prospect of survival at given wellbeing levels, has the same prudential value as the prospect of one's entire life being at those wellbeing levels. So, in single-person cases, if the fission parent

possible (though counter-intuitive) for morality and prudence to come apart, it is not possible for prudence and prudence to come apart.

In this sense, then, Prudential Ex Ante Pareto is more compelling than Moral Ex Ante Pareto: while there is a coherent moral theory on which Moral Ex Ante Pareto is false in single-person cases, no such theory is tenable in the prudential case. In particular, we can rule out that Prudential Ex Ante Pareto fails in the case where only one fission offspring exists. But this does not yet mean that we must accept Restricted Prudential Ex Ante Pareto. To get Restricted Prudential Ex Ante Pareto from its logically weaker single-person cousin, we need

> *Fission Prospect Separability* Let X and Y be any fission population prospects, and let Z be any fission population prospect disjoint from X and Y. Then $X \succeq Y$ if and only if $X + Z \succeq Y + Z$.[29]

The moral analogue of this principle is compelling: how could the relative values of two population prospects depend on what happens to people who

has no past history, the prudential value of a fission population prospect will be the same as the prudential value of that prospect for the fission offspring who exists in it. Now recall our assumption that the prudential value of a fission population prospect is independent of the past history of the fission parent. From this we can deduce that the ranking of fission population prospects when the agent has any history must be the same as the ranking of the same fission population prospects when the agent has no history, which in turn must be the same as the ranking of those prospects for the single fission offspring who exists in them.

[29] We can extend the notion of "disjointness" and the $+$ notation from fission populations to fission population prospects in the obvious way: two fission populations prospects X and Y over the same set of states of nature are disjoint if and only if they result in disjoint fission populations in each state of nature; $X + Y$ is the fission population prospect which, for each state of nature s, results in $X(s) + Y(s)$.

are totally unaffected by the choice between the two? To reject Separability for Population Prospects seems to go against the basic thought that what matters, morally speaking, is what happens to particular people, and fails to take seriously the distinctness of individuals: there is just no such thing as the point of view of society or the universe, from which separate individuals could form an organic unity, or gain the sort of holistic value that would be needed for Separability to fail.

However, the same cannot be said in the case of Fission Prospect Separability. Here there really *is* a perspective which amalgamates the lives and experiences of all fission offspring, namely the perspective of the fission parent. Unlike in the moral case, it does not seem that we can rule out, on theoretical grounds, the view that separate fission offspring contribute to holistic prudential value for the fission parent. A cautionary aside, though: recall that we are only considering cases where the lives of fission offspring contain no incomplete holistic goods. So it cannot be that the goods *themselves* have holistic value in virtue of what happens in the lives of separate fission offspring.[30] The view under consideration is instead that the *lives* of fission offspring have holistic value for the fission parent in virtue of the prospects faced by separate fission offspring, even though this holistic value

[30]Even if we did consider cases like this, we would not be able to apply Population Prospect Separability anyway. This is because the wellbeing levels of some of the fission offspring would change depending on whether the "unaffected" fission offspring are included in the fission population prospect.

is not prudentially significant for the fission offspring themselves. Since we cannot rule out this sort of view on the same grounds as we can in the moral case, Prudential Ex Ante Pareto is in this respect less compelling than Moral Ex Ante Pareto.

Summing Up

Most of the premises of the prudential analogue of the Harsanyi Aggregation Theorem are straightforwardly harder to reject than their moral counterparts. The exception is Prudential Ex Ante Pareto. This is more of a mixed bag: it is in one respect more plausible than its moral counterpart, because it is absurd that it should fail in the one-person case. However, it is in another respect less plausible than its moral counterpart, because while Population Prospect Separability is supported by theoretical considerations from the personal perspective and the separateness of persons, the same is not true of Fission Prospect Separability.

Still Prudential Ex Ante Pareto remains an intuitively compelling principle. The prudential analogue of Harsanyi's Aggregation Theorem thus has considerable force. But I cannot say that it is an entirely decisive argument. It might fail if we reject Fission Prospect Separability. And in any case, without that principle we cannot obtain Prudential Neutral Addition, and without Prudential Neutral Addition we cannot move from Same-Number Fission Totalism to Fission Totalism. Conversely, if we *do* accept Fission

Prospect Separability, then we get Prudential Neutral Addition almost for free, and there are no good grounds to reject Prudential Ex Ante Pareto or the other premises of the prudential Aggregation Theorem. The upshot is that we should accept Fission Totalism on the EUT scale if and only if we accept Fission Prospect Separability.[31]

In §6.5, I shall provide a separate argument for Fission Totalism, this time on the life-years scale of wellbeing. If this argument succeeds, then Fission Totalism is true on the life-years scale. In that case, Fission Totalism is true on the EUT scale if and only if the EUT scale and life-years scale coincide.

6.5 A Three-Step Argument for Fission Totalism

6.5.1 Overview of the Argument

Consider the following four outcomes:

Synchronous Fission A fission operation which produces two individuals. Both simultaneously have one day of good life. Afterwards, both die.

Asynchronous Fission A fission operation which produces two individuals. The first individual has one day of good life, and then dies. The second

[31]But recall that all this is predicated on the plausible assumption that the prudential betterness relation satisfies the axioms of Expected Utility Theory in ordinary cases, at least on a restricted range of wellbeing levels.

individual wakes for the first time on the second day, has one day of good life, and then dies.

Interrupted Survival Ordinary survival for two days of good life. However, at the beginning of the second day, the agent loses all memories and other psychological relations of prudential significance which pertain to her past self on the first day. The agent loses no memories or other psychological relations pertaining to her past self at any time before the first day.

Ordinary Survival Ordinary survival for two days of good life.

I shall argue that each outcome is prudentially equally as good as the next for the agent's person-stage existing just before commencement of the first day. That is:

(i) *Asynchronous Fission* $\sim$ *Synchronous Fission*

(ii) *Interrupted Survival* $\sim$ *Asynchronous Fission*

(iii) *Ordinary Survival* $\sim$ *Interrupted Survival*

From these three claims, it follows by transitivity that *Synchronous Fission* and *Ordinary Survival* are equally good. At no point will my argument rely on the fact that there are two *days* at stake, rather than other lengths of time (including two unequal lengths of time), nor will it rely on there being specifically two fission offspring in *Synchronous Fission*, nor will it rely on the

periods of time being at a certain, constant good quality. So if my argument succeeds, it generalises to show that splitting into multiple fission offspring who might exist simultaneously is prudentially equally as good as getting ordinary survival with an aggregate of the lives of all these fission offspring. Given Prudential Ex Post Pareto, this implies that Fission Totalism is true on the life-years scale of wellbeing.

The question, then, is whether we should accept claims (i) to (iii). Let us consider each in turn.

6.5.2 Step One

First, consider claim (i), according to which *Asynchronous Fission* is prudentially equally as good as *Synchronous Fission*. The two cases differ only in the timing of the life of the second fission offspring: the first fission offspring has exactly the same sort of life either way. So there can only be a prudential difference between these two cases if the timing of the life of the second fission offspring can make a prudential difference.

There are two reasons it might be thought that a change in timing could make a prudential difference. The first is that one might think it appropriate to have a pure time preference: to discount one's future wellbeing on the basis of the time that passes between now and then. The second is that one might think that it is better for the lives of one's fission offspring to

be spread out in time, rather than occurring simultaneously. However, both these views lead to implausible implications in the following case.

> *Vault Dwellers* You are about to undergo fission into two fission offspring, Lefty and Righty. Just before of your fission offspring wake up, they will be put into technologically advanced stasis pods, which will perfectly preserve them for certain lengths of time, which may be different. If you were to use these stasis pods while conscious, it would seem as though no time at all had passed on the inside, but one hundred, one thousand, or one hundred thousand years might have passed on the outside. After undergoing stasis, your fission offspring will wake up in entirely separate underground communities and live out their lives at a decent standard or living.

In cases like Vault Dwellers, does it matter to you how long the stasis lasts for each of your fission offspring? According to the two aforementioned views, it does. If you should employ pure time discounting, then it will be very bad for you if the stasis lasts a long time. If it matters to you whether your fission offspring are spread out in time, rather than living at the same time, it will be bad for you if the lengths of the two periods of stasis are about the same, so that the lives of your two fission offspring overlap. But how could it matter how long the stasis periods will last? It would not matter at all to Lefty, nor would it matter to Righty. They will wake up and enjoy exactly

the same sort of life, regardless of the length they spend in the stasis pod. They wouldn't even know how long the stasis lasted, unless they were told about it.

It might seem that this argument appeals to the principle of Prudential Ex Post Pareto mentioned in 6.2.2. I raised a potential problem for that principle: it implies that, if a very short but excellent life is equally as good as a much longer but mediocre life, then undergoing fission into millions of short, excellent lives is equally as good as undergoing fission into the same number of longer, mediocre lives. One might doubt that this implication is really true. However, we do not actually need Prudential Ex Post Pareto to support my claim about the Vault Dwellers case. Instead, we can appeal to a much weaker principle, namely

> *Supervenience on Life Types* If fission operations O_1 and O_2 produce exactly the same fission offspring, and these fission offspring have qualitatively identical lives in either case, then O_1 and O_2 are equally good for the fission parent.

This principle is eminently plausible, and it is neutral on the aforementioned controversial implication of Prudential Ex Post Pareto. By appealing to Supervenience on Life Types, we get the right result in the Vault Dwellers case. It also explains why a mere difference in the spatio-temporal location of a fission offspring cannot make an all-things-considered prudential difference

for the parent: such differences make no qualitative difference to the lives of the fission offspring.

6.5.3 Step Two

Next, consider claim (ii), which holds that *Interrupted Survival* and *Asynchronous Fission* are equally good. The only potentially relevant difference between these two cases is the difference in the identities of the people concerned. In the case of ordinary survival, the first-day-person is you, and the second-day-person is also you. In the case of fission, the first-day-person is not you, and the second-day-person is different from the first-day-person (and is also not you).

These differences cannot be prudentially significant on our assumption that Identity Does Not Matter, since they are only differences in identity. Since there are no prudentially significant differences between the two outcomes, they must be prudentially equally as good, in line with claim (ii). So we should accept claim (ii) *if* we believe that Identity Does Not Matter.

6.5.4 Step Three

Finally, consider claim (iii), which is that *Ordinary Survival* is equally as good as *Interrupted Survival*.

It's important here to keep in mind our assumption that the lives in

question contain no incomplete holistic goods on either day. If you have two days left to live, and it will take two days to finish your *magnum opus*, it will be very bad for you to lose your memories after the first day, leaving you unable to complete your work. But we are not considering cases like these. We are considering cases where the severing of psychological connections does not prevent the realisation of holistic goods, nor does it make any difference whatsoever to the wellbeing of the agent on either day.

Another point worth emphasising is that while there is a severing of psychological relations between the first-day-agent and the second-day-agent, there is no such severing of psychological relations between either the first or the second day agent, and the agent before these two days. In particular, the agent loses her memory between her first and second day, but her second-day-self remembers being her pre-first-day self perfectly clearly. In short, the pre-first-day agent bears the full relation of prudential concern to both her first-day-self and her second-day-self.

The basic argument for the equal goodness of *Ordinary Survival* and *Interrupted Survival* is that both cases equally deliver the goods. More precisely, the agent gets exactly the same wellbeing at each time either way, and the pre-first-day agent has full prudential concern for herself at all points at which her wellbeing is at issue. These are the only things that matter, prudentially speaking, for the pre-first-day agent.[32] Since these things are

[32]Of course, it's a different story for the first-day agent: it may be very bad for *her* if

the same in *Ordinary Survival* and *Interrupted Survival*, the two cases must be equally prudentially good.

That's the theoretical argument, which may or may not convince you. A more pragmatic point is that the claim that interruption in ordinary survival does not matter delivers what seem to be the right results in a range of practical cases. It will be easiest to see this if we compare what we might call the Total Verdict, on which interruptions in ordinary survival are prudentially irrelevant, with the Average Verdict, on which an Interrupted Survival which partitions one's future into pieces of equivalent value is just as good as getting exactly one of these pieces. (That is, *Interrupted Survival* is equally as good as surviving one day.)

Consider the case of an agent – let's call him Derek – who unfortunately has two days to live. Derek has difficulty sleeping, but can resolve this problem by taking sleeping pills. However, these sleeping pills have a side effect: they cause complete memory loss of the day in which they are taken. If Derek takes the pills after his first remaining day, he will wake up fresh and rested on his final day. Otherwise, he will be tired, and will enjoy his last day on earth less.

Derek should, of course, take the pills if we ignore their side effect. It also seems to me that he should take the pills even though they have the side effect of memory loss: as far as I can see, this makes no difference. It is

her psychological connections to her future self are severed.

very clear, though, that the Average Verdict is wrong in this case: it would imply that taking the pills and surviving two days would be equally as good as surviving for only *one* day.

Another relevant case has been dramatised in an episode of the British science fiction television series *Black Mirror*.[33] The episode depicts a form of cruel and unusual punishment in which the protagonist is hunted by merciless sadists while being watched and filmed by uncaring members of the public. At the end of a terrifying day, the protagonist is informed that she had committed a heinous crime in the past, and that she is being psychologically tortured in order to punish her for this crime. She is then sent to have her memories wiped, and the cycle repeats indefinitely.[34]

The audience is clearly expected to react with horror at the protagonist's predicament: her lack of psychological connections with her past and future selves is not supposed to make her any better off. This reaction is ruled out by the twin assumptions that Identity Does Not Matter and that fission offspring lack any prudential concern for each other, since the case is designed in such a way that it differs from a fission case only because the protagonist's person-stages bear the relation of personal identity to each other.[35] But we

[33]Brooker (2013).

[34]We can assume here that this process severs *all* prudentially relevant psychological connections with her past self, not just memories.

[35]Since the intended audience reaction seems intuitively correct, this might be taken to be an objection to this Parfitian view, and in favour of the view that physical continuity is sufficient for the relation of prudential concern. A similar argument for the same conclusion is given by Williams (1970); a detailed response is offered by Parfit (1984: 229–243).

can imagine a variant of the case in which all of the psychological connections between the protagonist's person-stages on different days of torture are severed, but each stage still retains all prudentially relevant psychological connections with the pre-torture person-stages. We can ask in this case: would it be worse for the pre-torture protagonist to receive many days of torture, rather than fewer days of torture? The Average Verdict incorrectly says: No. The Total Verdict correctly says: Yes.

6.5.5 Summing Up (Life Years)

If the preceding arguments succeed, then *Synchronous Fission* is equally as good as *Ordinary Survival*. We can say something more general. Since synchronicity makes no prudential difference, not does a mere change in identity, and nor does an interruption in ordinary survival (for an agent before the interrupted period of time), our conclusion is

> *Fission Totalism (aggregate life version)* Undergoing a fission operation which results in fission population X is equally as prudentially good as ordinary survival as a single individual whose life is an aggregate of all of the lifetimes of the fission offspring in X.

