TOPICS IN POPULATION ETHICS

JOAQUIN TERUJI THOMAS

St Cross College, University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term 2016

Abstract

Topics in Population Ethics

Joaquin Teruji Thomas
St Cross College, University of Oxford

Submitted for the Degree of Doctor of Philosophy, Trinity 2016
About 71,600 words

This thesis consists of several independent papers in population ethics. I begin in Chapter I by critiquing some well-known ‘impossibility theorems’, which purport to show there can be no intuitively satisfactory population axiology. I identify axiological vagueness as a promising way to escape or at least mitigate the effects of these theorems. In particular, in Chapter II, I argue that certain of the impossibility theorems have little more dialectical force than sorites arguments do. From these negative arguments I move to positive ones. In Chapter III, I justify the use of a ‘veil of ignorance’, starting from three more basic normative principles. This leads to positive arguments for various kinds of utilitarianism – the best such arguments I know. But in general the implications of the veil depend on how one answers what I call ‘the risky existential question’: what is the value to an individual of a chance of non-existence? I chart out the main options, and raise some puzzles for non-comparativism, the view that life is incomparable to non-existence. Finally, in Chapter IV, I consider the consequences for population ethics of the idea that what is normatively relevant is not personal identity, but a degreed relation of psychological connectedness. In particular, I pursue a strategy based in population ethics for understanding the controversial ‘time-relative interests’ account of the badness of death.

Acknowledgements

I thank my supervisors, Hilary Greaves and Ralf Bader, for their thoughtful engagement with my work, and support in countless other ways. Two other people have played a similar role: John Broome, who supervised and encouraged my work in moral philosophy during the BPhil, and David McCarthy, who has been enormously generous with his attention and advice.

Many other people have given me valuable feedback, including Gustaf Arrhenius, Tim Campbell, Robyn Kath, Per Algander, Stefan Riedener, John Halstead, and especially the other members of the Population Ethics project at the Future of Humanity Institute: Toby Ord, Theron Pummer, John Cusbert, Simon Beard, and Daniel Cohen.

That project, funded by the Leverhulme Trust, has been my intellectual home while writing this thesis. I was more materially supported by a graduate studentship from the Arts and Humanities Research Council. And I am grateful to the administrative staff at the philosophy faculty and FHI, especially Iris Geens, Sally Baume, and Kathryn Mecrow, for all their work behind the scenes.

This is also a fitting place to thank some of the philosophers who inspired and (perhaps unwittingly) abetted me in my turn towards philosophy. Besides Hilary Greaves and David McCarthy, I may mention David Wallace, Harvey Brown, Andrew Bacon, Ralph Wedgwood, and Maria Lasonen-Aarnio.

Finally, my whole family has offered unfailing support. Most of all I am grateful to Samantha Johnson and Ryukazu Thomas for keeping me company on this funny old adventure. I hope I am good company on theirs.

Contents

Introduction

This thesis consists of several independent papers on population ethics. This integrative chapter provides a thematic overview and points out a few of the many loose ends which I may address in future work. Since each later chapter is self-contained, I will keep these comments brief and work in broad strokes. (The one substantial argument is my criticism of Parfit's Imprecise Lexical View, starting on page 7.)

What is Population Ethics?

The fundamental question of population ethics is how we should take into account people whose very existence depends on our choices. This is important because practically everything we do has an effect, or a chance of an effect, on both how many people and *which* people eventually exist.

One of the main choice-points can be made vivid by considering the possibility of human extinction. Suppose that, if we did nothing, the Earth would be vaporised next year, killing everyone painlessly. Suppose we have a way of completely avoiding this catastrophe, ensuring that humanity will continue on as expected for a thousand generations. (Just for simplicity, I will ignore animal life, and assume that future people by and large will have good lives.) But avoidance has a cost: it will impose on each person currently alive a period of suffering slightly worse for each one than the prospect of painless death. Does the current generation have a moral obligation, in such a case, to endure great suffering and thus ensure the existence of future generations? If so, what is the nature and the strength of this obligation?

In thinking about this case there are two competing principles. One, which I call the *principle of neutrality*, is aptly summarised by the thought that 'We are in favour of making people happy, but neutral about making happy people' (Narveson, 1973, p. 80). It is far from clear how best to make sense of this initially attractive thought. But one obvious reading is that there are no moral reasons to ensure the existence of future generations, and very strong moral reasons to avoid present suffering worse than death. On balance we should accept annihilation.

The competing principle is that *more is better*: more good lives means, in particular, *more good.* (See, for example, Huemer (2008, §6).) We must ensure survival when the very existence of so many good lives, and so much good, is at stake, at almost any cost to the few billion people presently alive.

Impossibility Theorems

Both of these principles face intuitive and theoretical difficulties. For one thing, different people seem to have strongly opposed intuitions about the importance of avoiding extinction. More generally, though, a number of 'impossibility theorems' have appeared that show that it is difficult to reconcile various intuitive criteria. These theorems are the subject of Chapters I and

II, although they also have some bearing on Chapters III and IV.

Broadly speaking, the problem with the principle of neutrality is that it is hard not to make it too strong. We end up being neutral about too many things. For example, we might end up being neutral about 'making happy people' even when it benefits already existing people. (This violates a criterion called 'Dominance Addition', in the terminology of Arrhenius (2013).) Or we might end up neutral between creating good lives and creating bad ones. (This violates a version of 'Non-Sadism'.) Chapters III and IV of this thesis are oriented towards problems of this sort.

In contrast, Chapters I and II are largely concerned with difficulties faced by the principle that more is better. (To put it another way, in my discussion of the impossibility theorems, I largely take it for granted that a theory of population ethics should satisfy such conditions as Dominance Addition and Non-Sadism mentioned above.) The most famous difficulty of that sort is expressed by the so-called 'Repugnant Conclusion' (Parfit, 1986): for any world full of wonderful lives, it would be better, all else equal, to have a world of sufficiently many lives that are barely good at all. This conclusion, considered unacceptable by many, is certainly *suggested* by the principle that more is better – and indeed it turns out to be difficult to avoid.

Vagueness and Lexicality

Following a general evaluation of the impossibility theorems in Chapter I, I identify *axiological vagueness* as a promising way to mitigate their impact. The reason is that many of the impossibility theorems are formally akin to sorites arguments. This point has been noticed before, but not properly spelled out. I claim that once it has been spelled out, it is difficult to dismiss. That is the argument of Chapter II.

One project which I have had to leave for another occasion is to map out more completely the space of theories that are left over, given this response to the impossibility theorems. As it is, I have relied on some toy models to illustrate the mechanics and the internal consistency of the view. Unfortunately, these models do not seem very promising in themselves, so some pessimism is still warranted. Let me explain why.

Derek Parfit has recently advocated a view of the same kind, and since I did not have an opportunity to discuss his view directly in the main text, I may as well use it as the example here. His 'Imprecise Lexical View' is characterised by two statements.

> [1] Anyone's existence is in itself good if this person's life is worth living. Such goodness has non-diminishing value...[[1]]
>
> [2] If many people exist who would all have some high quality of life, that would be better than the non-existence of any number of people whose lives, though worth living, would be, in certain ways, much less good. (Parfit, 2016, p. 112)

The first statement is a natural strengthening of the thought that more is better. The second statement is a denial of the Repugnant Conclusion. So

[1] The sentence continues: 'so if there were more such people, the combined goodness of their existence would have no upper limit'. But I am not sure what he means by this, given the part of the view quoted next.

there are not many overt commitments here. Nonetheless, the view so expressed faces serious problems.

Let A be the world in which many people exist with a high quality of life, and Z a world with even more people but whose lives are barely worth living. Let Z+ be like Z, but with ten times as many people. Imagine a scenario in which either Z or A might arise by chance (call this 'the Risky Scenario'), and compare it to a scenario in which Z+ will arise for sure (call this 'the Safe Scenario'). Which of these scenarios is better?

Z+ is better than Z, by condition [1], and this counts in favour of the Safe Scenario. On the other hand, Z+ is worse than A, by condition [2], and this counts against the Safe Scenario. How do these things weigh up? In general it may not be clear, but suppose that the chance that A will arise in the Risky Scenario is only one chance in a billion, or one chance in a trillion, or one chance in a trillion trillions. Thus it is all but certain that the Risky Scenario will result in an outcome significantly worse than the outcome of the Safe Scenario. It is hard for me, and for many others, to believe that the Risky Scenario is not worse than the Safe Scenario, if the chance of A is small enough.