The aggregate life version of Fission Totalism implies Fission Totalism on the life-years scale of wellbeing, provided the assumption of Prudential Ex Post Pareto, which we made in §6.2.2, is correct. This is because, if we are

just talking about lives of some constant good quality, the aggregated life corresponding to a fission population will be of the same constant quality, while lasting as long as all of the lives in the fission population put together. The wellbeing associated with this life, on the life-years scale of wellbeing, will therefore be the sum of the wellbeing levels associated with all lives in the fission population. Prudential Ex Post Pareto extends this equivalence to cases in which the lives can be of different or varying qualities.[36] I shall continue to assume Prudential Ex Post Pareto. But it's worth pointing out that if this assumption were to be abandoned, the argument of this section would still support the aggregate life version of Fission Totalism.

6.6 Time-Separability and Risk-Neutrality

6.6.1 Two Kinds of Fission Totalism?

In §6.3 and §6.4, I argued for Fission Totalism on the EUT scale of wellbeing. In §6.5, I provided a different argument for Fission Totalism on the life-years scale of wellbeing. If these two arguments are both sound, then the EUT scale and the life-years scale of wellbeing coincide. Or, equivalently, if the EUT scale and the life-years scale of wellbeing do not coincide, then at least

[36]This applies to cases where all lives in the fission population are of the same valence, i.e., they are all good, all bad, or all neutral. Things get a little more complicated when the fission offspring can have lives of differing valences. These complications disappear if the life-years and EUT scales of wellbeing coincide, as I shall argue presently, so I shall ignore this issue.

one of the two arguments for Fission Totalism must be unsound. I shall argue that the two wellbeing scales do coincide. To make this claim more precise, and to make my arguments for it clearer, it will be helpful to introduce some more notation.

6.6.2 Life Segments and Risk-Neutrality

Define a *life segment* to consist of all wellbeing-relevant information about a life during some length of time, and let a *life prospect* be a probability distribution over life segments. We can say that one life prospect is at least as good as another if and only if a prospect which gives the agent the first life prospect, and nothing else, is better for her than a prospect which gives her the second life prospect, and nothing else.[37] If l_i and l_j are life segments, we can define (l_i, l_j) to be the life segment consisting of first l_i, then l_j. We can extend this notation to life prospects in the obvious way.

In keeping with §6.2.3, we shall assume that none of the life segments under consideration contain incomplete holistic goods, and in line with §6.2.4, we shall only consider life segments which have a life-years wellbeing representation.

We can introduce two functions from life segments to real numbers: for any life segment l, we can write $w(l)$ to denote the life-years wellbeing level

[37] This definition works because life segments contain *all* wellbeing-relevant information.

of l, and we can write $u(l)$ to denote the EUT wellbeing level of l.[38] We can assume that the life-years and EUT scales both assign level 0 to the empty life segment; or, more precisely, the wellbeing level assigned by each approaches zero as a life segment of constant good quality approaches zero. This defines the EUT scale up to positive scalar multiplication. We can pick out a unique utility function u by stipulating that $u(w^{-}(1)) = 1$: that is, a life receiving one unit of wellbeing on the life-years scale also receives one unit of wellbeing on the EUT scale.[39]

Note that, by definition, it is best to maximise the expectation of wellbeing on the EUT scale. Or, for the benefit of those readers who (for whatever reason) really like sigma notation:

> *Risk-Neutrality (EUT)* For any life prospects l_i and l_j over the same set of states of nature S, $l_i \succeq l_j$ if and only if:[40]
>
> $$\sum_{s \in S} p(s) \cdot u(l_i(s)) \succeq \sum_{s \in S} p(s) \cdot u(l_j(s))$$

Putting it this way makes it easy to see that the claim that the EUT

[38] Throughout this section, I shall assume that an EUT scale exists; that is, that the axioms of Expected Utility hold for prudential betterness on life prospects.

[39] Since we want utility 0 to correspond to the level at which one would be indifferent between death and ordinary survival at this level, assigning 1 to a good life lasting for one year does involve the assumption that such a life has positive utility: it would be better than death. (So long as this life has positive utility, we can scale the utility function so that it has utility 1.)

[40] If l_i is a life prospect and s is a state of nature, $l_i(s)$ denotes the life segment resulting from life prospect l_i in state of nature s.

and life-years scales of wellbeing coincide, i.e., the claim that $w = u$, really amounts to the following claim about how it is best to respond to wellbeing on the life-years scale in cases of risk:

> *Risk-Neutrality (life-years)* For any life prospects l_i and l_j over the same set of states of nature S, $l_i \succeq l_j$ if and only if:
>
> $$\sum_{s \in S} p(s) \cdot w(l_i(s)) \succeq \sum_{s \in S} p(s) \cdot w(l_j(s))$$

Henceforth, I shall just call this claim "Risk-Neutrality", since Risk-Neutrality (EUT) is not a substantive claim.

One argument for Risk-Neutrality is implicit in what has come before: since I have already argued for Fission Totalism on both the EUT scale and the life-years scales of wellbeing, these two argument, taken together, support the coincidence of these two wellbeing scales.

6.6.3 Time-Separability

My main argument for Risk-Neutrality appeals to the principle of

> *Time-Separability of Wellbeing* For any life prospects l_i, l_j and l_k, $l_i \succeq l_j$ if and only if either:
>
> $$(l_i, l_k) \succeq (l_j, l_k) \qquad \text{(i)}$$

or

$$(l_k, l_i) \succeq (l_k, l_j) \qquad \text{(ii)}$$

The basic structure of the argument is as follows.

(P1) Wellbeing is Time-Separable.

(P2) If wellbeing is Time-Separable, then Risk-Neutrality is true.[41]

(C) Hence Risk Neutrality is true (and so the EUT and life-years wellbeing scales coincide).

To understand what Time-Separability amounts to, it might help to consider what it would take for it to fail. Here's one view on which Time-Separability is false: we should maximise the expectation of the *square root* of the life-years wellbeing levels we shall receive, rather than straightforwardly maximising the expectation of our expected life years without applying a square root function.

[41] A somewhat similar argument for Risk-Neutrality is considered by Broome (1991: 224–230). Broome's argument appeals to an intrapersonal, temporal ex ante pareto-like principle (the "Principle of Temporal Good") rather than to Time-Separability (though these two principles are closely related). Broome denies the soundness of this argument on two grounds. The first is the existence of incomplete holistic goods. The second is the apparent reasonableness of theories of intrapersonal aggregation which are not Time-Separable, such as the family of diminishing marginal utility views we shall come to shortly. The first objection does not apply to my own argument, since I restrict myself to cases not involving incomplete holistic goods. As for Broome's second objection, I shall later argue directly that wellbeing is indeed Time-Separable. If my arguments succeed, then Broome's second objection can be dismissed: while non-Time-Separable theories of wellbeing might look reasonable at first blush, they are in fact mistaken.

Diminishing Marginal Utility of Life Years (square root version) For any life prospects l_i and l_j over the same set of states of nature S which certainly yield good life segments,[42] $l_i \succeq l_j$ if and only if:

$$\sum_{s \in S} p(s) \cdot \sqrt{w(l_i(s))} \succeq \sum_{s \in S} p(s) \cdot \sqrt{w(l_j(s))}$$

More generally, we might apply any strictly concave function g, rather than the square root function, and we'll get a corresponding view on which life years have diminishing marginal utility according to g, i.e., we take the expectation of $g(w(l_i))$ rather than $\sqrt{w(l_i)}$.

To see that utility diminishing in this way causes Time-Separability to fail, consider the life prospects represented by the table below.

	$s_1(p = 0.4)$	$s_2(p = 0.6)$
l_1	36	36
l_2	16	16
l_3	64	0

Compare l_2 and l_3. l_2 yields the equivalent of 16 years of good life for sure, and so has an expected utility of 4. l_3 has a 40% chance of yielding 64 years of good life (utility 8) and a 60% chance of yielding 0 years of good life (utility 0). So, the expected utility of l_3 is 3.2. According to the square root version of the diminishing marginal utility view, then, l_2 is better than l_3.

But now compare (l_1, l_2) with (l_1, l_3). The first yields the equivalent of

[42] I.e., for each state of nature s, $f(l_i(s)) > 0$ and $f(l_j(s)) > 0$.

52 years of good life for sure. The second yields a 60% chance of getting the equivalent of 80 years of good life, and a 40% chance of getting merely 16 years of good life. The utility of the first life prospect is $\sqrt{52} \approx 7.21$ units of wellbeing, while the utility of the second life prospect is $0.4 \cdot \sqrt{100} + 0.6 \cdot \sqrt{36} \approx 7.6$ units of wellbeing. (l_1, l_3) is therefore better than (l_1, l_2). We thus have a violation of Time-Separability: l_2 is better than l_3, but if we add l_1 into the mix, the reverse is true: (l_1, l_3) is better than (l_1, l_2).

We can see that Time-Separability fails in this specific case of non-neutrality about risk, but this does not show that Time-Separability implies Risk-Neutrality in general. We shall fill in the gap now.

6.6.4 Substitution of Equivalents and Intrapersonal Neutral Addition

Before we begin, we will need to prove two small lemmas. The first is that Time-Separability implies the

> *Substitution of Equivalents* For any life segments l_j and l'_j, and any l_i and l_k which are either life segments or empty, $l_j \sim l'_j$ if and only if
>
> $$(l_i, l_j, l_k) \sim (l_i, l'_j, l_k)$$

To show this, we can first apply Time-Separability in both directions to

show that $l_j \sim l'_j$ if and only if $(l_i, l_j) \sim (l_i, l'_j)$. (If l_i is empty, this step is vacuous; the same will be true of l_k.) We then do the same thing again to show that $(l_i, l_j) \sim (l_i, l'_j)$ if and only if $(l_i, l_j, l_k) \sim (l_i, l'_j, l_k)$.

Since Substitution of Equivalents is true, it will do no harm to use the following notation: we can write l_x to denote a life segment such that $w(l_x) = x$. Of course, there are many life segments l such that $w(l) = x$, but Substitution of Equivalents implies it does not matter which we pick: since all such life segments are equally good, they will have equal contributive value within larger life segments and within prospects.

The second is that Time-Separability implies

> *Intrapersonal Neutral Addition* For any life prospects l and l_0, if l_0 is a neutral life prospect (i.e., for some $l_i, (l_i, l_0) \sim l_i$), then, for any l,
>
> $$l \sim (l, l_0) \sim (l_0, l)$$

First note that because $(l_i, l_0) \sim l_i$, we must also have $(l_i, l_0, l_0) \sim (l_i, l_0)$ by Substitution of Equivalents. By Time-Separability, it follows that $(l_0, l_0) \sim l_0$. Now let l be an arbitrary life segment, and note that again, by Substitution of Equivalents, $(l, l_0) \sim (l, l_0, l_0)$. It follows by Time-Separability that $l \sim (l, l_0)$. Similarly, Substitution of Equivalents implies that $(l_0, l) \sim (l_0, l_0, l)$, and then Time-Separability yields $l \sim (l_0, l)$.

6.6.5 Time-Separability Implies Risk-Neutrality

Part 1

We will now show that for any natural number n and positive x, $u(l_x) = x$ if and only if $u(l_{nx}) = nx$. The cases $n = 0$ and $n = 1$ are trivial, so assume $n > 1$, and consider the following two prospects over the set of equi-probable states of nature $s_1, \ldots s_n$.

	s_1	s_2	$\ldots$	s_n
A	l_{nx}	l_0	$\ldots$	l_0
B	l_x	l_x	$\ldots$	l_x

Note that $u(l_x) = x$, $u(l_0) = 0$, and the probability of each s_i is $\frac{1}{n}$. Therefore, $u(B) = x$, and $u(A) = \frac{u(nx)}{n}$. It follows that $u(nx) = nx$ if and only if $A \sim B$. So it is sufficient for us to show that $A \sim B$.

To see this, begin by considering the following life prospects:

	s_1	s_2	$\ldots$	s_n
B_1	l_x	l_0	$\ldots$	l_0
B_2	l_0	l_x	$\ldots$	l_0
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
B_n	l_0	l_0	$\ldots$	l_x

Note that, by the axioms of Expected Utility Theory, each $B_i \sim B_j$: they each give a $\frac{1}{n}$ chance of getting l_x, otherwise l_0. Now consider two life prospects. First, define A' to be equal to B_1 concatenated with itself n times, i.e., $A' = (B_1, B_1, \ldots, B_1)$. Second, define B' to be equal to $(B_1, B_2, \ldots B_n)$.

By Substitution of Equivalents, these two life prospects are equally good. They are represented in the table below.

	s_1	s_2	...	s_n
A'	$(l_x, \ldots, l_x)$	$(l_0, \ldots, l_0)$	...	$(l_0, \ldots, l_0)$
B'	$(l_x, l_0, \ldots, l_0)$	$(l_0, l_x, \ldots, l_0)$	...	$(l_0, \ldots, l_x)$

By Intrapersonal Neutral Addition and Expected Utility Theory, these life prospects are equally good as ones in which the superfluous l_0 life segments are stripped out. Furthermore, since $(l_x, \ldots, l_x)$ just is a life lasting for nx years at quality q (or at any rate is equally as good as one, given Substitution of Equivalents), we have $(l_x, \ldots, l_x) \sim l_{nx}$. Hence $A' \sim A$ and $B' \sim B$. It follows that $A \sim B$, as required.

Part 2

We next prove Risk-Neutrality for all positive real-valued life-year wellbeing levels. We do this by contradiction: suppose that for some positive x, we have $u(l_x) \neq x$. Write $\epsilon = |u(l_x) - x|$; by the assumption that $u(l_x) \neq x$, we have $\epsilon > 0$. Let n be the least natural number such that $\frac{1}{n} < \epsilon$, and let m be the least natural number such that $\frac{m}{n} > x$.

We have defined u such that $u(l_1) = 1$. Since $u(l_x) = x$ iff $u(l_{nx}) = nx$, it follows that $u(l_{\frac{1}{n}}) = \frac{1}{n}$, and that additionally, $u(l_{\frac{k}{n}}) = \frac{k}{n}$ for any k. Now, by

our choice of m and n, we know that[43]

$$\frac{m}{n} > x > \frac{m-1}{n}$$

Since u is monotonic with respect to years of good life, we then have

$$u(l_{\frac{m}{n}}) > u(l_x) > u(l_{\frac{m-1}{n}})$$

$u(\frac{m}{n}) = \frac{m}{n}$ and $u(\frac{m-1}{n}) = \frac{m-1}{n}$; therefore, we have

$$\frac{m}{n} > u(l_x) > \frac{m-1}{n}$$

We also know that the sum of differences between x and these upper and lower bounds on $u(l_x)$, that is $(\frac{m}{n} - x) + (x - \frac{m-1}{n})$, is equal to $\frac{1}{n} < \epsilon$. It follows that $|u(l_x) - x| < \epsilon$, which is a contradiction.

This suffices to prove that $u(l_x) = x$ for all positive real numbers x, while the case where $x = 0$ is trivial.

Part 3

All that remains is to prove that $u(l_x) = x$ for negative x. Fortunately, doing so is easy. Recall that we defined the level $-x$ on the life-years scale to

[43] If $\frac{m-1}{n}$ were equal to x, we would be done, as we would have $u(x) = u(\frac{m-1}{n}) = \frac{m-1}{n} = x$.

be a level such that one would be indifferent between ordinary death (level 0 for sure) and a 50-50 gamble yielding either x years of good life or the $-x$ wellbeing level. This means that $u(l_0) = 0.5 \cdot u(l_x) + 0.5 \cdot u(l_{-x}) = 0$; hence $u(l_x) = -u(l_{-x})$. Therefore, for any positive x, since by the previous argument we have $u(l_x) = w(l_x) = x$, we must also have $u(l_{-x}) = -u(l_x) = -x = w(l_{-x})$.

6.6.6 Arguments for Time-Separability

We have seen that Time-Separability implies Risk-Neutrality. Well and good, but why accept Time-Separability?

Argument 1: No New Information, No Preference Change

Recall the example of a diminishing marginal utility view discussed in §6.6.3. We saw how it violated Time-Separability by considering the three prospects in the table below:

	$s_1(p = 0.4)$	$s_2(p = 0.6)$
l_1	36	36
l_2	16	16
l_3	64	0

We saw that, on the square root version of the diminishing marginal utility view, $l_2 \succ l_3$; however, $(l_1, l_3) \succ (l_1, l_2)$. Suppose, then, that you are facing a choice like this, and following the future-oriented version of this theory: at

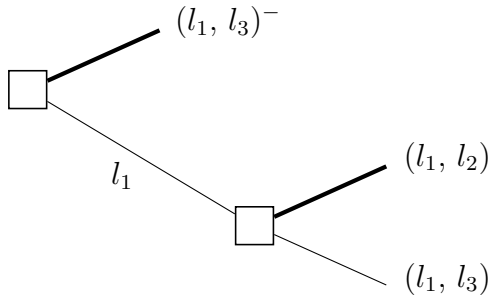
each time, you are trying to get the best future for yourself. Before having your first 36 years of good life, you would prefer to later get l_3, rather than l_2. After having those first 36 years of good life, you would instead prefer l_2 to l_3. But you have gained no new information of any sort about the probabilities involved: you have changed your preferences, apparently for no reason. Changing your preferences in this way is irrational, while choosing the expectedly best future for yourself is not irrational. Therefore, following the future-oriented diminishing marginal utility view does not ensure you are choosing the expectedly best future for yourself. So this view is false.

Argument 2: Money Pump

The previous problem can be sharpened. Suppose you are at the beginning of your 36 years of good life. You now prefer the risky gamble l_3 to the safe-but-mediocre option l_2. Suppose that, as things stand, you will get the risky gamble l_3 once your first 36 years are up. However, an exploiter is in a position to offer you a swap from l_3 to l_2 after 36 years have passed. They will, after 36 years have passed, make you this offer (which you will be perfectly at liberty to refuse) to swap from l_3 to l_2, unless you pay them a small sum of money now.

You know that, once the 36 years have passed, you would accept the exploiter's offer. At the moment, you prefer to get l_3, even with the loss of a small sum of money (or amount of wellbeing), compared to getting l_2.

So you accept your exploiter's offer, and pay him some money so that he will later leave you alone. This situation is illustrated by the decision tree below. The square boxes denote decision nodes, the leaves of the tree denote final outcomes, and the bold edges represent what would be chosen, on the future-oriented diminishing marginal utility view, at that decision node.



Following the future-oriented diminishing marginal utility view therefore renders you vulnerable to exploitation: you can end up paying for something you could have had for free. But you cannot be vulnerable to exploitation if you have rational preferences.[44] Hence, if you follow the future-oriented diminishing marginal utility view, you don't have rational preferences. If you were following the correct theory of prudence, you would have rational preferences. Therefore, the future-oriented diminishing marginal utility view isn't the correct theory of prudence.

Unlike some others, this money pump cannot be escaped by appealing to foresight or backwards induction.[45] It is foresight which leads you to pay

[44]See Gustafsson (nd).

[45]This is sometimes known as "sophisticated choice"; see for instance Pollak (1968).

your exploiter to leave you alone. The money pump could be escaped if we could appeal to a theory of *resolute choice*, on which you might turn down the exploiter's first offer, and then later turn down his offer to swap from l_3 to l_2.[46] But resolute choice is especially implausible in this case. We are talking about prudential *value*. How could it be prudentially rational to choose the alternative that is worse for you, just because of some previous decision you made? (The fact that you made that previous decision, after all, does not affect anything you care about *now*.)

Response: Choose On A Whole-Lifetime Basis

It might be replied that these sorts of problems can be avoided if we stipulate that an agent always ought to choose in line with *entire lifetime* wellbeing, rather than merely their future wellbeing. In this case, the 36 initial years of good life will always be considered. They will prefer l_3 before they get the 36 years in l_1, and they will still prefer l_3 after they get the 36 years in l_1.

Objection 3: What's Wrong With Future-Oriented Choice?

Choosing on a whole-lifetime basis does seem to be rational. But then again, provided our choice does not at all affect our past wellbeing, choosing with only the future in mind seems to be rational as well. So the initial challenge for this response is to explain exactly what is wrong with future-oriented

[46] See (McClennen, 1985, 2000).

choice. It is not enough to simply point at the rationality of whole-lifetime choice. Proponents of Time-Separability also accept that it is rational to choose on this basis. The point is that future-oriented choice and whole-lifetime choice *both* seem rational, yet only Time-Separable theories can account for this.

Objection 4: Childhoodology

Whole-lifetime choice for a non-Time-Separable theory forces you to take into account apparently irrelevant, unchangeable information about your past wellbeing. So it is vulnerable to a version of the well-known *Egyptology* objection to non-separable theories in population axiology.[47] At the risk of polluting the English language, let's call the prudential version of this objection the *Childhoodology* objection.[48]

Suppose you are coming to the end of your life, and you have a choice to make. You can either take a risky treatment for your otherwise terminal illness, which will be somewhat painful but which might, if you are lucky, give you a few additional years of good life. Alternatively, you can choose to take palliative care and die in a pain-free, dignified way. The prudential stakes might be balanced in such a way that, choosing on a whole-lifetime basis on the diminishing marginal utility view, whether you ought to take the

[47]See McMahan (1981: 115), Parfit (1984: 420).

[48]I owe this excellent name (which I do like, or I would not have chosen it) to Johan Gustafsson.

risky treatment or the palliative care depends on how your earlier life went. For example, it might depend on how much you enjoyed your childhood years which, unfortunately, you cannot now remember in much detail.[49] The whole-lifetime diminishing marginal utility view might then suggest that you peruse some old photographs of your childhood in order to jog your memory. But surely this cannot be what you ought to do. The facts about your forgotten childhood cannot be relevant to deciding between a risky treatment and palliative care.

6.6.7 Summary

I have argued that we should accept Time-Separability. And we have seen that Time-Separability implies Risk-Neutrality (that is, the coincidence of the EUT and life-years scaled of wellbeing).

Given Risk-Neutrality, the argument of §6.5 does not contradict the arguments of §6.3 and §6.4 after all. They are instead complementary. §6.3 and §6.4 showed that the EUT version of Fission Totalism is practically inescapable, given Fission Prospect Separability. §6.5 argued for the life-years version of Fission Totalism. Given Time-Separability, these two versions of

[49] I assume that, even if you can't remember your childhood years, they still affect your entire-lifetime prudential value. If you doubt this, substitute with some other case where your past wellbeing affects the entire-lifetime prudential value of your present person-stage, but the present person-stage cannot remember this past wellbeing. There must be at least some such cases: to borrow from Parfit (1984: 205), I cannot remember putting on my shirt yesterday, but anything that was bad for me-then would surely be bad for me-now in the whole-lifetime sense.

Totalism amount to the same thing.

That should not be a surprise, since Time-Separability and Fission Prospect Separability are also closely related. Given that Identity Does Not Matter, undergoing a fission operation into two fission offspring who live one after the other looks just like getting ordinary survival with one life after the other. If Time-Separability fails, then what happens in the earlier segment of the life (analogously, to the earlier fission offspring) affects the prudential value of prospects involving the later segment of the life (analogously, the later fission offspring), and the converse is true as well.

To summarise, the case for the EUT version of Fission Totalism seems to come down to separability. Either Time-Separability or Fission Prospect Separability will do, and we should expect that these principles stand and fall together. The case for the life-years version of Fission Totalism, meanwhile, seems to me to instead rest on the rather weaker principle of Prudential Ex Post Pareto. If we were to abandon even this principle, we may yet be left with a third, weaker form of Fission Totalism: the aggregate life version.

6.7 Objections to Fission Totalism

I shall now address two objections to Fission Totalism.

6.7.1 The Sequential Fusion Objection

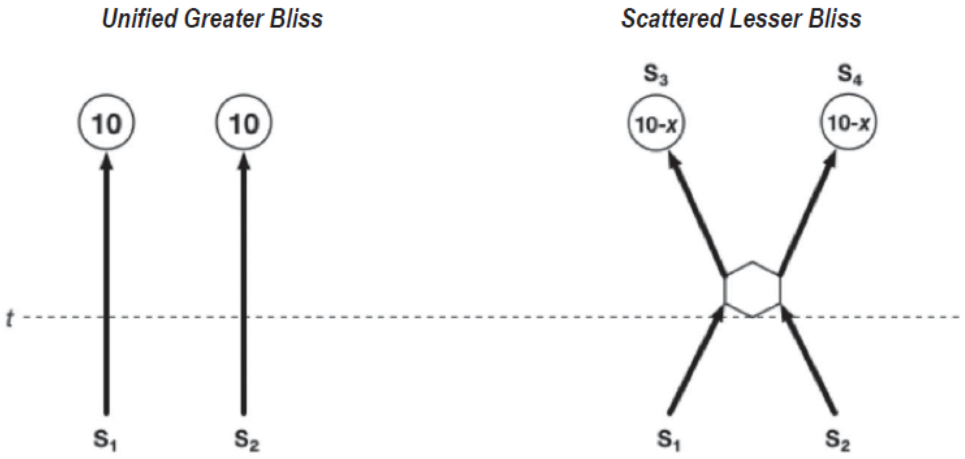
The first objection is due to Ross (2014). It has to do with what Fission Totalists should say about *fusion* cases: cases in which, rather than one person splitting into two or more fission offspring, two or more distinct fusion parents instead combine into one (or, I suppose, it could be more than one) fusion offspring. As in the case of fission, identity is not preserved, but the relation of prudential concern remains intact.

What does Fission Totalism say about fusion cases? Well, in fusion cases, two agents bear the full relation of prudential concern to a single future fusion offspring. If Lefty and Righty fuse into wholly, then Lefty and Righty do not care about (i.e, do not bear the relation of prudential concern to) each other, but they do both care about Wholly. Effectively, as far as each of Lefty and Righty is concerned, undergoing a fusion operation is just like undergoing a fission operation which results in only one future individual. That is, it is like undergoing an operation which severs the relation of personal identity, but leaves the relation of prudential concern untouched. Since Identity Does Not Matter, this is just as good as getting ordinary survival at Wholly's wellbeing level.

This view about fusion, when combined with the other verdicts of Fission Totalism, raises a problem. Roughly, the problem is that agents who successfully follow Fission Totalism may end up making themselves worse off in

sequential decision situations. Ross (2014: 254) considers the following case, which is reprinted below:

Figure 6.1: Unified or Scattered Bliss



In *Unified or Scattered Bliss*, there are two alternatives. In the first alternative, *Unified Greater Bliss*, person stages S_1 and S_2 get ordinary survival with 10 units of wellbeing. In the second alternative, *Scattered Lesser Bliss*, person stages S_1 and S_2 fuse and then instantaneously fission again into S_3 and S_4, who each get slightly less than 10 units of wellbeing.[50]

It seems that in this situation, S_1 and S_2 should have full concern for S_3 and S_4. Given this assumption, according to Fission Totalism, it would

[50]This is an unimportant departure from Ross's original example, which makes use of a "switcheroo" operation rather than fusion followed by fission. In a "switcheroo" operation, two qualitatively identical individuals have the left and right halves of their brains separated, and then fused with the opposite-side brain of their counterpart, resulting in two individuals, each of whom has the left half of their brain from one of the original individuals, and the right half of their brain from the other. See Ross (2014: 249).

be best for S_1 and S_2 to choose *Scattered Lesser Bliss*. This way, each will effectively get to survive as two fission offspring with slightly less than 10 units of wellbeing, i.e., a total of nearly 20 units of wellbeing, rather than getting ordinary survival with 10 units of wellbeing. However, Ross gives several arguments against the claim that it would be best for each of S_1 and S_2 to choose *Scattered Lesser Bliss* over *Unified Greater Bliss*.

Argument 1: It's Just Not Intuitive

The first argument is that it simply does not seem to be true that S_1 and S_2 could be made better off by effectively combining and then separating their body parts, at a cost of some total wellbeing (Ross, 2014: 255). But I do not see why we should share this intuition. S_1 and S_2 really *do* get something out of the deal: they get to bear the relation of prudential concern to two people rather than to one person, effectively surviving as two people rather than surviving as one person. It's not strange for this to make an important prudential difference.

Argument 2: Iterability

The second argument is that the problem can be iterated (Ross, 2014: 255–256). If it would be best for S_1 and S_2 to fuse and then separate into S_3 and S_4 at a lower wellbeing level, it would also be better for S_3 and S_4 to fuse and separate into S_5 and S_6 at a still lower wellbeing level, and so on

indefinitely. This would then seem to make Fission Totalism deficient by its own lights: if this process ends with the successors of S_1 and S_2 having less *total* wellbeing than S_1 and S_2 would have had individually, if they had they chosen *Unified Greater Bliss*, then successfully following Fission Totalism results in outcomes which are worse according to Fission Totalism.

However, merely pointing out these facts about betterness, taken in isolation, is not enough to show that the problem is genuinely iterable. In order to show that, we would need to provide an actual decision tree in which it would be best, at each node, for all participants to choose in such a way that we end up with two much worse off final fission offspring S_k and S_{k+1}.

The obvious decision tree to go for, and the one Ross presumably has in mind, is a decision tree in which S_1 and S_2 face a choice between Unified or Scattered Bliss, and then, if Scattered Bliss is chosen, S_3 and S_4 face a similarly structured choice between Unified or Scattered Bliss, and so on. This decision tree would show the iterability of the problem if one were to employ so-called *myopic* choice, in which one chooses without anticipating the choices of one's future person-stages (or, in this case, one's future fission offspring).[51] On this dynamic choice rule, S_1 and S_2 would fuse and then fission, because that is myopically best for them; next, S_3 and S_4 would fuse and then fission, because that is myopically best for them; and so on.