Suppose, for concreteness, that the Risky Scenario is worse than the Safe Scenario when the chance of A is one in one hundred. Then it is safe to say that the value difference between A and Z is less than a million times greater than the value difference between Z+ and Z. Otherwise, the small chance of A would still easily make the Risky Scenario better than the Safe Scenario.

On the other hand, according to Parfit's formulation of the lexical view, the value of the lives in Z is non-diminishing. Let Z++ be a world like Z+, but with a trillion times more lives. I do not know what 'non-diminishing' means if it does not imply that the value difference between Z++ and Z is *at least* a million times greater than the value difference between Z+ and Z. Indeed, naively, the first value difference must be about a trillion times greater than the second.

Therefore the value difference between Z++ and Z is unambiguously greater than the value difference between A and Z. So Z++ is better than A, and the imprecise lexical view is false.

Perhaps there are some ways to block this argument. But the fact that value is vague or (in Parfit's terminology) imprecise is no objection, as far as I can tell. I did not say anything that presupposes precision. For example, the claim that one value difference is more than a million times greater than another does not presuppose that the first value difference is some precise number of times greater than the second.

Uncertainty, Utilitarianism, and Non-Existence

The argument I just gave illustrates how powerful considerations of uncertainty can be in evaluating normative theories. The argument thus goes well beyond the austere formalism used in the standard impossibility theorems. Indeed, that is the main moral of Chapter I: the indispensability of a wider circle of considerations, both formal and substantive.

Uncertainty is also a crucial ingredient in some of the best *positive* arguments in population ethics. In Chapter III, I give an argument of this kind, drawing on joint work with David McCarthy and Kalle Mikkola. In the first instance, the argument is for a 'veil of ignorance principle'. A bit roughly, this associates the moral point of view with the point of view of rational self-interest in the face of self-locating ignorance. As I explain, this idea has a long history, but it has lacked both a clean formulation and a clear justification. The implications of the principle depend on how one answers what I call 'the risky existential question': what is the value to an individual of a chance of non-existence? Some ways of answering this question lead to efficient arguments for generalisations of total utilitarianism. My critique of Parfit above meshes well with this point of view: lexical versions of utilitarianism correspond behind the veil to the denial of one of standard axioms of decision theory.

This chapter also raises a puzzle. How should one answer the risky existential question if one accepts the principle of neutrality along with the cognate view that life cannot be better or worse for a person than non-existence? I explore several options here, all problematic. It seems quite plausible that the best way to understand the principle of neutrality requires a more significant departure from the kind of axiological principles that we use to derive the veil of ignorance.

Neutrality and Identity

The final chapter also centers around the principle of neutrality. I in fact sketch two theories ('Complex Necessitarianism' and 'Regret Minimization') that validate the principle in relatively plausible ways. But getting the optimal version of the principle of neutrality is not my main concern in this chapter.

Rather, my treatment of it is motivated by two other problems.

It has become popular to think that the normative role traditionally played by personal identity is instead played by some relation of psychological connectedness. Moreover, psychological connectedness, unlike personal identity, comes in degrees. There is a well-known account of prudential rationality that incorporates this idea. This is Jeff McMahan's 'time-relative interests' account of prudential value. (It is most often discussed as an account of the badness of death.) Even though it is initially an account of prudential value, it is also supposed to have certain ethical consequences. But serious questions have been raised about the coherence and plausibility of the resulting ethical view. Really, it is unclear what the view amounts to. The first motivating problem in Chapter IV is how best to understand the theory of time-relative interests.

This is (so I claim) intertwined with a more general problem. Formal discussions of population ethics most often deal in *lifetime welfare* – how good a whole life is for the person who lives it. But this way of doing population ethics conflicts with the idea that it is not personal identity but degrees of psychological connectedness that matter. The second problem is how to reimagine population ethics in light of this idea.

This is where the principle of neutrality comes back in. My suggestion is that the problems faced by the time-relative interests account coincide with the problems faced by the principle of neutrality, when the latter is reinterpreted in terms of psychological connectedness. Making good sense of the principle of neutrality is the key step to systematizing (and ultimately evaluating) McMahan's ethical view.

Some Loose Ends

Let me conclude this overview by mentioning a few of the themes which I wish I could have developed in detail, but could only touch on. They provide some directions for future work.

First, throughout this thesis I assume that there are only finitely many people who are relevant in any given choice. Infinite populations raise additional problems. In a sense, these problems are sui generis, and easy to set aside, as I do here. But they are also potentially devastating, and deserve proper consideration.[2]

Second, given my preoccupation with vagueness in Chapter II, as well as some discussion of indeterminate personal identity in Chapter IV, the question arises what one ought to *do* when faced with axiological indeterminacy. If it is indeterminate whether X is better than Y, is it indeterminate which of them one ought to choose? Or are both of them in some sense permissible? Several different approaches to this kind of question have recently appeared in the literature, but very little consensus has been reached.[3]

Finally, the principle of neutrality is especially difficult to understand in situations of uncertainty. I touch on these difficulties in Chapters III and IV, but there is much more to say. The basic issue can again be made vivid by the prospect of extinction. It is only in rather extreme cases that a course of action will lead to extinction with anything like certainty. But many things we do can affect the *risk* of extinction in the medium term. How should we apply the principle of neutrality in the face of existential risk?

[2] See Bostrom (2011) for an influential survey.

[3] See Williams (2014a); Rinard (2015); Dunaway (2016) for three very different recent approaches.

I | Some Possibilities in Population Axiology

ABSTRACT. It is notoriously difficult to find an intuitively satisfactory rule for aggregating welfare. Standard examples, like total utilitarianism, either entail the Repugnant Conclusion or run afoul of some other intuition of similar strength. Several philosophers have presented formal arguments that seem to show that this happens of necessity: our core intuitions stand in contradiction. This paper assesses the state of play, focusing on the most powerful of these 'impossibility theorems', as developed by Gustaf Arrhenius.

Narrowly construed, the goal of these theorems is to establish a conflict between intuitions that are so strong and widespread it would seem repugnant to set them aside. I argue that, even accepting the force of these intuitions, the theorems fall short of their goal. Some of them appeal to a supposedly egalitarian condition which, however, does not properly reflect egalitarian intuitions; others rely on a background assumption about the structure of welfare which cannot be taken for granted.

More broadly construed, the theorems remain important: they give insight into the difficulty, if not the impossibility, of constructing a satisfactory population axiology. We should aim for reflective equilibrium between intuitions and more theoretical considerations. I conclude by highlighting one possible ingredient in this equilibrium, which, I argue, leaves open a still wider range of acceptable theories: the possibility of vague or otherwise indeterminate value relations.

1 Introduction

It sometimes happens that one possible population is better than another with respect to the distribution of welfare. For example, suppose that the two populations have the same size, and that in the first population everyone has a happy and fulfilling life, while in the second population every life is unhappy and devoid of meaning. Then – always as far as welfare goes – the first population is better than the second. A population axiology, in a sense I will later make precise, is a theory of such comparisons.[1]

It turns out to be hard to find such a theory that accords with certain strong and widely held intuitions. For example, consider a large population of happy and fulfilling lives, and a second, perhaps larger one, in which life is barely worth living. Many people strongly intuit that the first population must be better than the second. On the other hand, it appears that, if the second population is sufficiently large, it will inevitably have more *total* welfare than the first. And so one obvious criterion for betterness seems to entail what Parfit (1986) calls *the Repugnant Conclusion*: the second population may be better than the first.[2]

[1] A few clarificatory points. First, I will speak of 'happy' lives, and so on, just to mean those lives with a high level of welfare, without commitment to any particular theory of wellbeing. The core intuitions are meant to hold for a wide range of views about what constitutes welfare. Second, some people might prefer to frame things in terms of what we ought to do, or in terms of reasons. They might say, for example, that as far as welfare goes, we have more reason to bring about the first population than the second. Such a re-framing would not change the specific arguments of this paper. Finally, where I say that one population is better than another with respect to the distribution of welfare, some prefer to say that a population with one distribution of welfare would be better than a population with another distribution, all else equal. I prefer my formulation, but nothing is supposed to hang on it here.