However, if one instead employs foresight in one's decision-making, which

[51] See Dow (1984: 96).

seems quite rational, this decision tree becomes non-iterable. Suppose for instance that S_1 and S_2, starting with 10 units of wellbeing each, may fuse and then fission into S_3 and S_4 at 7 units of wellbeing, and then a similar decision can produce S_5 and S_6 at 4 units of wellbeing. Myopically, each pair of fission offspring would want to go further in the process. However, S_1 and S_2 know that S_3 and S_4 would choose the second fusion-then-fission operation. S_1 and S_2 therefore know that the actual result of undergoing their own fusion-then-fission operation would be that they will eventually survive as S_5 and S_6, with 4 units of wellbeing each. So, employing foresight, S_1 and S_2 would choose to remain at 10 units of wellbeing each, rather than go through with their fusion-then-fission operation. So this decision tree does not show that the problem is iterable for agents who choose with foresight.

Argument 3: Non-Ratifiability

The third argument, which is adapted from its original presentation in Ross (2014: 255–257), goes as follows. If S_1 and S_2 were to choose *Scattered Lesser Bliss*, this choice would be *non-ratifiable* in the sense that "all the future person-stages that [bear the relation of prudential concern] to the person-stage making the choice will have quasi-self-interested reason to regret this choice." (Ross, 2014: 255) But, the correct theory of prudence (allegedly) should not recommend any non-ratifiable choices. So any theory of prudence which recommends *Scattered Lesser Bliss*, such as Fission Totalism, must be

false.

One might respond to this objection by claiming that the relation of prudential concern is asymmetric, extending forwards but not backwards, so that S_3 and S_4 do not bear the relation of prudential concern to either S_1 or S_2; see Karhu (forthcoming) for a detailed defence of this view. While the asymmetric concern view plus Fission Totalism does not lead to non-ratifiability in *Unified or Scattered Bliss*, it does, as Ross (2014: 256–257) notes, give rise to non-ratifiability in other cases. However, on the asymmetric concern view, *every* plausible theory of prudence will be non-ratifiable. To see this, consider a case in which S_1 and S_2 can either each experience a wellbeing level of $-1,000$, or they can fuse into a single individual S_3, who will get -1 units of wellbeing. Clearly, on every reasonable theory of prudence in fusion cases, fusing is in S_1 and S_2's interests. But on the asymmetric concern view, the choice to fuse would be non-ratifiable, since S_3 will not be prudentially concerned for either S_1 or S_2, and will regret her own negative wellbeing level. If the asymmetric concern view is true, then, ratifiability cannot be a criterion of rational choice.

I shall therefore assume in what follows that the asymmetric concern view is false, and in particular that fission and fusion offspring always have prudential concern for their fission parents, and for at least one of their fusion parents. On this assumption as well, I maintain that every reasonable theory of prudence in fission and fusion cases is non-ratifiable in Ross's sense. The

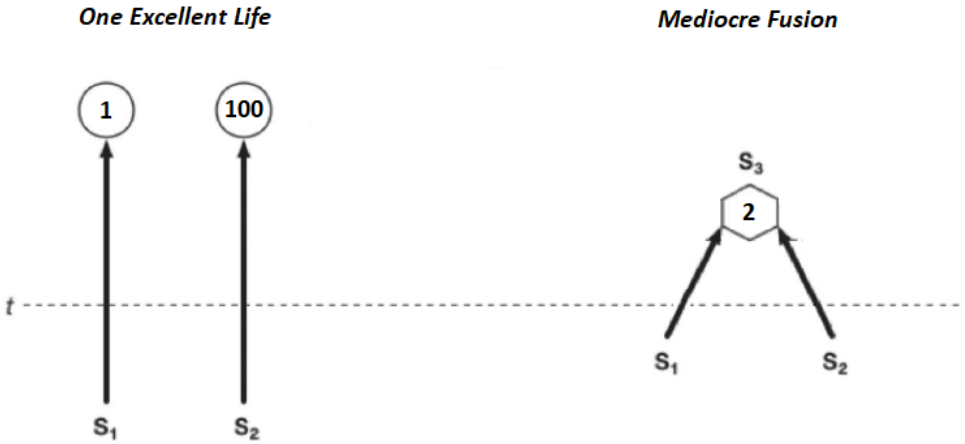
presence of non-ratifiability of this sort is therefore not a serious objection to any particular theory of prudence.

Here's a first argument to that effect. As Gustafsson and Kosonen (forthcoming) point out, two fission offspring might end up playing a game of Prisoner's Dilemma against each other. Provided each fission offspring lacks prudential concern for the other, as seems to be the case, both will then end up worse off if they both follow the recommendations of any plausible theory of prudence (that is, if both defect). Of course, the collective choice of both defecting will be regretted by both fission offspring.

This argument might seem to be not quite on target, since the choice between *Unified Greater Bliss* and *Scattered Lesser Bliss* may be being made by a single individual (for instance, S_1). One might respond by insisting that the ratifiability criterion should only apply to *individual* choices, not choices made collectively by multiple individuals.

However, even if we restrict the ratifiability criterion in this way, it remains the case that all plausible prudential theories are non-ratifiable. Consider the case represented by the diagram below.

Figure 6.2: To Fuse or Not to Fuse



In *To Fuse or Not to Fuse*, S_1 alone is able to choose between *One Excellent Life*, in which S_1 exists at wellbeing level 1 and S_2 exists at wellbeing level 100, and *Mediocre Fusion*, in which S_1 and S_2 fuse into S_3, who subsequently enjoys 2 units of wellbeing. Clearly, on any plausible theory of prudence in fission and fusion cases, it is in S_1's interests to choose *Mediocre Fusion*. However, equally clearly, S_3 will regret the choice of *Mediocre Fusion* on any plausible theory. For while S_1 lacks prudential concern for S_2, and therefore doesn't care about S_2's loss of 100 units of wellbeing in return for the mediocre fusion, by symmetry S_3 *must* have prudential concern for S_2[52]. On any plausible theory, the combined loss of S_1's one unit of well-

[52]More precisely, symmetry tells us that S_3 has prudential concern for S_2 if and only if S_3 has prudential concern for S_1. Since we have assumed that S_3 must have prudential concern for *some* fusion parent, it follows that S_3 has prudential concern for both S_1 and S_2.

being and (much more importantly) S_2's one hundred units of wellbeing is worse from S_3's perspective than the gain of two units of wellbeing from the fusion. S_3 therefore regrets the choice of *Mediocre Fusion*, and so S_1's choice is non-ratifiable.

Diagnosis

The deeper reason why non-ratifiability occurs in *To Fuse or Not to Fuse* is that the non-ratifiable choice is being made by S_1, for whom the relation of prudential concern does not extend to S_2, and then regretted by S_3, for whom the relation of prudential concern does extend to S_2. Similarly, in the original *Unified or Scattered Bliss* case, fusion-then-fission is preferable for both S_1 and S_2, since each lacks prudential concern for the other, but is regrettable for S_3 and S_4, since these individuals each have prudential concern for both S_1 *and* S_2.

The root cause of non-ratifiability in both cases is the fact that, although an earlier person stage might be fully prudentially concerned for a later fusion or fission offspring, this does not mean that the two agree on the matter of who to be prudentially concerned for. More precisely, let us say that two person-stages S_i and S_j are *equivalently concerned* if and only if, for any person stage S_k, S_i is prudentially concerned for S_k if and only if S_j is prudentially concerned for S_k; and moreover, S_i and S_j have the same degree of prudential concern for S_k. The reason non-ratifiability arises is that it is

possible for a person-stage to be in a situation where, although they might bear the relation of full prudential concern to many future person-stages, they are not equivalently concerned with any of them. It should not be a surprise that non-ratifiability is then possible: the decision-making person-stage has different aims to all her future person-stages, so why should there be agreement between them?

To illustrate the point, it might help to consider yet another case: Wholly is about to undergo fission into Lefty and Righty, and can choose whether to give 10 units of wellbeing to Lefty and -1000 to Righty, 10 units to Righty and -1000 to Lefty, or 9 units of wellbeing to both. It is obviously in Wholly's interests to give 9 units of wellbeing to both. But both Lefty and Righty will regret this choice, and wish that they had instead got 10 units of wellbeing. Why? Because Lefty and Righty both have different aims to Wholly. While Wholly aims to make both Lefty *and* Righty better off, Lefty and Righty instead have the aim to make only *themselves* better off.

The upshot is that the ratifiability criterion should not apply to all of those future person-stages towards whom one bears the relation of full prudential concern. Instead, it should apply only to those future person-stages with whom one is equivalently concerned. This is a subtle distinction, and it is easy to see why it has been missed. If personal identity is what matters, then one is fully prudentially concerned for another person stage if and only if one is personally identical with that person stage. Since the personal

identity relation is symmetric and transitive, it follows that one person stage is fully prudentially concerned for another if and only if the two stages are equivalently concerned. So, if personal identity is what matters, the full prudential concern relation and the equivalent concern relation are co-extensive. But this is not true on the view that Identity Does Not Matter; on this view, we need to be careful to distinguish between full prudential concern and equivalent concern.

6.7.2 The Fission Repugnant Conclusion

The second objection to Fission Totalism is that it implies a prudential version of the Repugnant Conclusion. According to the Fission Repugnant Conclusion, rather than getting ordinary survival with an excellent life, it would be better to instead split into a sufficiently large number of fission offspring, each of whom would have a life only barely worth living. In fact, Fission Totalism implies an especially "repugnant" conclusion of this sort, on which it would be better to split if the combined wellbeing of one's fission offspring were only *slightly* higher.

Most people find variants of the Repugnant Conclusion hard to believe. However, it is easier to refuse to believe a principle than to provide a sensible theory which avoids it. As we have seen, the cases of population axiology and the axiology of prudence in fission cases are precisely structurally analogous.

This means, in particular, that all impossibility theorems and other results pertaining to the difficulty of avoiding the Repugnant Conclusion also apply to the case of fission. I won't repeat those arguments in this chapter; suffice it to say that, on balance, accepting the Fission Repugnant Conclusion seems to me less bad than the logically possible alternatives.

6.8 Conclusion

I have argued for Fission Totalism on the EUT scale of wellbeing, and also on the life-years scale of wellbeing. I have also argued directly for Risk-Neutrality: the claim that these scales coincide. Putting these claims together, we have a view on which, provided no incomplete holistic goods are at stake and there are no relations of partial prudential concern, the prudentially best prospect is the one which maximises the expected sum of wellbeing, measured on the life-years scale, of each non-overlapping person-stage for whom one has full prudential concern. This view has a natural extension to the case of partial prudential concern: maximise the sum of expected life-years wellbeing of each person stage, weighted by the extent to which one has prudential concern for that person-stage. I suspect that this extended view is also true, though I have given no arguments for it here.

I think that the arguments I have given for Fission Totalism are quite compelling *if* the assumptions I have made are sound. But which of them

are least secure? Setting aside the well-trodden ground of rejecting transitivity, I think two possibilities for avoiding Fission Totalism are particularly noteworthy.

The first is to reject Prudential Ex Post Pareto. This might at first seem like a non-starter: it might even seem conceptually confused to deny that the value of undergoing a fission operation supervenes on the prudential values obtained by one's fission offspring. But perhaps doing so has more going for it than meets the eye. It might be prudentially rational to be concerned with the overall pattern of the lives of our fission offspring in a way that allows for interrelations between the shapes of these lives. For example, if our attitude to wellbeing over time is asymmetrical, so that it is important for our lives to have an overall upward curve of wellbeing, it might be that the presence of one fission offspring with a strongly upward-curving life lessens our prudential reasons to prefer upward curves in the lives of other fission offspring.[53] (On the other hand, this view would seem to be vulnerable to the objections I provided in §6.6.6 against theories which are not Time-Separable.) If we were to reject Prudential Ex Post Pareto, then of course Fission Totalism, on the EUT and life-years scales of wellbeing (or on any scale), would be false. But as far as I can see, there would be no particular challenge to my argument for the aggregate life version of Fission Totalism.[54]

[53]See for instance Velleman (1991).

[54]With one caveat: I suppose if Time-Separability were to be denied, there may be some issues concerning the determination of the appropriate temporal order in which to

The second possibility is to reject Fission Prospect Separability. As we have seen, this principle is tightly bound up with Time-Separability, so this would likely mean denying both separability principles. If one were to do this, one would certainly deny Fission Totalism on the EUT scale of wellbeing. But, as far as I can see, rejecting these principles would provide no reason to doubt my argument for Fission Totalism on the life-years scale of wellbeing.

construct the aggregated life.

Chapter 7

Concluding Arguments

Abstract

This chapter summarises the results of the previous chapters and provides several direct arguments for Totalism in population axiology. The first is an argument from the Neutral Addition Principle and Same-Number Totalism, akin to the argument for Fission Totalism given in sections 6.3 and 6.4 of this thesis. The second is an extension of the argument from Different-Number Egalitarian Dominance given in §2.4. The third and final argument is a fully intrapersonal argument which appeals to principles of the sort considered in Chapter 5.

7.1 Taking Stock

Throughout this thesis, I have argued for various aspects of Totalism:

– In *Anonymity and Non-Identity Cases*, I argued for the principle of Anonymity, on which changes in identity make no all-things-considered difference to evaluations of populations.

– *The Welfare Diffusion Objection to Prioritarianism* argued against axiological Prioritarianism, implicitly defending the Totalist position that a given unit of wellbeing makes the same evaluative difference no matter who receives it.

– *In Favour of Making Happy People* argued against the Principle of Neutrality, on which creating happy people never makes an outcome better. Put another way, it defended the controversial Totalist claim that creating happy people is one way of making the world a better place.

– *Repugnance Without Mere Addition* and *Intrapersonal Arguments for the Repugnant Conclusion* provided, respectively, a new interpersonal and a new intrapersonal argument for the Repugnant Conclusion. These arguments help to rebut probably the most common objection to Totalism: that it implies the allegedly false Repugnant Conclusion. If the Repugnant Conclusion is *true*, as these arguments suggest, then the

fact that Totalism implies this true proposition is no objection to the view.

– Lastly, *Prudence in Different-Number Fission Cases* provided several arguments for Totalism as a view about prudence in cases of fission.

I have not, however, focused on giving direct arguments *for* Totalism in population axiology. Several arguments for Totalism are stated in, can be inferred from, or are closely related to the material covered in earlier chapters of this thesis. In the interests of collecting these arguments together into one place, I shall present these arguments in this chapter.

To avoid complicating the discussion, I shall assume throughout that prudential (or "personal") betterness and moral or overall betterness are both transitive and option-set-independent. And to avoid the usual discussion of formalism, I shall generally re-use notation from earlier chapters without explanation. There will be one small exception: as we learned in *Prudence in Different-Number Fission Cases*, it sometimes matters which scale of wellbeing is implicitly being talked about. I shall be non-committal about the choice of wellbeing scale *until* it matters for a particular argument, which it sometimes will.

A final caveat. In previous chapters, I have tried to avoid making claims that are not warranted by the evidence. Here I shall relax this policy somewhat. I shall, so to speak, move somewhat away from the lectern and some-

what towards the soapbox. The point of what follows is not to provide knock-down arguments for Totalism which everyone should accept. Rather, the point is to give a broad picture of what I take to be some of the best reasons to be a Totalist.

7.2 The Neutral Addition Argument

The first argument is exactly analogous to the argument for Fission Totalism on the EUT scale given in Chapter 6. It appeals to the following two principles:

> *Neutral Addition* If X and Y are disjoint populations and Y contains only neutral lives, then $X + Y \sim X$.

> *Same-Number Totalism* If X and Y are same-number populations, then $X \succeq Y$ if and only if X contains at least as much total wellbeing as Y.

From these two principles, we can conclude:

> *Totalism* For any populations X and Y, $X \succeq Y$ if and only if X contains at least as much total wellbeing as Y.

Let me briefly mention some reasons to accept these two premises.

7.2.1 Neutral Addition

In my view, the best reason to accept the principle of Neutral Addition is that it follows from

> *The Equivalence of Personal and Moral Contributive Value*[1] Other things being equal, the addition of a life
>
> – makes a population better if and only if the life is at a good well-being level,
>
> – makes a population worse if and only if the life is at a bad well-being level, and
>
> – leaves the value of the population unchanged if and only if the life is at a neutral well-being level.[2]

The Equivalence of Personal and Moral Contributive Value captures a compelling idea about the value of populations: when only one person is affected and all else is equal, doing something that is good for the person makes things better, doing something bad for the person makes things worse, and doing something neutral for the person makes no evaluative difference.

There are, broadly speaking, two ways of denying the Equivalence of Personal and Moral Contributive Value. The first is to deny that the the moral

[1] This principle is taken from Gustafsson (2020: 87).

[2] We also need to assume that there *are* neutral lives, and that these correspond to level 0 on our chosen wellbeing scale. While Gustafsson (2020) accepts the Equivalence of Personal and Moral Contributive Value, he denies these further claims.

betterness relation is "personal" in any way which would lead us to expect that prudential value and moral value match up. For instance, a proponent of the Average view or a Variable Value view might be interested in average wellbeing or a hybrid of average and total wellbeing *as such*, rather than being interested in average or hybrid wellbeing as a mere derivative consequence of being interested in the wellbeing of people. Of course, Totalism itself can also be construed in this impersonal way: one might subscribe to what Parfit (1984: 400) calls the "Milk-Production Model" of morality. On this view, one is concerned with total wellbeing as such, not with the wellbeing levels of people, just as a person interested in maximising the total quantity of milk is concerned with how much milk there is, not with how much milk is held by each individual container.[3]

The second way of denying the Equivalence of Personal and Moral Contributive Value, which seems to me more promising, involves two moves. The first is to appeal to *Existence Non-Comparativism*: the view that an outcome in which a person exists can never be better or worse for her than an outcome in which she does not exist. The next move is to claim that the personal view of moral betterness is best captured by

[3]Parfit's "Milk-Production Model" is a play on the "Steam-Production Model" of morality he credits to MacKaye (1906).

The Equivalence of Personal and Moral Contributive Comparative Value[4]

Other things being equal, the addition of a life

- makes a population better if and only if the life is better than non-existence,
- makes a population worse if and only if the life is worse than non-existence, and
- leaves the value of the population unchanged if and only if the life is equally as good as non-existence.

Furthermore, the same is true regarding additions of multiple lives, all of which are better, worse, or equally as good as non-existence respectively.

Those who accept the Equivalence of Personal and Moral Contributive Comparative Value seem to have done enough to avoid the charge of being objectionably impersonal. However, this way of rejecting the Neutral Addition Principle comes with some serious extensional drawbacks. The conjunction of Existence Non-Comparativism and The Equivalence of Personal and Moral Contributive Comparative Value will imply the

Extreme Evaluative Principle of Neutrality If A and B are any disjoint populations, then A and $A + B$ are incomparable.

[4]A principle like this is stated by Rabinowicz (2009: 391) (who seems to accept Existence Comparativism), and has more recently been endorsed by Bader (2022a: 18–19; 2022b: 263) (who does not).

The Extreme Evaluative Principle of Neutrality immediately implies that adding arbitrarily many arbitrarily tortured lives to a population does not make things worse. Given minimal principles for making same-person comparisons, it also implies that transforming a population of excellent lives, into a larger population of equally tortured lives, including the whole original population, does not make things worse either.[5]

It might be thought that these sorts of problems can be escaped if we accept an *asymmetric* view on the Equivalence of Personal and Moral Contributive Comparative Value. But, setting aside the difficulties associated with giving a proper theoretical foundation for such a view, we have seen already in Chapter 3 that such views will not be extensionally plausible. In particular, they will still violate the

> *Absolute Value Principle* If X is a population consisting only of good lives, and Y is a population consisting only of bad lives, then X is better than Y.

A view which countenances different-number comparisons in general, but does not satisfy the Absolute Value Principle, seems objectionably impersonal. So, it seems that those who insist on rejecting Neutral Addition are

[5]Consider a population X consisting of ten billion people at level 100. First, add ten trillion people at level $-10,000$, resulting in $X + Y$. Finally, let Z consist of the $X + Y$ people, all at level $-1,000$. According to the Extreme Evaluative Principle of Neutrality, $X + Y$ is not worse than X. Clearly, Z is better than $X + Y$. Hence Z is not worse than X either.

left with two options. They might embrace an impersonal moral theory, denying there is a tight link between what is good or bad for people and what is good or bad, morally speaking. Or, they might replace the Equivalence of Personal and Moral Contributive Value with its Comparative equivalent, at the cost of being left with what many would consider to be an extensionally implausible moral theory.

7.2.2 Same-Number Totalism

The best argument for Same-Number Totalism is John C. Harsanyi's (1955) Aggregation Theorem; more precisely, the version of this argument I shall discuss is Theorem 2 of Fleurbaey (2009: 301). I discussed the prudential analogue of this theorem in §6.4, so I won't repeat the details here. I shall instead simply state the result. Assuming some domain conditions which we shall pass over without comment, it can be shown that Same-Number Totalism, on the EUT scale of wellbeing, follows from the following four claims:

> *Expected Utility Theory* The prudential betterness relation on prospects satisfies the axioms of Expected Utility Theory.

> *Anonymity* If populations X and Y are anonymously equivalent, then they are equally good.

Moral Ex Ante Pareto (Same-Person) Let X and Y be any population prospects which guarantee the existence of the same set of individuals. If, for all existing individuals i, $X \succeq_i Y$, then $X \succeq Y$. If, additionally, $X \succ_i Y$ for some i, then $X \succ Y$.

Moral Statewise Dominance (Same-Person) Let X and Y be any population prospects over the same set of states of nature S, which guarantee the existence of the same set of individuals. If, for each $s \in S$, $X(s) \succeq Y(s)$, then $X \succeq Y$. If additionally $X(s) \succ Y(s)$ for some s, then $X \succ Y$.

I won't discuss whether we should accept Expected Utility Theory here (though I think we should), and I argued extensively for Anonymity in Chapter 1. Assuming these two principles, Same-Number Totalism follows from Moral Ex Ante Pareto and Moral Statewise Dominance.

This is an immensely important argument. I would go so far as to say that it is the most important argument in value theory. That is because Moral Ex Ante Pareto and Moral Statewise Dominance are extremely compelling premises, while Same-Number Totalism is an extremely controversial conclusion.

We should not reject Moral Statewise Dominance, for two main reasons. First, because doing so would generate a preference structure that is at odds

with rationality; we can see this by the fact that such a preference structure is exploitable.[6] Second, because you should care ultimately about final outcomes, and only derivatively about prospects: violating Moral Statewise Dominance would put the cart before the horse.[7]

That leaves us with Moral Ex Ante Pareto. The main reason to accept this principle is, once again, that failing to do so constitute an objectionable kind of impersonalism.[8] Note that Moral Ex Ante Pareto follows from its restriction to the case where only one person is affected (cf §6.4.5). To deny Moral Ex Ante Pareto, then, is to deny

> *The Equivalence of Personal and Moral Contributive Value, Part 2*
> Other things being equal, a change in a person's prospects
>
> – makes a population better if and only if the new prospect is better for the person than the old prospect,
> – makes a population worse if and only if the new prospect is worse for the person than the old prospect, and
> – leaves the value of the population unchanged if and only if the the new prospect is equally as good for the person as the old prospect.

[6] See Gustafsson (nd).

[7] For a detailed argument along these lines, see Schoenfield (2014).

[8] This might not seem too bad, because it might not seem implausible that there are impersonal values, such as the value of biodiversity, of cultural or scientific achievements, or perhaps an impersonal value of the survival of humanity (Frick, 2017). But remember we are comparing the value of *populations*, that is, we are interested in value when *all else is equal* except for people's wellbeing. It is here that impersonal moral value seems most objectionable.

The main reason to accept this principle is that it is obviously true – just look at it! But let me add a further argument which, admittedly, is probably not more compelling than the principle itself. It goes as follows.

If one prospect would be better for you than another, and all else is equal, you are rationally obligated to choose the first prospect. Similarly, you are morally obligated to bring about a morally better prospect rather than a worse one, provided all else is equal. Now imagine that personal and moral contributive value come apart in the sense of Part 2. It might then be rationally obligatory to choose a prospect which is better for yourself, but morally obligatory to choose a prospect which is worse for yourself. Since it is rationally obligatory to choose the prudentially better prospect, following your moral obligation would be irrational. But following your moral obligations cannot be irrational. So prudential and moral contributive value cannot come apart in this way.

7.3 The Argument from Different-Number Egalitarian Dominance

This second argument comes from Chapter 2. It is most compelling when understood on the life-years scale of wellbeing. It appeals to three premises. The first two are lifted directly from Chapter 2:

Different-Number Egalitarian Dominance Let X and Y be any populations. If

(i) X is a perfectly equal non-empty population of good or neutral lives;

(ii) each person in X is at least as well off as each person in Y;

(iii) each person in X exists in Y (and is therefore at least as well off in X as in Y);

(iv) X has at least as much total wellbeing as Y,

then X is at least as good as Y.

*Mere Addition** For any populations X and Y, if Y consists only of good or neutral lives, then $X + Y$ is not worse than X.

The third premise is an implication of the Pigou-Dalton principle used in Chapter 2, but has the advantage of being easier to state and understand:

Non Anti-Egalitarianism[9] If populations X and Y contain exactly the same people, everyone in X is equally well off, and X has at least as much total wellbeing as Y, then X is at least as good as Y.

I won't repeat the argument from §2.4, only state the result, which is that these three principles, plus Anonymity, jointly imply

[9] As far as I know, this principle was first named and stated by Ng (1989: 238).

Totalism for Good Populations Suppose that non-empty populations X and Y contain only good or neutral lives. Then X is at least as good as Y if and only if X contains at least as much total wellbeing as Y.

7.3.1 Different-Number Egalitarian Dominance

We should accept Different-Number Egalitarian Dominance because it is intuitively compelling on the life-years scale of wellbeing. The principle applies when, comparing X and Y, we find that X is perfectly equal, makes everyone in Y at least as well off and has at least as much total wellbeing as Y. Intuitively X is then at least as good as Y, since everything that seems like it could matter looks to be either equal between the two, or in X's favour. This kind of justification might not seem very principled, since it does not explain exactly *what* about X makes it at least as good. It rather throws as many axiological considerations as possible at the comparison, hoping that some of them will stick. Still, it seems to me that they *do* stick.

7.3.2 Mere Addition

Since Mere Addition* is an implication of Neutral Addition, everything which supports Neutral Addition also supports Mere Addition*. But there is something additional which can be said for Mere Addition*. As we saw, there is a way of holding onto a personal view of moral value while denying Neutral

Addition, namely to deny The Equivalence of Personal and Moral Contributive Value in favour of its Comparative counterpart. However, since Mere Addition* asserts *non-worseness* rather than betterness, the Comparative principle also implies Mere Addition*. Because of this, as far as I can see there is no way whatsoever to deny Mere Addition* without thereby adopting an objectionably impersonal moral theory.

7.3.3 Non Anti-Egalitarianism

Once again, the best argument for Non Anti-Egalitarianism might just be that it is a compelling principle, on the life-years scale of wellbeing. Sheer intrinsic plausibility seems a sufficient reason to accept it. It can also be argued for, though once more, it is hard to find premises which are much more compelling than Non Anti-Egalitarianism.

The first argument for Non Anti-Egalitarianism is the one that was implicit in §2.4: it follows from the version of Pigou-Dalton which says that pure non-rank-switching transfers of wellbeing from the better-off to the worse-off always make an outcome at least as good. This version of Pigou-Dalton seems to me obviously true, but I am not sure that it is more obviously true than Non Anti-Egalitarianism.

A second argument for Non Anti-Egalitarianism is intrapersonal, and works on the EUT scale of wellbeing. It goes via a version of the Harsanyi

Aggregation Theorem with weaker versions of the Moral Ex Ante Pareto and Moral Statewise Dominance premises. In particular, we can weaken these premises in the following way:

> *Moral Ex Ante Pareto (Egalitarian, Same-Person)* Let X and Y be any population prospects over the same set of states of nature S, which guarantee the existence of the same set of individuals. Suppose that, for each $s \in S$:
>
> (i) $X(s)$ is perfectly equal,
>
> (ii) X gives each person the same expected wellbeing, and
>
> (iii) $X \succeq_i Y$.
>
> Then $X \succeq Y$.
>
> If, additionally, $X \succ_i Y$ for some i, then $X \succ Y$.

> *Moral Statewise Dominance (Egalitarian, Same-Person)* Let X and Y be any population prospects over the same set of states of nature S, which guarantee the existence of the same set of individuals. Suppose that, for each $s \in S$:
>
> (i) $X(s)$ is at least as equal as $Y(s)$: there exists some population X' such that X' is anonymously equivalent to $X(s)$ and $Y(s)$ can

be obtained from X' via a (possibly empty) series of pure non-ranking-switching transfers of wellbeing from better-off to worse-off.

(ii) X gives each person the same expected wellbeing, and

(iii) $X(s) \succeq Y(s)$.

Then $X \succeq Y$.

If additionally $X(s) \succ Y(s)$ for some s, then $X \succ Y$.

To see how the argument works in the two-person case, consider the three population prospects X, Y and Z illustrated by the tables below.

X	s_1	s_2
Adam	100	100
Eve	0	0

Y	s_1	s_2
Adam	100	0
Eve	0	100

Z	s_1	s_2
Adam	50	50
Eve	50	50

In each state of nature s_i, $Y(s_i)$ and $X(s_i)$ are anonymously equivalent. Therefore, each $Y(s_i)$ is at least as equal as $X(s_i)$, and additionally $Y(s_i) \succeq X(s_i)$ by Anonymity. Furthermore, Y gives each person the same expected

wellbeing. The egalitarian version of Moral Statewise Dominance therefore implies that Y is at least as good as X.