[2] I say only 'seems to entail' because I will later consider a theory of total welfare that does not entail the Repugnant Conclusion. The well-known 'critical-level' theories (for

Can we avoid the Repugnant Conclusion? By itself, of course. But concrete attempts to do so have turned out to violate other intuitions of comparable strength. This has led several authors to produce formal arguments that seem to rule out any completely satisfactory population axiology.[3] These arguments reach their culmination in Gustaf Arrhenius's *Population Ethics*, intended to be the major survey of the current state of the field.[4] His six increasingly subtle 'impossibility theorems' claim to show that our core intuitions stand in contradiction. The implications of this claim are potentially profound. At a basic level, we are simply learning what kinds of bullets we must bite. But if we cannot adjudicate between the core intuitions, we may be pushed into a wider methodological and meta-ethical inquisition.

In this paper I will review some of the basic ideas used in these impossibility theorems, and identify some problems with them. Because Arrhenius has approached the subject systematically and aimed for the best possible results, it is convenient to focus on his work. However, as I will make clear along the way, these problems undermine all arguments I have seen of a similar kind.

Those arguments deploy two basic strategies, and I will divide up my analysis accordingly. The first basic strategy relies on the background assump-

which see Blackorby et al. (1995)) are also arguably of this kind. Of course, one may wonder what it means to 'total' welfare at all; I will raise a related issue in section 4.

[3] Examples include Ng (1989); Carlson (1998); Kitcher (2000); Tännsjö (2002). Most of the ideas are ultimately derived from Parfit (1986).

[4] Arrhenius's book, forthcoming from Oxford University Press, is well known in draft form. In fact, all of the relevant parts have been published in previous work (subject to some irrelevant revisions). I will refer to the theorems as they are enumerated in the manuscript, but cite the published discussions. The first four theorems are essentially those developed in (Arrhenius, 2000a); the fifth is from (Arrhenius, 2003); the sixth is from (Arrhenius, 2009, 2011). I note that the proof of his favoured sixth theorem contains an error in the derivation of the 'Restricted Quality Addition Condition' (Arrhenius, 2011, Lemma 1.3). As far as I know, one has to slightly strengthen the premisses of the theorem, in a way unlikely to cause further controversy. This will not affect my discussion.

tion – typically suppressed – that one can get from a low welfare level to any higher welfare level by a finite number of appropriately 'small' increments. I will call this premiss 'Small Steps'. In section 3, I argue that Small Steps could be denied, and show that, without it, there are counterexamples to four of Arrhenius's six theorems. One is a lexic version of total utilitarianism.

The second basic strategy is used in the remaining two theorems, and is very popular in the wider literature. It appeals to egalitarian intuitions. But, as I argue in section 4, the key 'egalitarian' condition is poorly motivated, and does not capture any useful notion of egalitarianism. Indeed, my example of total lexic utilitarianism satisfies two other egalitarian conditions that are often supposed to be logically stronger. In particular, none of the six theorems tells strongly against total lexic utilitarianism.

Where does this leave us? The impossibility theorems I will discuss embody a particular top-down methodology. Instead of building on foundations, the focus is on identifying high-level and very abstract intuitions as hard constraints. These intuitions should be so strong and widespread it would seem repugnant to set them aside. For the purposes of this paper, I will take for granted that Arrhenius's adequacy conditions meet this test, with the exception I discuss at length in section 4.[5] But the key background assumption, Small Steps, does not fit well with this methodology. Its denial,

[5] There is lively disagreement about whether the Repugnant Conclusion itself is truly repugnant; see especially Huemer (2008), who invokes (for one thing) the kind of argument for RC that I discuss. It is worth noting that the most nuanced fifth and sixth theorems involve intuitive strengthenings of the Repugnant Conclusion, which strike many, including me, as more problematic. For example, the fifth invokes the 'Very Repugnant Conclusion': any large population of blissful lives is worse than one consisting of (say) ten times as many people in terrible agony as well as a lot of others whose lives are barely worth living. But such strengthenings would not change the main points I will make, and in general I will avoid irrelevant complications.

while perhaps surprising, would hardly be repugnant. Nor does it have the quasi-logical character of some other background assumptions, like the one that 'better than' is transitive. To evaluate Small Steps, we have to look beyond the intuition-matching rubric to more deeply theoretical considerations about the nature and structure of wellbeing.

I will sketch some of these considerations in section 5, and then explain a final way to mitigate the impact of the theorems. I suggest that a careful treatment of axiological vagueness could leave open a still wider range of acceptable theories. I show that, even assuming Small Steps, there need not be *determinate* counterexamples to the most important adequacy conditions. At a minimum, vagueness provides a conservative way to weaken those conditions – conservative because our intuitions may not easily distinguish between certain cases of determinate and indeterminate truth.

2 The Framework and Key Examples

To set the stage, I will first describe the framework in which all of the impossibility theorems take place. The idea is to assume as little as possible beyond what is needed to state the main adequacy conditions. Although I have omitted some possible nuances, this framework is meant to be common ground. I will start a bit informally, and then say officially what data constitute a population axiology.

The first idea is that one life may be better for the person living it than another – it may have higher *welfare*. The relation 'at least as good as' is assumed to be a preorder, that is, reflexive and transitive. (It is not required to be complete, although it will be in my examples.) If two lives are equally good, I say they have the same *welfare level*. So a welfare level can be understood as an equivalence class of lives, all equally good.

A population is a collection of lives, and a population axiology will give a preorder on populations – a specification of when one possible population is at least as good as another. However, when I compare populations, I am – by stipulation – only interested in the welfare levels of the lives that they contain. Let me then define a *distribution* to be a finite, unordered list of welfare levels (perhaps containing repetitions). Each finite population determines a distribution, and the population axiology amounts to a preorder on these distributions. (Of course, one might also be interested in infinite populations, but these raise *sui generis* problems – see, for example, Bostrom (2011).) Although nothing ultimately turns on it, it is convenient to assume that every logically possible distribution – every finite unordered list of welfare levels – lies in the domain of the preorder. Here is some useful terminology. If a is a welfare level, then a population or distribution 'at level a' is one in which only level a occurs (perhaps many times). And if A and B are distributions, then $A \cup B$ is the distribution obtained by concatenating the lists A and B. Finally, the *size* of a distribution is the number of people involved, i.e. the length of the list.

The impossibility theorems articulate various adequacy conditions for a population axiology, and state that these adequacy conditions are mutually incompatible. It is important to realise that these adequacy conditions are officially about the *form* of the population axiology; they don't explicitly concern themselves with the *interpretive* question of which welfare levels correspond to which lives. However, to formulate the adequacy conditions in an understandable way, it helps if we can refer to a few broad features of this correspondence. For example, one of the main adequacy conditions is that the population axiology must avoid the Repugnant Conclusion. The Repugnant Conclusion in turn refers to a class of happy, fulfilling (henceforth 'blissful') lives and a class of barely worth living (henceforth 'drab') lives. Officially:

> **The Repugnant Conclusion.** For any distribution at a blissful welfare level, there is a better one at a drab level.

It is possible to eliminate this classification by quantifying over sufficiently high and sufficiently low welfare levels, as Arrhenius effectively does. But I will pragmatically take the classification as a part of the axiology.

Officially, then, a population axiology consists of the following data. First, a set of welfare levels; second, a preorder on that set. Third, a preorder on the corresponding set of distributions. Fourth, a particular welfare level, singled out as 'neutral'. We can then say that a welfare level is 'positive' or 'worth living' if it is higher than the neutral one.[6] Fifth, among the positive welfare levels there is a class of 'blissful' ones, and a disjoint class of 'drab' ones. For my purposes, the only further assumptions are that there exists a blissful level, and that for any blissful level there is a lower drab level and another even lower drab level. Some of Arrhenius's more complicated arguments require three blissful and three drab levels, but those complications are irrelevant to what I shall say.

[6] How exactly to understand the neutral level is one of the key substantive questions; I will say nothing about it here.

To illustrate this framework, let me lay out a few examples. The first is *total utilitarianism*. According to this axiology, the welfare levels are indexed by integers, positive or negative, and ordered in the obvious way.[7] The neutral welfare level is indexed by 0; at least the levels 1 and 2 are drab and 100 is blissful. One distribution (or population) is at least as good as another just in case its total welfare – the sum of integers – is at least as high. A variation on this axiology is *average utilitarianism*. It rules that one distribution is at least as good as another just in case its *average* welfare is at least as high.