Z guarantees perfect equality in each state of nature, gives each person the same wellbeing, and gives each person at least as much (in fact, the same amount of) expected wellbeing as Y. Therefore, the egalitarian version of Moral Ex Ante Pareto implies that Z is at least as good as Y. Putting these claims together via transitivity, we can conclude that Z is at least as good as X.[10]

This argument can of course be generalised. We can move from any unequal distribution of wellbeing to an equal distribution over the same people, with at least as much total wellbeing, in two steps. First, we move from the unequal population to a population prospect in which everyone gets a $\frac{1}{n}$ chance of getting each wellbeing position in the initial distribution (where n is the number of people). By Anonymity and the egalitarian version of Moral Statewise Dominance, this prospect will be at least as good as the initial population. Next, we move from the prospect in which each person gets an equal chance at each wellbeing position to one in which everyone gets the average wellbeing level (or a greater level than this) for sure. This population will be at least as good for each person, and will have perfect ex post and ex ante equality. Therefore, it will be at least as good according to

[10]Technically, we need the principle known as Certainty Equivalence, discussed in Chapter 5, in order to conclude that the population guaranteed by Z is at least as good as the population guaranteed by X.

the egalitarian version of Moral Ex Ante Pareto.

The advantage of this argument over the standard Harsanyi theorem given earlier is that the egalitarian versions of Moral Ex Ante Pareto and Moral Statewise Dominance are unimpeachable on Egalitarian and Prioritarian grounds. Some philosophers believe that it is worth embracing an ex ante impersonal view (or, less plausibly, that it is worth sacrificing standard principles of rationality like Statewise Dominance) in order to maintain that considerations of equality or priority have axiological significance.[11] Such moves do not undermine the egalitarian versions of these principles.

Note that the intrapersonal argument supports Non Anti-Egalitarianism on the EUT scale of wellbeing. But recall that Different-Number Egalitarian Dominance is most compelling on the life-years scale. What happens if we put together the EUT version of Non Anti-Egalitarianism, Mere Addition* and the life-years version of Different-Number Egalitarian Dominance? Technically, not much. We do not get Totalism, even for good populations. If wellbeing on the EUT scale is a strictly convex function of wellbeing on the life-years scale, then the argument will not go through. But this is not a very plausible situation: it would mean that we are prudentially required to be risk-seeking with respect to wellbeing on the life-years scale. For example, views of this sort might imply that it is better to get a 50-50 gamble yielding

[11] See for example Rabinowicz (2002) for the first view, and McCarthy (2006) for an exposition of the second view.

either 100 years of good life or nothing, rather than 60 years of good life for sure. If we assume that additional years of good life do *not* have increasing marginal utility in this way, then the life-years version of Different-Number Egalitarian Dominance should be enough to give us Totalism for Good Populations.[12]

7.3.4 Extending the Argument

Can we move from Totalism for Good Populations to Totalism writ large? Not without further principles. Suppose, for instance, we accept Same-Number Totalism. In that case, we can immediately extend our result to

> *Totalism for Good-On-Average Populations* Suppose that non-empty populations X and Y have non-negative total wellbeing. Then X is at least as good as Y if and only if X contains at least as much total

[12]The only condition in Different-Number Egalitarian Dominance for which the choice of wellbeing scale (beyond choice of the neutral level) matters is condition (iv), that X, a perfectly equal and possibly larger population, has at least as much total wellbeing as Y; we can also assume that Y is perfectly equal, since this is the case as the principle is applied in the argument for Totalism for Good Populations.

Suppose that (iv) holds of the EUT scale but not of the life-years scale. In that case, the sum of the *utilities* in the smaller population is at least as great as the sum of utilities in the larger population, but the same is not true of life years. Therefore, life years must have increasing marginal utility: the smaller sum of life years in the smaller population transforms into a greater sum of utility, because it is more concentrated. Assuming that marginal utility does not increase in this way, then, (iv) must hold of the life-years scale whenever it holds of the EUT scale. This means we can derive the EUT version of Different-Number Egalitarian Dominance from the life-years version (with the additional condition that Y is perfectly equal): whenever condition (iv) for the EUT version is satisfied, condition (iv) must also be satisfied for the life-years version, and hence the relevant pairwise comparison will hold.

wellbeing as Y.

We can go further if we strengthen the principle of Different-Number Egalitarian Dominance so that condition (iii), which requires that each person in the dominating population exists in the dominated population, no longer operates.[13] Together with Same-Number Totalism, this strengthened Dominance principle will then imply that any population with negative total wellbeing will be worse than any population with non-negative total wellbeing.

To see this, note that according to Same-Number Totalism, each population with positive total wellbeing is equally as good as an equal population in which everyone has a good life, and similarly for populations with negative or neutral total wellbeing. if X is an equal population of bad lives, and Y is an equal population of neutral or good lives, then Y has more total wellbeing than X, and each person in Y is better off than each person in X. Y must therefore be at least as good as X by our strengthened Dominance principle.

Next, note that by Same-Number Totalism, X is worse than the same-person population X' which assigns to each person half their negative wellbeing level in X. Since we have shown that *every* equal population of good or neutral lives is better than *every* equal population of bad lives, we can conclude that Y is at least as good as X'; since X is worse than X', we can

[13]Given Anonymity, this just means that we allow Different-Number Egalitarian Dominance to apply when X is a larger population than Y.

conclude that X is worse than Y.

This tells us that populations with non-negative total wellbeing must be ranked according to Totalism, and that populations with negative total wellbeing rank below population with non-negative total wellbeing (also in line with Totalism). But it leaves open how to rank populations with negative total wellbeing. We can get the full Total view here if we strengthen our assumptions in two respects. First, we can again strengthen Different-Number Egalitarian Dominance so that the dominating population is no longer required to consist solely of good or neutral lives. (If this sounds like it will make the principle much less plausible, note that the dominating population is still required to have higher personal and total wellbeing than the dominated population.) Second, we assume the

> *Negative Addition Principle* If X is any population, and Y is any population consisting solely of bad or neutral lives, then X is at least as good as $X + Y$.[14]

We can now complete the argument for Totalism. Suppose X and Y are populations with negative total wellbeing. If X and Y are the same

[14]This principle might seem overly strong in that it implies not only that additions of bad lives are for the worse, but also that additions of neutral lives are worse or equally good (thus cannot be incomparable). Without this feature of the principle, we would (as far as I know) be unable to prove full-strength Totalism. However, this is only because we would be unable to prove that two populations with the same negative total wellbeing, but different sizes, are equally good; it might be that one is better than the other. This would not matter much, since we would still know that any slight improvement to either of the two competing populations would render the improved population better. As I noted in Chapter 1, this is the distinctive feature of equal goodness.

size, then Same-Number Totalism implies they should be ranked according to total wellbeing. So assume, without loss of generality, that X contains more people than Y. By Same-Number Totalism, we can assume that X and Y are perfectly equal and everyone existing in Y exists in X. There are then three cases:

(i) X contains more negative total wellbeing than Y.

(ii) X contains less negative total wellbeing than Y.

(iii) X and Y have the same total wellbeing.

Consider first case (i). There is a population X', with the same people as X, in which everyone who exists in Y has the same wellbeing levels as they have in that population, and everyone else has the excess negative total wellbeing in X spread evenly between them. By Same-Number Totalism, X' is equally as good as X. The Negative Addition Principle and Same-Number Totalism imply that X' is worse than Y; hence, X is worse than Y.[15]

Next, consider case (ii). Our strengthened principle of Different-Number Egalitarian Dominance immediately implies that X is at least as good as Y. This is because X is perfectly equal, each person in X is better off than each person in Y, and X contains at least as much total wellbeing as Y. Now let X' be just like X, but with half the difference in negative wellbeing between

[15]This is because, if we construct X'', a version of X' in which the additional people have slightly less negative wellbeing, then the Negative Addition principle implies that $Y \succeq X''$; but Same-Number Totalism implies that $X'' \succ X'$, hence $Y \succ X'$.

X and Y added to each person. From what went before, we know that X' must be at least as good as Y. But X' is worse than X by Same-Number Totalism. Hence, X is better than Y.

Finally, consider case (iii). Strengthened Different-Number Egalitarian Dominance immediately implies that X is at least as good as Y. We can construct a new population X' in the manner of case (i) which is equally as good as X; by the Negative Addition Principle, Y is at least as good as X'. From this we can conclude that X and Y must be equally good. This completes the case for Totalism, since we have considered all possibilities for the two populations to be compared.

To summarise, the strengthened principle of Different-Number Egalitarian Dominance, the Negative Addition Principle, Mere Addition* and Same-Number Totalism imply full-strength Totalism.

7.4 The Fully Intrapersonal Argument

7.4.1 The Big Picture

Our third argument for Totalism is fully intrapersonal. It appeals to principles about prudence in the sorts of risky existence cases we considered in detail in Chapter 5. We first need some slightly stronger versions of the most important premises of the Harsanyi Aggregation Theorem:

Moral Ex Ante Pareto (risky existence) Let X and Y be any population prospects. If, for all individuals i with a positive probability of existence in either X or Y, $X \succeq_i Y$, then $X \succeq Y$. If, additionally, $X \succ_i Y$ for some i, then $X \succ Y$.

Moral Statewise Dominance (full strength) Let X and Y be any population prospects over the same set of states of nature S. If, for each $s \in S$, $X(s) \succeq Y(s)$, then $X \succeq Y$. If additionally $X(s) \succ Y(s)$ for some s, then $X \succ Y$.

Additionally, we will help ourselves to Anonymity. Finally, we need a principle which, for reasons which will soon become clear, I shall call

Risky-Existence Totalism Let X and Y be any population prospects, and let i be an individual with positive probability of existence in both X and Y. $X \succeq_i Y$ if and only if, treating the outcomes in which i does not exist equivalently to existence at zero wellbeing, the expected wellbeing of i in X is greater than or equal to the expected wellbeing of i in Y.

Obviously, this sort of principle could only be true on the EUT scale of wellbeing, so we shall use this scale throughout this section. In particular, we shall assume that Expected Utility Theory is true for prospects guaranteeing existence.

We will now see that Risky-Existence Totalism, Anonymity, the risky existence version of Moral Ex Ante Pareto and full-strength Moral Statewise Dominance imply population-axiological Totalism. I shall show how this works in one concrete case to give an idea of how it goes, before explaining how to generalise the argument.

Suppose that we are comparing two populations, X and Y, which involve three possible people. We can write a vector (w_1, w_2, w_3), where w_i can be a wellbeing level or can instead be Ω, representing non-existence, to denote any population involving these three people. In particular, let's say $X = (12, 11, 10)$ and $Y = (15, 14, \Omega)$. Now consider the four population prospects over equiprobable states of nature s_1, s_2 and s_3, represented in the table below.

	s_1	s_2	s_3
X	$(12, 11, 10)$	$(12, 11, 10)$	$(12, 11, 10)$
Y	$(12, 11, 10)$	$(11, 10, 12)$	$(10, 12, 11)$
Z	$(15, 14, \Omega)$	$(14, \Omega, 15)$	$(\Omega, 15, 14)$
W	$(15, 14, \Omega)$	$(15, 14, \Omega)$	$(15, 14, \Omega)$

By Anonymity and Moral Statewise Dominance, $X \sim Y$ and $Z \sim W$. By Moral Ex Ante Pareto, Y is better than Z provided Y is better for each person than Z; Risky-Existence Totalism implies this is indeed the case. It follows that Y is better than Z, and from this, that X is better than Y.

To generalise this argument, consider any two arbitrary populations X and Y. We can turn each into a prospect which gives each person who

might exist in *either* population an equal chance of existing at each wellbeing position in the population in question, *including* non-existence (which we can think of as another wellbeing position). Call these $\mathcal{X}$ and $\mathcal{Y}$ respectively. By Anonymity and Moral Statewise Dominance, $\mathcal{X} \sim X$ and $\mathcal{Y} \sim Y$. By Moral Ex Ante Pareto, $\mathcal{X} \succeq \mathcal{Y}$, provided $\mathcal{X} \succeq_i \mathcal{Y}$ for each person i; and vice versa. According to Risky-Existence Totalism, the two prospects are ranked, for each person, according to their total wellbeing divided by the total number of people existing in *either* X or Y. Since this is just a constant, the two prospects are ranked in terms of prudential value by the total wellbeing of the populations they derived from. Therefore, if X has at least as much total wellbeing as Y, then $\mathcal{X} \succeq \mathcal{Y}$, and vice versa. Hence, if X at least as much total wellbeing as Y, then $X \succeq Y$; if Y has at least as much total wellbeing as X, then $Y \succeq X$. Similarly, if one population has more total wellbeing than the other, then the corresponding prospect is better for each person; consequently, Moral Ex Ante Pareto implies that this prospect is better overall, and transitivity yields that the first population is then better than the second.

Notice that Risky-Existence Totalism comes into play at exactly one stage: in determining whether or not $\mathcal{X}$ is better than $\mathcal{Y}$. If we substituted Risky-Existence Totalism with a different prudential axiology for risky existence cases, it would generate a different population axiology. More generally, there is a one-to-one correspondence between prudential axiologies in

risky existence cases and population axiologies: with sufficient background assumptions, each uniquely determines the other. This result is due to McCarthy et al. (2020); what has been said here is just a special case.

7.4.2 The Neutral Addition Argument for Risky-Existence Totalism

I will not repeat my arguments for the other premises: regarding the strengthened versions of Moral Ex Ante Pareto and Moral Statewise Dominance, suffice to say that these seem just as compelling to me as the versions which appear in the original Aggregation Theorem. The important question, as I see it, is whether we should accept Risky-Existence Totalism.

Let me quickly introduce some notation. Say that two populations X and Y give i the same outcome, denoted $X =_i Y$, if either X and Y both give i non-existence, or X and Y both give i existence at the same wellbeing level.

Risky-Existence Totalism follows from the following two principles, which are risky-existence analogues of the premises of the population-axiological Neutral Addition Argument:

> *Same-State Risky-Existence Totalism* Let X and Y be any population prospects over the same set of states of nature. Let i be an individual with positive probability of existence in both X and Y, and who exists in the same states of nature in each case. $X \succeq_i Y$ if and only if,

conditional on existence, the expected wellbeing of i in X is greater than or equal to the expected wellbeing of i in Y.

Principle of Neutral Non-Existence Let X and X' be any population prospects over the set S of states of nature. Let i be an individual with a positive probability of existence in both X and X'. Suppose that, for some $s \in S$, and all $s' \neq s, s' \in S$:

(i) i does not exist in $X(s)$.

(ii) i exists at a neutral level in $X'(s)$.

(iii) $X(s') =_i X'(s')$.

Then $X \sim_i X'$.

The risky-existence Neutral Addition Argument goes as follows. Suppose X and Y are any population prospects giving i a positive probability of existence. The Principle of Neutral Non-Existence implies that X and Y are equally as good for i as X' and Y' respectively, where these prospects are exactly like X and Y, but give i neutral existence instead of non-existence wherever applicable. Hence $X \succeq_i Y$ if and only if $X' \succeq_i Y'$. Same-State Risky Existence Totalism implies that $X' \succeq_i Y'$ if and only if the expected wellbeing of i in X' is greater than or equal to the expected wellbeing of i in Y', which is precisely the criterion given by Risky-Existence Totalism for the claim that $X \succeq_i Y$.