Here is a second example, which I call *total lexic utilitarianism* (TLU). Let me begin with an informal picture. (As I emphasise below, this picture is not an official part of the axiology.) There are two things that make life good – call them 'love' and 'money'. The neutral level of welfare corresponds to a life with no love and no money; a blissful life has at least a little love, while a drab life has none. Moreover, a little love is worth any amount of money. So one population, or one life, is at least as good as another if it contains either more love in total, or the same amount of love and at least as much money.

Formally now, TLU claims that welfare levels can be represented by pairs of integers (corresponding, in the picture above, to quantities of love and money respectively). These pairs are ordered lexicographically: (a_1, a_2) is at least as good as (b_1, b_2) just in case either $a_1 > b_1$ or else $a_1 = b_1$ and $a_2 \geq b_2$. The neutral welfare level is $(0, 0)$. I will stipulate that $(1, 0)$ is a blissful level, and that the drab welfare levels are those of the form $(0, m)$, with $m > 0$. Two welfare levels can be added together: $(a_1, a_2) + (b_1, b_2) = (a_1 + b_1, a_2 + b_2)$. It

[7] It is quite common to think of welfare levels as indexed by arbitrary real numbers rather than integers. It makes very little difference to this example, but sticking to integers here and below will be useful in the discussion of Small Steps.

thus makes sense to speak of the total welfare of a distribution. As in total utilitarianism, one distribution is at least as good as another just in case its total welfare is at least as high.

To be clear, this paper is not a defence of total lexic utilitarianism. Rather, I introduce it because it illustrates a wide variety of important ideas. Before getting into more specific issues in the next section, let me explain in general terms why the impossibility theorems get so little traction on this theory.

The first point is that TLU does not entail the Repugnant Conclusion. According to my stipulations above, a drab life contains no love. Therefore a *population* of drab lives contains no love, and must be worse than any population of lives at the blissful level (1,0). At this point one might object that $(1,0)$ could not reasonably correspond to a blissful life. After all, a life with a minimal amount of love is not much better than a life with none at all. We must recognise that a life at level $(1,0)$ is barely worth living, and then we will obtain a version of the Repugnant Conclusion.

But this objection is based on a misunderstanding. My interpretation in terms of 'love' and 'money' was picturesque and convenient. I will continue to use these terms in informal discussion. But this interpretation is not part of the axiology. All the axiology says is that welfare levels can be *represented* by ordered pairs. The welfare levels can be represented in other ways, too; the view is not commited to a two-component analysis of wellbeing. For example, we could represent the welfare levels by real numbers. Concretely, instead of using a pair (a_1, a_2) of integers, we could use the real number $2a_1 + \arctan a_2$. The ordering of welfare levels would then correspond to the standard ordering of these real numbers. This new representation allows me to answer the preceding objection. Despite what the picturesque interpretation might suggest, there is no fundamental sense in which the welfare level $(1, 0)$ is adjacent to the neutral level $(0, 0)$. There are, indeed, infinitely many welfare levels in between them. There is no reason $(1, 0)$ cannot represent a high level of welfare.

The final point is that total lexic utilitarianism has a very desirable property in common with ordinary total utilitarianism: it is a *separable* axiology. Here is what that means. Suppose that populations A and B have a subpopulation C in common. Let A' be the rest of A, and let B' be the rest of B. Then separability says that A is at least as good as B if and only if A' is at least as good as B'. To see why this is plausible, suppose that C is on the other side of the universe from A' and B'. Or suppose that C existed a long time before A' and B'. It seems that in these cases, at least, we should be able to ignore C when comparing the merits of A and B. The intuition is particularly strong in so far as this comparison influences which population we should bring about. We should not have to worry about the unaffected welfare of far-off aliens or ancient Egyptians when making such a choice (Parfit, 1986, p. 420).

Indeed, many standard objections to particular population axiologies involve violations of separability. For example, suppose that A' consists of many people, all with good lives, and that B' is a smaller population of people with truly appalling lives. Average utilitarianism will rightly judge A' to be better than B'. But, contrary to separability, average utilitarianism may also judge B to be better than A (whether it does so depends on C).

This looks fatal for average utilitarianism and similar theories.[8] But total lexic utilitarianism, like ordinary total utilitarianism, is immune to such objections.

Having said all that, separability is not one of the adequacy conditions appearing in the impossibility theorems I will discuss. It is a strong condition,[9] and perhaps one could countenance some small violations. In particular, some egalitarians are happy to reject separability. They think that inequality in the population matters, and some natural ways of making it matter violate separability.[10] So I will henceforth put separability aside.

3 Against Small Steps

Now let me begin my analysis of the impossibility theorems. One strategy used by these theorems relies on the following principle.

[8] This argument essentially shows that average utilitarianism fails the 'non-Sadism' conditions used in Arrhenius's third, fourth, and sixth impossibility theorems. A comprehensive critique of average utilitarianism along similar lines is given in Hurka (1982a,b).

[9] In fact, any separable axiology allows for a quantitative representation of welfare, such that the value of a population is represented by the total welfare. (For this I assume that the preorder on one-person populations coincides with the preorder on welfare levels.) The 'quantities' involved may not be real numbers, however. They may be some more unfamiliar objects, such as, in our case, pairs of integers. See Pivato (2014) for a general result. To avoid the Repugnant Conclusion, a separable axiology can either (i) incorporate a lexical element, i.e. accept that there are positive quantities $a > b > 0$ such that no multiple of b is greater than a; or (ii) introduce a critical level, i.e. deny that neutral lives contribute zero to the welfare total. (They may contribute negative value, or a value incomparable to zero.) Critical level views face further problems (see e.g. Arrhenius (2000b) or Mulgan (2002)), and in that sense lexical views like total lexic utilitarianism occupy a special place in the constellation of population axiologies.

[10] See chapter 9 of Broome (1991). As I will argue in section 4, it is hard to say which population axiologies, in our very abstract sense, are egalitarian. Still, the underlying point is that the degree of inequality is not itself separable, however it is measured. By way of illustration, suppose that the lives in A' are at welfare level a, and the lives in B' and C are all at welfare level b. Then A' and B' are each perfectly egalitarian, but A is more unequal than B. Separability is nonetheless compatible with egalitarianism, as argued by McCarthy (2015).

> **Small Steps.** Any blissful welfare level a can be reduced to any lower drab level z in a finite sequence of small steps.[11]

Of course, 'small' is context-dependent. It invokes an implicit standard for smallness. The standard must be weak enough to make Small Steps true. But it must also be strict enough to make various adequacy conditions seem compelling. For example, here is a simplified version of Arrhenius's first impossibility theorem.[12] The simplified claim is that no population axiology can avoid the Repugnant Conclusion while satisfying

> **The Quantity Condition.** Suppose that a and b are positive welfare levels, that b is lower than a, and that the difference between a and b is small. Then, for any distribution at level a, there is a larger, better distribution at level b.

Informally, we should accept a small decrease in welfare levels in exchange for a sufficient increase in population size. It is easy to see how one might argue from the Quantity Condition to the Repugnant Conclusion. Starting from a distribution at a blissful level a, we should accept a small decrease in welfare levels in exchange for a sufficient increase in population size. But, according to Small Steps, a sequence of such small decreases can lead us from a to a drab level z. We should therefore accept a decrease in welfare levels from

[11] More formally: there is a finite decreasing sequence $a = a_1 > a_2 > \cdots > a_n = z$ of welfare levels such that the difference between consecutive terms is small. Arrhenius does not state Small Steps explicitly; he relies instead on Discreteness, below, and tends to speak of 'slight' differences.

[12] See Arrhenius (2000a, §10.3). His argument (including his version of the Quantity Condition) is more nuanced than the version I give here, as he is keen to make his premisses as weak as possible. I have eschewed certain subtleties in the interests of clarity, while preserving the features of the argument that I wish to discuss. A similar comment applies to my discussion of his other theorems.

a to *z* in exchange for a sufficient increase in population size. That is the Repugnant Conclusion. So no population axiology can satisfy the Quantity Condition *and Small Steps* while avoiding the Repugnant Conclusion.

The first, fourth, fifth, and sixth of Arrhenius's theorems follow this sort of strategy, proceeding through a sequence of small steps. (Recognising that the Quantity Condition is open to criticism, the later three theorems rely on the so-called Non-Elitism Condition, which I will discuss in section 4.) Each theorem is implicitly of the following form.

> Given any population axiology, there can be no standard of smallness according to which (a) Small Steps is true; and (b) certain intuitively compelling adequacy conditions are all true, perhaps including the Quantity Condition or the negation of the Repugnant Conclusion.

There are several types of objections one might make to such a theorem, while conceding the force of the underlying intuitions. One possible objection is that an argument via Small Steps has the character of a sorites argument. Although I think this objection has merit, its force is not immediately clear; for a critical discussion, see Temkin (2012, chapter 9). I will make some related comments in section 5. A second objection calls into question the way in which the adequacy conditions formalise the underlying intuitions. Let me say a little about this second objection, and then focus on the main line of thought: we can easily defuse the impossibility theorems by giving up Small Steps.

To see why the formal adequacy conditions might not properly reflect the underlying intuitions, consider the case of the Quantity Condition. As far as I can tell, the intuition is that any sufficiently small decrease in the quality of lives can be compensated by a sufficiently large increase in the quantity of lives. But that is not what the Quantity Condition actually says. The intuition as just stated is better represented by the weaker

> **Tradeoff Condition.** Suppose given a distribution at a positive welfare level a. Suppose that b is a lower welfare level, but still positive. If the difference between a and b is sufficiently small, then there is a larger, better distribution at level b.

The worry is that the Quantity Condition gains plausibility by conflation with the strictly weaker Tradeoff Condition. At any rate, it is not obvious to me that intuition supports the Quantity Condition over and above the Tradeoff Condition. The Tradeoff Condition is weaker because what counts as a 'sufficiently small' difference between a and b can depend on the size of the population at level a. The Quantity Condition, in contrast, requires there to be a single standard of smallness that works for all populations. Small Steps also presupposes such a universal standard. To see that this matters, consider Ng's axiology, which he calls 'Theory X'' (Ng, 1989). He assumes that welfare levels are represented by real numbers, with 0 as the neutral level. We might take level 100 to be blissful, and those between 0 and 2 to be drab. Ng then gives a rule for aggregating these numbers, with populations ranked by their aggregate scores. The rule, in one concrete version, is that a population of n people with average welfare a has aggregate score

$$(1-0.99^n)a.$$

It is easy to see that this axiology satisfies the Tradeoff Condition and avoids (as Ng argues) the Repugnant Conclusion. It also satisfies Small Steps, for any standard of smallness (see footnote 28 below). However, Arrhenius's first impossibility theorem rules out Theory X' because it does not satisfy the Quantity Condition for any standard of smallness. The failure of Theory X' to satisfy the Quantity Condition over and above the Tradeoff Condition does not, in itself, seem like compelling grounds for criticism. One can raise similar worries about the other adequacy conditions that refer to small differences.[13]

Now let me turn to my main objection. My main objection is that Small Steps is not itself compelling enough to be considered a basic adequacy condition. Faced with the impossibility theorems, the simplest answer is just to give up Small Steps.

First, let me show that the use of Small Steps is no mere convenience; without it, the impossibility theorem fails. Consider the example of total lexic utilitarianism that I introduced in section 2. I explained there that this axiology does not entail the Repugnant Conclusion. On the other hand, it does satisfy the Quantity Condition, for an obvious standard of smallness. The standard is that the difference between a and b is small if and only if a and b have the same amount of love.[14] (As I explained in section 2, there is

[13] I thank John Broome for first alerting me to this kind of objection, in personal communication.

[14] To see that the Quantity Condition applies, suppose that $a = (x, y)$ and $b = (x, z)$ are positive welfare levels whose difference is small. A population A of m people at level a has aggregate welfare (mx, my). If $x > 0$, then a population B of $m + 1$ people at level b has aggregate welfare $((m + 1)x, (m + 1)z)$; this is better than A. Or if $x = 0$, then y and z must both be positive. If we choose a number n such that $nz > my$, then a population B of n people at level b is better than A, as the Quantity Condition requires.

no reason to think of differences in love as in any way small.) Thus the only objection that the first impossibility theorem raises against TLU is that it violates Small Steps. In fact, TLU satisfies all the adequacy conditions (save Small Steps) that are required by Arrhenius's first, fourth, fifth, and sixth impossibility theorems.

Why, then, accept Small Steps? One might, with Arrhenius, endorse

> **Discreteness.** For any welfare levels a and b, there are at most finitely many welfare levels worse than a and better than b.[15]

According to Discreteness, one can get from a to b through a finite sequence of consecutive welfare levels. Presumably, the difference between consecutive levels counts as small by any standard. The question, though, is why we should believe Discreteness. Arrhenius has very little to say about this. He claims that the alternative to Discreteness is

> **Denseness.** For any welfare levels A and B, if B is better than A, then there is a welfare level that is better than A and worse than B.

Arrhenius finds Denseness 'improbable', and so favours Discreteness. The rejection of Denseness is already controversial, since it is very common to model welfare levels by real numbers. Be that as it may, Discreteness and Denseness do not exhaust the options. Total lexic utilitarianism satisfies neither. I have already explained why it does not satisfy Discreteness. It does

[15] My discussion here refers to Arrhenius (2011, §1.2), as well as to parallel sections in his other cited works. I have slightly simplified his formulations of Discreteness and Denseness below.

not satisfy Denseness, because each welfare level has an immediately higher one, consisting of the same amount of love and a penny more of money.[16]

Are there some other, more convincing arguments in favour of Small Steps? No doubt; I will mention one such argument in section 5. The point, however, is that Small Steps does not have the same status as the other premisses of the best impossibility theorems. It is not backed up by a strong normative intuition, like the one many people have against the Repugnant Conclusion. Nor is it a quasi-logical precondition for doing axiology, as transitivity is often thought to be.[17] It is not a basic adequacy condition, but something that requires support.

4 Against the Inequality Aversion Condition

Now let me consider the second basic strategy of the impossibility theorems. It involves the following adequacy condition.[18]

> **The Inequality Aversion Condition.** Suppose that a, z, b are welfare levels, with a higher than z and z higher than b. For any distribution A at level a, there are distributions Z at level z and B at level b, such that Z has the same size as $A \cup B$ and Z is better than $A \cup B$.

[16] Although Arrhenius favours Discreteness, he does not rely on its being true. He suggests that, even if Discreteness is not true, we can focus attention on some subset of all welfare levels in which Discreteness holds, and in which the difference between consecutive levels is small. But this claim is little more than restatement of Small Steps. It gives no new argument.

[17] See Broome (2004, Chapter 4) for a forceful statement of this view.

[18] Cf. Arrhenius (2000a, §10.5). Again, I have made some harmless simplifications in order to focus on the key issues.

Informally: the decrease of some people's welfare levels from a to z can be compensated by the increase of sufficiently many others' from b to z.

Before discussing the merits of this condition, let me explain how it is used in the impossibility theorems. The simplest argument appeals to the following principle.

> **The Mere Addition Principle.** Suppose that A and B are distributions containing only positive welfare levels. Then $A \cup B$ is at least as good as A.

The claim is that no population axiology can satisfy the Inequality Aversion Condition and the Mere Addition Principle while avoiding the Repugnant Conclusion. This is a simplified version of Arrhenius's second impossibility theorem.[19] Here is how the argument goes. Take a to be a blissful welfare level, z to be a drab level lower than a, and b to be a drab level even lower than z. (Recall that my official definition of 'population axiology' in section 2 stipulated that such levels exist.) For any distribution A at level a, the Inequality Aversion Condition gives us distributions Z and B, with Z at least as good as $A \cup B$. The Mere Addition Principle tells us that $A \cup B$ is better than A. So, by transitivity, Z is better than A. That is the Repugnant Conclusion.

[19] The theorem replaces the Mere Addition Principle by the 'Dominance Addition Condition'. While 'mere addition' simply adds the population B of positive lives, 'dominance addition' (or 'benign addition' in the terminology of Huemer (2008)) simultaneously improves the lives in A. Curiously, Kitcher's impossibility theorem (Kitcher, 2000) includes an adequacy condition ('DVA', p. 567) that directly denies the Mere Addition Principle as applied here: he insists that, if the Repugnant Conclusion is false, then adding drab lives to a large, blissful population decreases its value. This significantly reduces the interest of his theorem, since the Mere Addition Principle is widely seen as (at least) the default hypothesis, or even '*obviously* true' (Tännsjö, 2002, p. 357). Kitcher also relies on Small Steps, recognizing, however, that this may be problematic (§9).