7.4.3 Same-State Risky-Existence Totalism

Same-State Risky-Existence Totalism follows from Expected Utility Theory applied to prospects guaranteeing existence, plus the

> *Sure-Thing Principle (risky-existence version)* Let X, X', Y and Y' be population prospects over the same set S of states of nature. Let $s \in S$, and let i be an individual with positive probability of existence in all four population prospects. Suppose that, for all $s' \neq s, s' \in S$,
>
> (i) $X(s') =_i X'(s')$
>
> (ii) $Y(s') =_i Y'(s')$
>
> (iii) $X(s) =_i Y(s)$
>
> (iv) $X'(s) =_i Y'(s)$
>
> Then $X \sim_i Y$ if and only if $X' \sim_i Y'$.

The statement of this principle looks quite complicated, but the idea is actually rather simple. The idea is that if two prospects give i the same outcome in some state of nature, we can replace that outcome with a different outcome for i in both prospects without affecting how these prospects rank in terms of prudential value for i.[16]

[16]It has been argued by Kowalczyk and Masny (nd) that the Sure-Thing Principle might be inadmissible in these contexts, since it can imply that certain non-existence is equally as good for an individual as certain non-existence, whereas many Existence Non-

Same-State Risky Existence Totalism follows from these two principles because if we are comparing any two prospects X and Y involving non-existence in exactly the same states of nature, the Sure-Thing Principle allows us to replace non-existence in these states of nature with existence at a neutral level of wellbeing, resulting in prospects X' and Y' respectively. Comparing these two prospects, we will then find that $X' \succeq_i Y'$ if and only if, conditional on the existence of i, $X \succeq_i Y$.

7.4.4 The Principle of Neutral Non-Existence

The other distinctive feature of Risky-Existence Totalism is that it implies prudential indifference between non-existence in a state of nature and existence at a neutral level in the same state of nature. This principle might seem uncomfortably close to Existence Comparativism, the view that certain existence and certain non-existence can be compared in terms of prudential value (presumably, with non-existence being equally as good as existence at the neutral level).

We can argue for the Principle of Neutral Non-Existence by appealing to an old friend from Chapter 5, namely the

Comparativists would deny this. My version of the Sure-Thing Principle does not imply this conclusion, because it only applies to prospects which give i a positive probability of existence. This restriction might seem ad hoc at first glance. However, for Existence Non-Comparativists who deny reflexivity in the case of non-existence, there is a natural justification for this restriction: it might be thought that speaking of prudential value for a person who certainly does not exist is nonsensical, while speaking of prudential value for a person who gets a positive probability of existence is not.

Conditional Value Principle For any population prospects X and Y, if X is conditionally good for i and Y is conditionally bad for i, then $X \succ_i Y$.

(Recall that prospect X is conditionally good (or bad or neutral) for i if and only if X gives i positive (or negative or neutral) wellbeing in each state of nature in which i exists; and i does exist in some state of nature.)

We also need the risky-existence version of the Sure-Thing Principle. And we need one more principle. To state it, we need some more notation. Let X_k be an infinite sequence of prospects over the set S of states of nature which give everyone except i the same outcomes in every state of nature. Suppose that i exists in the same states of nature in each member of the sequence X_k, and that the wellbeing levels of i, in each state of nature in which i exists, converge to certain values (which may be different in each state) as the sequence tends to infinity. Write X for the prospect which gives everyone except i whatever they get in the X_k, gives i their convergent wellbeing values in each state of nature for which they exist in the X_k, and gives i non-existence in all other states of nature. We then say that the sequence X_k *converges* to X for i.

We can now state the final principle we need, which is the

Convergence Principle Let X_k be a sequence of population prospects which converges to X for i, and let Y be any population prospect. Suppose that for each X_k, $X_k \succ_i Y$. Then $X \succeq_i Y$. Similarly, if $X_k \prec_i Y$ for each X_k, then $X \preceq_i Y$.

The basic idea behind the Convergence Principle is that arbitrarily small differences in i's wellbeing cannot make the difference between a prospect being better for i and its being worse (or incomparable) for i.

From the Conditional Value Principle and the Convergence Principle, we can derive the

Conditional Neutrality Principle If population prospects X and Y are conditionally neutral for i, then $X \sim_i Y$.

To see this, let X and Y be any population prospects which are conditionally neutral for i. Some quick notation: for any real number x, write $X + x$ to denote the prospect which gives x more units of wellbeing to i whenever i exists, and is otherwise exactly the same. Consider the sequences $X_j = X + 2^{-j}$ and $Y_k = Y - 2^{-k}$, which converge to X and Y respectively. By the Conditional Value Principle, each X_j is better for i than each Y_k. The Convergence Principle thus implies that X is better for i than each Y_k.[17] Applying the Convergence Principle again, we have that $X \succeq_i Y$. Obviously,

[17] X is at least as good as each Y_k, and each Y_k is worse than Y_{k+1}. Hence, X is better than each Y_k.

we can do the same thing the other way round in order to show that $Y \succeq_i X$; we can conclude that $X \sim_i Y$, as required.

Finally, we use the Conditional Neutrality Principle and the risky-existence Sure-Thing Principle to derive the Principle of Neutral Non-Existence. Suppose that the antecedent of this principle is fulfilled: let X and Y be any population prospects over the set S of states of nature, let $s \in S$, and let i be an individual with a positive probability of existence in both X and Y. Suppose that i exists at a neutral level in $Y(s)$ but does not exist at all in $X(s)$. Assume also that i gets the same outcomes in X and Y in all other states of nature. (We know that i exists in at least one of these other states, since otherwise i has zero probability of existence in X.) We want to show that $X \sim_i Y$.

Write X' for the population prospect which is exactly the same as X, except that it gives i a neutral life in all states except s. Similarly, write Y' for the population prospect which is the same as Y, except that it gives i neutral existence in all states except s. (Actually, Y' guarantees a neutral existence for i.) If there are n states of nature in total, then by applying the risky-existence Sure-Thing Principle $n - 1$ times, we have that $X \succeq_i Y$ if and only if $X' \succeq_i Y'$, and vice versa. But X' and Y' are both conditionally neutral for i; hence they are equally good for i by the Conditional Neutrality Principle. Therefore, $X \sim_i Y$.

This completes the argument for Risky-Existence Totalism. This argu-

ment appealed to four premises: Expected Utility Theory for prospects guaranteeing existence, the risky-existence version of the Sure-Thing Principle, the Conditional Value Principle and the Convergence Principle.

7.4.5 Doing Without the Convergence Principle

The Convergence Principle is perhaps the least well-motivated of these four premises. We can do without it, without sacrificing much of importance. In particular, our three other premises imply

> *Risky-Existence Totalism (strict, overlapping states only)* Let X and Y be any population prospects over the same set S of states of nature, and suppose that for some $s \in S$, i exists in $X(s)$ and in $Y(s)$. Then if X gives i greater expected wellbeing than Y, treating non-existence as though it were zero wellbeing, it holds that $X \succ_i Y$.

The general argument will be easier to understand if we see a concrete case first. Consider the population prospects, in which only i exists, illustrated by the following table. We are trying to show that, since X gives i greater expected wellbeing (with neutral non-existence) than Y, X is better for i than Y. As usual, Ω represents the non-existence of i.

	$s_1(p=\frac{1}{2})$	$s_2(p=\frac{1}{4})$	$s_3(p=\frac{1}{4})$
X	24	24	Ω
Y	Ω	32	32
X'	2	68	Ω
Y'	Ω	68	-4
X''	2	0	Ω
Y''	Ω	0	-4
X'''	$\frac{4}{3}$	$\frac{4}{3}$	Ω
Y'''	Ω	-2	-2

Since X has greater total wellbeing than Y, we can find a quantity of wellbeing which, in s_2, exceeds the total expected wellbeing in Y, but is less than the total expected wellbeing in X. We then shift the wellbeing levels of i so that i has this quantity of wellbeing in state s_2, while preserving the total expected wellbeing. In this case, the quantity of wellbeing to be shifted to state s_2 (the average expected total wellbeing, divided by the probability of state s_2) is 68. This moves us from X and Y to X' and Y'; Same-State Risky-Existence Totalism implies that $X \sim_i X'$ and $Y \sim_i Y'$.

Next, we obtain X'' and Y'' by replacing existence at wellbeing level 68 with existence at wellbeing level 0; the Sure-Thing Principle implies that $X' \succeq_i Y'$ (or $Y' \succeq_i X'$) if and only if $X'' \succeq_i Y''$ (or $Y'' \succeq_i X''$).

Finally, we equalise the wellbeing of X'' and Y'' without affecting their total expected wellbeing. We then arrive at X''' and Y''', and we apply Same-State Risky-Existence Totalism again to show that $X'' \sim_i X'''$ and $Y'' \sim_i Y'''$. Since X''' is a conditionally good prospect and Y''' is a conditionally bad

prospect, the Conditional Value Principle implies that $X''' \succ_i Y'''$; it follows that $X \succ_i Y$.

Next, let's see the general case. Suppose that the antecedent of this version of Risky-Existence Totalism is fulfilled. We have that i exists in both $X(s)$ and $Y(s)$. Assume this is also true of some other state s' (if necessary, we can ensure this by splitting s into two equi-probable states of nature). Let p be the probability of state s. For any population prospect Z, write $w_i(Z)$ to denote the expected wellbeing of i in Z, treating non-existence as zero. We know that $w_i(X) \succ w_i(Y)$. Define x to be equal to $\frac{w_i(X)+w_i(Y)}{2p}$.

We can construct population prospects X' and Y', which are just like X and Y respectively, except that they give x units of wellbeing to i in state s, and give i the remaining wellbeing left to make up $w_i(X)$ or $w_i(Y)$ respectively, spread evenly over the other states of nature for which i exists. Since i exists in s', we know that there are other states in which i exists, so it is possible to do this.

The contribution to expected wellbeing for both X' and Y' in state s is $p \cdot x = \frac{w_i(X)+w_i(Y)}{2}$. Since $w_i(X) > w_i(Y)$, $p \cdot x$ is greater than $w_i(Y)$ and less than $w_i(X)$. Hence, the remaining wellbeing is negative in the case of X' and positive in the case of Y'. By Same-State Risky-Existence Totalism, $X \sim_i X'$ and $Y \sim_i Y'$.

Next, consider prospects X'' and Y'', which give i wellbeing 0 rather than x in state s, but which are otherwise exactly like X' and Y'. The Sure-Thing

Principle implies that $X' \succeq_i Y'$ if and only if $X'' \succeq_i Y''$, and $Y' \succeq_i X'$ if and only if $Y'' \succeq_i X''$.

Now let X''' be the i-equalisation of X'': it is the unique population prospect which gives the same outcomes in all states of nature to everyone except i, gives i existence in the same states of nature as X'', gives i the same wellbeing level in all of these states of nature, and preserves i's expected wellbeing: that is, $w_i(X''') = w_i(X'')$. Similarly, let Y''' be the i-equalisation of Y''. Same-State Risky-Existence Totalism implies that $X''' \sim_i X''$ and $Y''' \sim_i Y''$, and the Conditional Value Principle implies that $X''' \succ_i Y'''$. Putting all these claims together, we have that $X \succ_i Y$.

The concludes the argument for the strict, overlapping states version of Risky-Existence Totalism. The overlapping states condition is annoying, but it actually makes no difference for the argument from Risky-Existence Totalism to Totalism. This is because, when we transform two populations into prospects which give everyone the same ex ante position, we can arrange things such that each person exists in both prospects for some state of nature s. Assuming Anonymity, the risky existence version of Moral Ex Ante Pareto and Moral Statewise Dominance, the strict, overlapping states version of Risky-Existence Totalism therefore implies

> *Totalism (strict betterness only)* For any populations X and Y, if X contains more total wellbeing than Y, then $X \succ Y$.

Putting everything together: Anonymity, the risky existence version of Moral Ex Ante Pareto, full-strength Moral Statewise Dominance, Expected Utility Theory for prospects guaranteeing existence, the risky-existence version of the Sure-Thing Principle and the Conditional Value Principle jointly imply the strict-betterness version of Totalism.

The strict-betterness version of Totalism is slightly logically weaker than full-strength Totalism. But the difference between the two is not very large.

Concluding Remarks

Totalism is a simple, elegant and powerful theory. It has many controversial implications. But I believe that we should accept these implications: as we have seen, the price of avoiding them is just too high.

The best positive arguments for Totalism crucially involve appeals to principles of rationality, like the principle of Moral Statewise Dominance, and to principles of benevolence, like the principle of Moral Ex Ante Pareto. Taking rationality for granted, the crucial question becomes whether we should accept Moral Ex Ante Pareto. If we do, this takes us much of the way towards Totalism.

This might come as a surprise to some. Totalism is a Utilitarian axiology. It is often thought that Utilitarianism fails to respect the separateness of persons, that it treats people as mere containers of value, or that it is otherwise objectionably impersonal. Yet avoiding this objectionable impersonalism requires us to accept the sort of tight connection between prudential and moral value imposed by principles like Moral Ex Ante Pareto. If the choice between Totalism and the rest comes down to Moral Ex Ante Pareto, then it is non-Utilitarian population axiologies which are objectionably impersonal, not Totalism.

Bibliography

Adler, M. D. and N. Holtug. 2019. Prioritarianism: A reponse to critics. *Politics, Philosophy & Economics 18*(2): 101–144.

Ahmed, A. 2017. Exploiting cyclic preference. *Mind 126*(504): 975–1022.

Anglin, B. 1977. The repugnant conclusion. *Canadian Journal of Philosophy 7*(4): 745–754.

Arrhenius, G. 2000. An impossibility theorem for welfarist axiologies. *Economics and Philosophy 16*(2): 247–266.

Arrhenius, G. 2003. The very repugnant conclusion. In *Logic, Law, Morality: Thirteen Essays in Practical Philosophy in Honour of Lennart Åqvist*, eds. S. Krister and R. Sliwinski, 167–180. Uppsala: Uppsala University Press.

Arrhenius, G. 2009. One more axiological impossibility theorem. In *Logic, Ethics, and All That Jazz: Essays in Honour of Jordan Howard Sobel*, eds. L.-G. Johansson, J. Österberg, and R. Sliwinski, 23–37. Uppsala: Uppsala University.

Arrhenius, G. 2011. The impossibility of a satisfactory population ethics. In *Descriptive and Normative Approaches to Human Behaviour*, eds. E. N.

Dzhafarov and L. Perry, Chapter 1, 1–26. Singapore: World Scientific Publishing Co.

Arrhenius, G. 2016. Population ethics and different-number based imprecision. *Theoria* 82: 166–181.

Arrhenius, G. n.d. Population ethics: The challenge of future generations. Unpublished manuscript.

Asheim, G. B. and S. Zuber. 2014. Escaping the repugnant conclusion: Rank-discounted utilitarianism with variable population. *Theoretical Economics 9*(3): 629–650.

Bader, R. M. 2022a. The asymmetry. In *Ethics and Existence: The Legacy of Derek Parfit*, eds. J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan, Chapter 1, 15–37. Oxford: Oxford University Press.