Arrhenius's second, third, fourth, fifth, and sixth theorems elaborate this basic strategy. They replace the Mere Addition Principle with intuitive weakenings ('Dominance Addition' and two versions of 'Non-Sadism'). But they all rely on the Inequality Aversion Condition, either as a premiss or a lemma. It is a thread common to almost all the impossibility theorems.

Is the Inequality Aversion Condition intuitively compelling? To answer that question, consider how the condition is used in the argument just described. It is used to show that large penalties for some people (the people in A, reduced from blissful to drab lives) can be compensated by small benefits to sufficiently many others (the people in B, raised from one drab level to another). I think it is far from clear that we should favour such trade-offs. For example, can all-but-imperceptible benefits to sufficiently many people make up for the loss of all real joy in the world? If, as it seems to me, the answer is intuitively negative or unclear, we should not accept the Inequality Aversion Condition as a fundamental adequacy condition.

Perhaps one can argue for the Inequality Aversion Condition from intuitively compelling premisses. Arrhenius attempts two such arguments. As the name suggests, the basic thought is that the Inequality Aversion Condition is necessary for an appropriate degree of inequality aversion. Before considering the specific arguments, let me raise some general doubts.

It seems to me that the Inequality Aversion Condition has little to do with inequality aversion. Consider the two examples of ordinary and lexic total utilitarianism. On the face of it, these two axiologies express the very same attitude towards inequality. On either theory, adding one quantum of welfare to the population improves the population by one quantum, regardless of whether the quantum is given to a well-off person, a badly-off person, or to a new person. These theories are alike with respect to inequality aversion. But the Inequality Aversion Condition distinguishes between them: ordinary total utilitarianism satisfies the condition, while the lexic version does not.[20]

Thus the Inequality Aversion Condition makes distinctions for which there is no apparent egalitarian rationale. Of course, one might think that the condition nonetheless *correctly* rules out TLU. Then we should also rule out total utilitarianism on egalitarian grounds. But this cannot be correct. For it makes no sense to ask whether total utilitarianism, as I have defined it, is egalitarian or not. This is because I did not specify what sort of lives correspond to each numerical index of welfare. Whether or not the theory is egalitarian depends on this correspondence, which is not part of the formal axiology. More precisely, an axiology in the present sense only contains *ordinal* information about welfare levels (it tells us *whether* a is better than b) but not *cardinal* information (it does not tell us *how much* better a is). For example, in the present version of total utilitarianism, we should not suppose that welfare level 10 is better than welfare level 9 to the same degree that welfare level 9 is better than welfare level 8. All we know is that 10 is better than 9 and 9 is better than 8. In contrast, criteria for egalitarianism typically presuppose a cardinal scale.[21] For example, they require that it be a net

[20] Indeed, if again a is the blissful level and z and b are drab levels, then the population $A \cup B$ is always better than the population Z, according to total lexic utilitarianism.

[21] An exception to this rule is the criterion discussed by McCarthy (2015). But his criteria make fundamental use of uncertainty, which is again absent from the current framework. Indeed, the strongly separable egalitarian theories that McCarthy considers would be indistinguishable from total utilitarianism in the present framework.

improvement to decrease the welfare of a well-off person by a small amount while increasing the welfare of a worse-off person by that same amount. The sameness of the two amounts is a cardinal fact. Whether or not the present version of total utilitarianism is egalitarian depends on how the numerical indices of the welfare levels correspond to the cardinal facts that matter to egalitarians. If the cardinal difference between level n and level $n+1$ is independent of n, then the theory will be equality-neutral. (This is what I supposed in the previous paragraph: on an obvious interpretation, both total and total lexic utilitarianism are equality-neutral.) But if the difference increases as n increases, the theory will be egalitarian. It would make no sense to rule out total utilitarianism, in the present sense, on egalitarian grounds. More generally, it seems hard to say anything meaningful about egalitarianism within our purely ordinal framework.

With these general observations in mind, let me turn specifically to Arrhenius's two arguments for the Inequality Aversion Condition. The first argument (Arrhenius (2000a, §6.1), following Ng (1989)) starts from

> **The Non-Anti-Egalitarianism Principle.** A perfectly equal distribution is better than an unequal distribution of the same size and with lower total (and thus lower average) welfare.

It is important to note that the Non-Anti-Egalitarianism Principle requires a notion of total welfare.[22] The basic framework I introduced in section 2 does

[22] This in turn requires the sort of cardinal facts I mentioned earlier. If we can compare totals, we can compare differences. (For example, $a-b=c-d$ if and only if $a+d=b+c$.) The *motivation* for the Non-Anti-Egalitarianism Principle does not require talk of 'total' or 'average' welfare. (I thank Ralf Bader for pressing this point.) The idea is that the betterness relation might combine two values – let us call them 'utility' and 'equality'. The principle is

not include such a notion. Be that as it may, many people agree with Ng that the Non-Anti-Egalitarianism Principle is extremely compelling, and, like him, use it to derive the Repugnant Conclusion via the Inequality Aversion Condition.[23]

On the latter count, it is not true in general that the Non-Anti-Egalitarianism Principle entails the Inequality Aversion Condition. After all, I defined total lexic utilitarianism using a notion of 'total welfare', and with respect to that notion, it satisfies the Non-Anti-Egalitarianism Principle. But it does not satisfy the Inequality Aversion Condition.

Of course, proponents of the Non-Anti-Egalitarianism Principle invariably assume that we must represent welfare levels by real numbers. They define 'total welfare' in terms of the sum of such real numbers. On that assumption, the Non-Anti-Egalitarianism Principle does indeed entail the Inequality Aversion Condition.[24] This raises two questions. First, why real numbers instead of something else? Second, granting for the sake of argument that we have a real-numbered representation, why should we accept

that a population that is better with respect to utility and better with respect to equality is better overall. This formulation makes sense whether or not utility (whatever that is) can be summed. However, this more abstract version of the Non-Anti-Egalitarian Principle only slightly ameliorates the issues raised below.

[23] See e.g. Huemer (2008) and the 'first trilemma' of Carlson (1998). Carlson's second trilemma uses a similar argument to derive the 'Reverse Repugnant Conclusion': for any population of truly awful lives, there is a worse one consisting of lives only just below the neutral level. The derivation is a simple adaptation of Ng's to deal with negative welfare levels, and so faces the same worries. In fact, Carlson must appeal to an adequacy condition even stronger than Non-Anti-Egalitarianism: one population is better than another if it has higher total and higher average welfare.

[24] Here is the argument. Suppose that A has m people at welfare level a, and B has n people at welfare level b. Then the total utility of $A \cup B$ is $ma + nb$, while that of Z is $mz + nz$. Thus the latter has higher total utility (and the Non-Anti-Egalitarianism Principle says it is better than the former) so long as $n(z - b) > m(a - z)$. If a, b, z are real numbers, then this inequality will hold for all sufficiently large n, but it need not ever hold if they are (say) lexicographically ordered pairs.

the Non-Anti-Egalitarianism Principle?

All I need for my purposes is to note that neither of these questions admits a simple answer. That is enough to show that we should not take the Inequality Aversion Condition as a fundamental adequacy condition. For example, the most common way to produce a real-numbered representation of welfare levels is to argue that welfare satisfies the axioms of expected utility theory. But the axioms of expected utility theory, while rather plausible, are subject to dispute. In any case, by introducing uncertainty, they take us far beyond the minimal framework of the impossibility theorems. Then, once we have tied down a real-numbered representation, we can talk about total and average welfare. But at this stage there is no reason to think that *these* notions of 'total' and 'average' are directly relevant to population axiology, let alone relevant in the precise way described by the Non-Anti-Egalitarianism principle. Indeed, different real-numbered representations of welfare levels lead to different, mutually incompatible versions of the Non-Anti-Egalitarianism Principle. To see this, suppose that I assign to each welfare level a a numerical value $u(a)$. If, in a certain population, the people have welfare levels $a_1, \ldots, a_n$, then I interpret the 'total welfare' to be the sum $u(a_1) + \cdots + u(a_n)$. That yields one form of the Non-Anti-Egalitarianism Principle. But now suppose I assign numbers in a different way, according to a function v, and interpret the 'total welfare' as $v(a_1) + \cdots + v(a_n)$. That yields a different, and contradictory form of the Non-Anti-Egalitarianism Principle which contradicts the first. Which one of these two principles is supposed to be intuitively compelling? It is impossible to tell without further information. It may be that for *some particular* choice of representing function u, the Non-Anti-Egalitarianism Principle is compelling. But we need an explanation of what that choice is, and an argument for why the principle is compelling for that, as opposed to other, choices.[25]

Here is an analogy. Consider the weights in grams of various coins. The weights are real numbers. I can very well add them up to find the total weight of a pile of coins. But the total weight of a pile of coins tells me very little about its value in the ordinary sense. (This is true even if we suppose that, the heavier a single coin, the greater its face value.) So too, the mere fact that we have assigned real numbers to welfare levels (thus measuring welfare on a 'ratio scale') does not mean that the total of these real numbers carries any particular significance.

Here is a final, less technical worry about the Non-Anti-Egalitarianism Principle. The principle concerns cases in which considerations of equality, total utility, and average utility coincide. The thrust of the principle is that, in these cases, no additional considerations can make a difference. But that is not obvious. For example, perhaps it is also relevant how many lives are above a certain level of sufficiency. In particular, in our application of the Inequality Aversion Condition, populations like $A \cup B$ contain many high-quality lives, while populations like Z contain none. When thinking about the Repugnant Conclusion, a natural idea is that this very fact has

[25] The abstract version of the Non-Anti-Egalitarianism principle mentioned in footnote 22 suggests one kind of argument. It might be claimed that betterness with respect to utility is separable, and therefore can be represented by a 'total welfare' formula (cf. footnote 9). This notion of total welfare would, by definition, validate the Non-Anti-Egalitarianism Principle – at least assuming that only utility and equality matter to population axiology. To derive the Inequality Aversion Condition, one would still need to argue that total welfare was given by real number addition. An argument along these lines for the significance of total utility in the sense of expected utility theory is provided by Broome's development of Harsanyi's famous aggregation theorem (Broome, 2004; Harsanyi, 1976).

overriding axiological significance. More generally, considerations of utility and equality may, in some cases, be overridden by others. This possibility undermines the Non-Anti-Egalitarianism Principle.

Now let me turn to Arrhenius's second (and rightly favoured) argument for the Inequality Aversion Condition (Arrhenius, 2000a, §6.3). He claims to deduce it from

> **The Non-Elitism Condition.** Suppose that a, z, b are welfare levels, with a better than z and z better than b, and that that the difference between a and z is small. Then, for some number n, it is always a net improvement to reduce one person's welfare from a to z while increasing that of n others from b to z.[26]

I agree that this condition sounds compelling. It should also be clear, at least in outline, how the argument from Non-Elitism to Inequality Aversion goes. The basic difference between the two conditions is that the Non-Elitism Condition only allows us to compensate for the loss of a *small* amount of *one* person's welfare, while the Inequality Aversion Condition allows us to compensate for the loss of a *large* amount of *many* people's welfare. But a sequence of small losses to one person at a time can amount to large losses

[26] Arrhenius has two versions of the Non-Elitism Condition (one of them 'General'), but the difference between either of them and the version I have given here is nugatory. Parfit's arguments (1986, §142ff) also use a version of the Non-Elitism Condition: small losses to some are compensated by at least as large gains to others. (Parfit justifies such compensation using heuristics about utility and equality akin to the Non-Anti-Egalitarianism Principle; these heuristics inspired Ng's work.) He does not derive the Inequality Aversion Condition *per se*, but my objection applies to his argument as well, since it relies on a series of small steps. The second and third arguments for the Repugnant Conclusion in Tännsjö (2002) are variations on Parfit. (Tännsjö's first argument uses the Quantity Condition as in section 3, following Arrhenius.)

for many people. Thus recursive application of Non-Elitism leads to the Inequality Aversion Condition.[27]

However, this argument implicitly requires Small Steps: it assumes that a sequence of small decreases in welfare can amount to a large one. If, as I have already suggested, we deny Small Steps, the argument from Non-Elitism to Inequality Aversion fails. Indeed, total lexic utilitarianism does not satisfy the Inequality Aversion Condition, but it *does* satisfy the Non-Elitism Condition. Recall that the difference between a and z is 'small' only if they have the same amount of love. Suppose then that a and z differ only by m units of money. Since the difference between z and b is at least one unit of money, a decrease in the welfare of one person from a to z can be compensated by an increase in the welfare of $m+1$ people from b to z.

In conclusion, the Inequality Aversion Condition is not clearly related to inequality aversion. It is not strongly supported by direct intuition. Nor is it, in general, a consequence of the Non-Elitism Condition or of the Non-Anti-Egalitarianism Principle, and the latter principle has problems of its own.

5 Small Steps and Indeterminacy

Recall the story so far. I have argued that the Inequality Aversion Condition should not be one of the fundamental adequacy conditions of population axiology. On the other hand, this condition follows from the much more

[27] Elabourations of this argument appear as lemmas in the proofs of the fourth, fifth, and sixth theorems. See Arrhenius (2000a, Lemma 5.1.1; 2003, Lemma 1; 2011, Lemma 1.1.1).

compelling Non-Elitism Condition, if we accept Small Steps. We have also seen another impossibility theorem, based on the Quantity Condition, which employs Small Steps. I have argued that the assumption of Small Steps is no harmless technicality, nor is it clearly justified. Still, it would be uncomfortable to pin the hopes of population ethics on the falsity of Small Steps. Small Steps follows if the welfare levels are ordered like the real numbers or the integers.[28] Many of the quantities we detect in the world around us have that sort of structure, and it is (if nothing else) often assumed that welfare is the same.

One can say more in favour of Small Steps. That is not the project of this paper, but here is the kind of argument that I find most compelling. For simplicity, let us be hedonists. (Non-hedonists can tell a similar story, but there may be complications.) Stipulate that a life is 'blissful' if it consists of one hundred years of intense pleasure followed by a single neutral minute, completely devoid of pleasure and of pain. Stipulate that a life is 'drab' if it begins with one minute of that same intense pleasure, followed by one hundred neutral years. (These stipulations are appropriate as long as the Repugnant Conclusion seems repugnant when understood in terms of these kinds of lives.) Then we find a natural continuum between drab lives and blissful ones, as we lengthen the initial period of pleasure from one minute to one hundred years. Consider two lives on this continuum that differ by only

[28] For the real-number case, I assume that, for any real number x, the real numbers that differ from x by a 'small' amount include all those in some open interval around x. The Heine-Borel theorem says that any closed, bounded interval of real numbers is contained in the union of finitely many of these small intervals. Thus any closed, bounded interval can be traversed in a finite number of small steps. Note that this argument depends on the assumption that *all* real numbers in an appropriate interval correspond to welfare levels; it will not work if there are gaps.

one millisecond's worth of pleasure. The difference in welfare between two such lives is surely 'small' in the relevant sense. But then there are finitely many small steps between a blissful welfare level and a drab one, since there are finitely many milliseconds between one minute and one hundred years.[29]

Given the plausibility of such arguments, the impossibility theorems do pose genuine difficulties for population axiology. We should presumably aim for some kind of reflective equilibrium between basic intuitions and more theoretical considerations. In concluding this paper, I want to highlight one possible ingredient in this equilibrium that has not been widely addressed in the literature: the possibility of vague or otherwise indeterminate axiologies. Such indeterminacy has been considered before, but only in a limited way. For example, Broome (2004) advocates a version of total utilitarianism with a vague critical level. As he recognises, this move only partly mitigates the impact of the impossibility theorems. I suggest that vagueness has a more general role to play in balancing competing intuitions.[30]

Why is vagueness relevant at all? Let me begin with an analogous case. Suppose I believe that Fred – whom I have never met – is tall; I know precious little else about him. I walk into a room, certain that Fred will be there. But there are *two* men present. Which one is Fred? The first man I see is determinately not tall. If the second man were determinately tall, I would infer that he was Fred. But the second man is only borderline tall. Still, all else

[29] A similar argument can be run with probabilities instead of times: vary the *chance* of pleasure, rather than its *length*. The key point is that probabilities, like times, standardly have the structure of a real continuum, and this has implications for the structure of welfare. Indeed, this is essentially the strategy of expected utility theory for identifying welfare levels with (some) real numbers, as mentioned in section 4.