Bader, R. M. 2022b. Person-affecting utilitarianism. In *The Oxford Handbook of Population Ethics*, eds. G. Arrhenius, K. Bykvist, T. Campbell, and E. Finneron-Burns, Chapter 11, 251–270. Oxford: Oxford University Press.

Bales, A., D. Cohen, and T. Handfield. 2014. Decision theory for agents with incomplete preferences. *Australasian Journal of Philosophy 92*(3): 453–470.

Barrett, J. 2020a. Efficient inequalities. *The Journal of Political Philosophy 28*(2): 181–198.

Barrett, J. 2020b. Is maximin egalitarian? *Synthese 197*(2): 817–837.

Blackorby, C., W. Bossert, and D. Donaldson. 1995. Intertemporal population ethics: Critical-level utilitarian principles. *Econometrica 63*(6): 1303–1320.

Blackorby, C., W. Bossert, and D. Donaldson. 1996. Quasi-orderings and population ethics. *Social Choice and Welfare 13*(2): 129–150.

Blackorby, C., W. Bossert, and D. Donaldson. 2003. The axiomatic approach to population ethics. *Politics, Philosophy & Economics 2*(3): 342–381.

Blackorby, C., W. Bossert, and D. Donaldson. 2005. *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. New York: Cambridge University Press.

Blackorby, C. and D. Donaldson. 1984. Social criteria for evaluating population change. *Journal of Public Economics 25*(1): 13–33.

Boonin, D. 2014. *The non-identity problem and the ethics of future people*. New York: Oxford University Press.

Boonin-Vail, D. 1996. Don't stop thinking about tomorrow: Two paradoxes

about duties to future generations. *Philosophy & Public Affairs 25*(4): 267–307.

Brooker, C. (writer). 2013. White bear. Black Mirror, series 2, episode 2.

Broome, J. 1991. *Weighing goods: equality, uncertainty and time*. Oxford: Basil Blackwell.

Broome, J. 1999. *Ethics Out Of Economics*. Cambridge: Cambridge University Press.

Broome, J. 2004. *Weighing Lives*. Oxford: Oxford University Press.

Broome, J. 2005. Should we value population? *Journal of Political Philosophy 13*(4): 399–413.

Bykvist, K. 2007. The benefits of coming into existence. *Philosophical Studies 135*(3): 335–362.

Carlson, E. 1998. Mere addition and two trilemmas of population ethics. *Economics and Philosophy 14*(2): 283–306.

Chang, R. 2002. The possibility of parity. *Ethics 112*(4): 659–688.

Chang, R. 2016. Parity, imprecise comparability and the repugnant conclusion. *Theoria* 82: 182–214.

Crisp, R. 2003. Equality, priority, and compassion. *Ethics 113*(4): 745–763.

Cusbert, J. 2017. Acting on essentially comparative goodness. *Thought: A Journal of Philosophy 6*(2): 73–83.

Diamond, P. A. 1967. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility: comment. *Journal of Political Economy* 75: 765–766.

Dow, G. K. 1984. Myopia, amnesia, and consistent intertemporal choice. *Mathematical Social Sciences 8*(2): 95–109.

Fishburn, P. C. 1982. *The Foundations of Expected Utility*. Dordrecht: Reidel.

Fleurbaey, M. 2009. Two variants of harsanyi's aggregation theorem. *Economics Letters 105*(3): 300–302.

Frick, J. 2014. *'Making People Happy, Not Making Happy People': A Defense of the Asymmetry Intuition in Population Ethics*. Ph.D. thesis, Harvard University.

Frick, J. 2015. Contractualism and social risk. *Philosophy & Public Affairs 43*(3): 175–223.

Frick, J. 2017. On the survival of humanity. *Canadian Journal of Philosophy 47*(2-3): 344–367.

Frick, J. 2020. Conditional reasons and the procreation asymmetry. *Philosophical Perspectives: Ethics* 34: 53–87.

Frick, J. 2022. Context-dependent betterness and the mere addition paradox. In *Ethics and Existence: the Legacy of Derek Parfit*, eds. J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan, Chapter 9, 232–263. Oxford: Oxford University Press.

Gert, J. 2003. Requiring and justifying: Two dimensions of normative strength. *Erkenntnis 59*(1): 5–36.

Goodsell, Z. 2021. A St Petersburg paradox for risky welfare aggregation. *Analysis 81*(3): 420–426.

Greaves, H. 2015. Antiprioritarianism. *Utilitas 27*(1): 1–42.

Greaves, H. 2016. Cluelessness. *Proceedings of the Aristotelian Society 116*(3): 311–339.

Greaves, H. and W. MacAskill. 2021. The case for strong longtermism. GPI Working Paper No. 5–2021.

Gustafsson, J. E. 2010. A money-pump for acyclic intransitive preferences. *Dialectica 64*(2): 251–257.

Gustafsson, J. E. 2020. Population axiology and the possibility of a fourth category of absolute value. *Economics and Philosophy 36*(1): 81–110.

Gustafsson, J. E. n.d. Money-pump arguments. Unpublished manuscript.

Gustafsson, J. E. and P. Kosonen. forthcoming. Do lefty and righty matter more than lefty alone? *Erkenntnis*.

Gustafsson, J. E. and W. Rabinowicz. 2020. A simpler, more compelling money pump with foresight. *The Journal of Philosophy 117*(10): 578–589.

Hare, C. 2010. Take the sugar. *Analysis 70*(2): 237–247.

Harsanyi, J. C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy 63*(4): 309–321.

Herstein, I. N. and J. Milnor. 1953. An axiomatic approach to measurable utility. *Econometrica 21*(2): 291–297.

Holtug, N. 2010. *Persons, Interests, and Justice*. Oxford: Oxford University Press.

Huemer, M. 2008. In defence of repugnance. *Mind 117*(468): 899–933.

Huemer, M. 2012. Against equality and priority. *Utilitas 24*(4): 483–501.

Karhu, T. forthcoming. What justifies our bias towards the future? *Australasian Journal of Philosophy*.

Klocksiem, J. 2016. How to accept the transitivity of better than. *Philosophical Studies* 173: 1309–1334.

Kolodny, N. forthcoming. Saving posterity from a worse fate. In *Ethics and Existence: the Legacy of Derek Parfit*, eds. J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan. Oxford: Oxford University Press.

Kowalczyk, K. and M. Masny. n.d. The risky existential question and the repugnant conclusion. Unpublished manuscript.

Lazar, S. 2018. Limited aggregation and risk. *Philosophy & Public Affairs 46*(2): 117–159.

MacKaye, J. 1906. *The Economy of Happiness*. Boston: Little, Brown, and Company.

McCarthy, D. 2006. Utilitarianism and prioritarianism I. *Economics and Philosophy 22*(3): 335–363.

McCarthy, D., K. Mikkola, and T. Thomas. 2020. Utilitarianism with and without expected utility. *Journal of Mathematical Economics* 87: 77–113.

McClennen, E. F. 1985. Prisoner's dilemma and resolute choice. In *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, eds. R. Campbell and L. Sowden, Chapter 5, 94–104. Vancouver: University of British Columbia Press.

McClennen, E. F. 2000. The rationality of rules. In *Rationality, Rules, and Structure*, eds. J. Nida Rümelin and W. Spohn, Chapter 2, 17–33. Berlin: Springer.

McMahan, J. 1981. Problems of population theory. *Ethics 92*(1): 96–127.

McMahan, J. 2009. Asymmetries in the morality of causing people to exist. In *Harming Future Persons: Ethics, Genetics and the Nonidentity Problem*, eds. M. A. Roberts and D. T. Wasserman, Chapter 3, 49–68. Dordrecht: Springer.

McMahan, J. 2013. Causing people to exist and saving people's lives. *The Journal of Ethics 17*(1): 5–35.

Morgenstern, O. and J. Von Neumann. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Muñoz, D. 2021. Three paradoxes of supererogation. *Noûs 55*(3): 699–716.

Narveson, J. 1967. Utilitarianism and new generations. *Mind 76*(301): 62–72.

Narveson, J. 1973. Moral problems of population. *The Monist 57*(1): 62–86.

Nebel, J. M. 2018. The good, the bad, and the transitivity of better than. *Noûs 52*(4): 874–899.

Nebel, J. M. 2019. An intrapersonal addition paradox. *Ethics 129*(2): 309–343.

Nebel, J. M. 2022. Totalism without repugnance. In *Ethics and Existence: The Legacy of Derek Parfit*, eds. J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan, Chapter 8, 200–231. Oxford: Oxford University Press.

Ng, Y.-k. 1989. What should we do about future generations? impossibility of parfit's theory x. *Economics and Philosophy 5*(2): 235–253.

Ord, T. 2015. A new counterexample to prioritarianism. *Utilitas 27*(3): 298–302.

Otsuka, M. 2018. How it makes a moral difference that one is worse off than one could have been. *Politics, Philosophy & Economics 17*(2): 192–215.

Otsuka, M. 2022. Prioritarianism, population ethics, and competing claims. In *Ethics and Existence: the Legacy of Derek Parfit*, eds. J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan, Chapter 19, 527–551. Oxford: Oxford University Press.

Parfit, D. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Parfit, D. 1997. Equality and priority. *Ratio 10*(3): 202–221.

Parfit, D. 2004. Overpopulation and the quality of life. In *The Repugnant Conclusion: Essays on Population Ethics*, eds. J. Ryberg and T. Tännsjö, 145–164. London: Kluwer Academic.

Parfit, D. 2011. *On What Matters*. Oxford: Oxford University Press.

Parfit, D. 2012. Another defence of the priority view. *Utilitas 24*(3): 399–440.

Parfit, D. 2016. Can we avoid the repugnant conclusion? *Theoria 82*(2): 110–127.

Parfit, D. 2017. Future people, the non-identity problem, and person-affecting principles. *Philosophy & Public Affairs 45*(2): 118–157.

Persson, I. 2011. Prioritarianism, levelling down and welfare diffusion. *Ethical Theory and Moral Practice 14*(3): 307–311.

Persson, I. 2012. Prioritarianism and welfare reductions. *Journal of Applied Philosophy 29*(4): 289–301.

Pollak, R. 1968. Consistent planning. *The Review of Economic Studies 35*(2): 201–208.

Portmore, D. W. 2021. *Morality and Practical Reasons*. Elements in Ethics. Cambridge: Cambridge University Press.

Qizilbash, M. 2007a. The mere addition paradox, parity and vagueness. *Philosophy and Phenomenological Research 75*(1): 129–151.

Qizilbash, M. 2007b. The parity view and intuitions of neutrality. *Economics and Philosophy 23*(1): 107–114.

Qizilbash, M. 2018. On parity and the intuition of neutrality. *Economics and Philosophy 34*(1): 87–108.

Rabinowicz, W. 2002. Prioritarianism for prospects. *Philosophical Issues 14*(1): 2–21.

Rabinowicz, W. 2009. Broome and the intuition of neutrality. *Philosophical Issues* 19: 389–411.

Rachels, S. 1998. Counterexamples to the transitivity of *Better Than*. *Australasian Journal of Philosophy 76*(1): 71–83.

Rachels, S. 2001. A set of solutions to parfit's problems. *Noûs 35*(2): 214–238.

Rachels, S. 2004. Repugnance or intransitivity: A repugnant but forced choice. In *The Repugnant Conclusion: Essays on Population Ethics*, eds. J. Ryberg and T. Tännsjö, 163–186. London: Kluwer Academic.

Rawls, J. 1999. *A Theory of Justice* (Revised edition ed.). Cambridge, Massachusetts: Harvard University Press.

Raz, J. 1986. *The Morality of Freedom*. Oxford: Clarendon Press.

Redacted n.d. Redacted. Unpublished manuscript.

Roberts, M. A. 2003. Is the person-affecting intuition paradoxical? *Theory and Decision 55*(1): 1–44.

Roberts, M. A. 2011. The Asymmetry: A Solution. *Theoria 77*(4): 333–367.

Ross, J. 2014. Divided we fall: Fission and the failure of self-interest. *Philosophical Perspectives: Ethics 28*(1): 222–262.

Ross, J. 2015. Rethinking the person-affecting principle. *Journal of Moral Philosophy 12*(4): 428–461.

Rüger, K. 2020. Aggregation with constraints. *Utilitas* 32: 454–471.

Savage, L. J. 1954. *The Foundations of Statistics*. New York: John Wiley & Sons.

Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge, Massachusetts: Harvard University Press.

Scheffler, S. 1994. *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions* (2nd ed.). Oxford: Clarendon Press.

Schoenfield, M. 2014. Decision making in the face of parity. *Philosophical Perspectives* 28: 263–277.

Spears, D. and M. Budolfson. 2021. Repugnant conclusions. *Social Choice and Welfare 57*(3): 567–588.

Temkin, L. S. 1987. Intransitivity and the mere addition paradox. *Philosophy & Public Affairs 16*(2): 138–187.

Temkin, L. S. 1996. A continuum argument for intransitivity. *Philosophy & Public Affairs 25*(3): 175–210.

Temkin, L. S. 2012. *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. Oxford: Oxford University Press.

Thomas, T. 2016. *Topics in Population Ethics*. DPhil thesis, University of Oxford.

Thomas, T. 2018. Some possibilities in population axiology. *Mind 127*(507): 807–832.

Thomas, T. 2019. The asymmetry, uncertainty, and the long term. GPI Working Paper No. 11–2019.

Thomas, T. 2022. Separability and population ethics. In *The Oxford Handbook of Population Ethics*, eds. G. Arrhenius, K. Bykvist, T. Campbell, and E. Finneron-Burns, Chapter 12, 271–295. Oxford: Oxford University Press.

Thomas, T. forthcoming. On evaluative imprecision. In *Ethics and Existence: The Legacy of Derek Parfit*, eds. J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan. Oxford: Oxford University Press.

Thomas, T. n.d. Are spectrum arguments defused by vagueness? Unpublished manuscript.

Thomson, J. J. 1976. Killing, letting die, and the trolley problem. *The Monist 59*(2): 204–217.

Thornley, E. 2021. The impossibility of a satisfactory population prospect axiology (independently of finite fine-grainedness). *Philosophical Studies 178*(11): 3671–3695.

Thornley, E. forthcoming. Critical levels, critical ranges, and imprecise exchange rates in population axiology. *Journal of Ethics and Social Philosophy*.

Velleman, J. D. 1991. Well-being and time. *Pacific Philosophical Quarterly 72*(1): 48–77.

Voorhoeve, A. 2013. Vaulting intuition: Temkin's critique of transitivity. *Economics and Philosophy* 29: 409–423.

Voorhoeve, A. 2014. How should we aggregate competing claims? *Ethics 125*(1): 64–87.

Voorhoeve, A. 2017. Why one should count only claims with which one can sympathise. *Public Health Ethics 10*(2): 148–156.

Williams, B. 1970. The self and the future. *The Philosophical Review 79*(2): 161–180.

Zuber, S. et al. 2021. What should we agree on about the repugnant conclusion? *Utilitas 33*(4): 379–383.