[30] I develop these ideas from a different angle in Chapter II of this thesis.

equal, I will be inclined to think that the second man is Fred. This matches my prior beliefs and my new evidence better than the alternative. Moreover, some borderline tall people are taller than others. Some are borderline tall but *almost* determinately tall. The closer the second man is to being determinately tall, the more inclined I will be to believe that he is Fred. So too in the case of population axiology. If I think the Quantity Condition (for example) is compelling, then, all else equal, I should be inclined to favour a theory according to which that condition is borderline, but almost determinately true, over a theory according to which it is determinately false.

This picture is strengthened if we observe that the impossibility theorems that use Small Steps are structurally similar to sorites arguments.[31] The impossibility theorem I discussed in section 3 repeatedly invokes the Quantity Condition to derive the Repugnant Conclusion. Similarly, a sorites argument might claim to prove that every tree is tall by repeatedly invoking

> **The Tolerance Condition.** Suppose that a and b are heights, that b is lower than a, and that the difference between a and b is small (less than one millimeter, say). Then it cannot be the case that a is tall for a tree and b is not.

Theories of vagueness that respect classical logic must accept that there are counterexamples to the Tolerance Condition. But they must also explain the strong intuition in its favour. A typical explanation is that the Tolerance Con-

[31] I will develop and defend this analogy in other work. For a critical view, see Temkin (2012, Chapter 9). Here I only rely on a broad similarity to amplify the considerations of the preceding paragraph. In my informal survey, a few philosophers resist the very idea of moral vagueness. But many others take it to be an obvious and widespread phenomenon: see Constantinescu (2014); Dougherty (2014); Schoenfield (2015) for recent examples.

dition has no *determinate* counterexamples. Every instance of the Tolerance condition is at least borderline true, and indeed close to determinately true.[32] If this kind of story explains why the Tolerance Condition is compelling, then it may help with the Quantity Condition and the Non-Elitism Condition as well. Even if these conditions admit counterexamples, they need not admit determinate counterexamples. That may go some way towards explaining their attraction.

Let me now illustrate these general considerations with a toy model.[33]

In this axiology, welfare levels are indexed by integers. Let us suppose that 0 is the neutral level, 1 and 2 correspond to drab lives, and 100 to a blissful life. Small Steps is bound to hold, assuming that the difference between consecutive welfare levels counts as 'small'. The ranking of populations will have a sufficientarian flavour. There is some positive welfare level S, the level of sufficiency, above which life is 'very good'; lives below $-S$ are 'very bad'. Populations are ranked, in the first instance, by the number of very good lives minus the number of very bad lives. Then ties are broken by total welfare, the sum of integers.

Does such an axiology entail the Repugnant Conclusion? No – not as long as the blissful lives are above S and the drab lives are below it. By the first impossibility theorem (section 3) we know that the Quantity Condition must fail. Indeed, it fails when, and only when, the small decrease in welfare

[32] See e.g. Keefe (2000, pp. 185–6) in the case of supervaluationism. Note that for Keefe, as for many supervaluationists, plain-old-truth is what I have called determinate truth. For a theory that emphasises closeness to determinate or 'clear' truth, see Edgington (1996). Of course, other treatments of the sorites are available, including treatments that forgo classical logic; I cannot give a survey here.

[33] The model here resembles the views defended by Qizilbash (2005) and Knapp (2007), and also has some affinity to the 'Imprecise Lexical View' sketched by Parfit (2016).

from a to b brings us from a life that is very good to one that is not. But this is where vagueness can soften the blow. If the level of sufficiency is vague, then there are no consecutive welfare levels a and b such that, determinately, a is very good and b is not. Thus the Quantity Condition has no determinate counterexamples. Each instance is at worst borderline true, failing on at most one of the many possible precisifications of S.

For the same reason, the Non-Elitism Condition has no determinate counterexamples. Indeed, of the adequacy conditions appearing in Arrhenius's theorems, only one is entirely invalidated by this axiology. That is the Inequality Aversion Condition, which, I have already argued, we should not accept on its own merits.

6 Conclusion

Narrowly construed, the aim of the impossibility theorems is to establish a conflict between core axiological intuitions, given only minimal background assumptions like the transitivity of 'better than'. As I have argued, they fail in that aim: first, because the background assumption of Discreteness is unjustified, as is, consequently, the use of Small Steps; second, because the Inequality Aversion Condition does not properly reflect egalitarian concerns, and could reasonably be rejected.

More broadly construed, however, the theorems remain important: they insightfully illustrate the difficulty, if not the impossibility, of constructing a satisfactory population axiology. From that point of view, the aim of this paper has been to focus attention on some ways forward that respect as far as possible the adequacy conditions of these theorems. We can reject Small Steps, or, if driven to accept it, appeal to axiological vagueness to mitigate the inevitable conflict of intuitions.

II | Vague Spectra

ABSTRACT. I explore the relationship between two types of paradoxical argument: sorites arguments for predicates like 'tall' and spectrum arguments for comparatives like 'better than'. It has often been claimed that spectrum arguments are structurally different from sorites arguments. I argue that vagueness may still be the source of paradox in each case. I show that for each spectrum argument there is a closely related argument indisputably of sorites form. It is natural to think that this latter argument is indeed a sorites argument: vagueness explains the intuitive force of its premisses, even though the argument is unsound. But then vagueness must also lie at the heart of the original spectrum argument.

1 Introduction

Derek Parfit's 'Repugnant Conclusion' (RC) claims that, for any possible population in which every life is excellent, there is a better one in which every life is barely worth living (Parfit, 1986).[1] The Repugnant Conclusion is widely, although not universally, regarded as unacceptable. There are, nonetheless, many arguments in favour of RC. Some of these are based on relatively specific axiological theories. For example, the most obvious versions

[1] Parfit's original formulation of RC contains a *ceteris paribus* clause. I prefer to understand 'populations' as abstract welfare distributions falling under a relation of intrinsic betterness, in which case the *ceteris paribus* clause is unnecessary. But the reader may reformulate to taste.

of total utilitarianism entail RC. I am going to consider some arguments for RC which, in contrast, appeal to very general intuitions about the structure of the good.

These so-called 'spectrum arguments' were pioneered by Parfit and honed to an art by Gustaf Arrhenius (2013). They have also been pressed, from a slightly different vantage point, by Stuart Rachels (2004) and Larry Temkin (2012).[2] I think that everyone in this literature has noticed – but also, in general, rapidly dismissed – the objection that these spectrum arguments are at least superficially similar to sorites arguments, and that, therefore, they might be unsound in the same way that sorites arguments are unsound. Just as sorites arguments confirm that tallness and redness and baldness are vague, so too, the thought goes, spectrum arguments do no more than confirm that betterness is vague. We should not accept that everyone is tall, or that there are no heaps, on the basis of the sorites; no more should we accept RC on the basis of the spectrum arguments. Call this *the indeterminacy thesis*. Despite having occurred to almost everyone, I do not think that the indeterminacy thesis has been properly spelled out. And once it is spelled out, it is – so I shall argue – very hard to dismiss.

Let me introduce my argument by considering two preliminary issues.

First, since the most common examples of vague predicates, like 'tall', are unary predicates rather than relations, it may not be obvious what it means to claim that 'better than' is vague. The simplest way to think about

[2] In this paper I will focus on the Repugnant Conclusion to keep things relatively concrete. But spectrum arguments similar to the ones I will discuss have been put forth in many different contexts – see Temkin (2012, ch. 2 and 5) for a survey – and almost everything I say could be adapted to the general case.

this is that, for some precise state of affairs X, the unary predicate 'better than X' is vague.[3] This helps us navigate the following common point of confusion. One might think that 'tall' is vague, but the comparative 'taller than' is precise; in analogy, one might think that 'good' is vague, but 'better than' must be precise. What I agree with in this thought is that often 'good' means something like 'better than the standard', and one reason why 'good' is vague is that the standard is vague.[4] Even if 'better than' *were* precise, 'good' would still be vague, since the standard would be vague. But this does not show that 'better than X' is precise when X is precise. For example, it might be indeterminate whether X includes some particular good-making feature. Or there might be numerous factors with respect to which something can be better than X, and it might be indeterminate how these factors weigh up. Some non-normative comparatives, like 'hairier than', exhibit vagueness of both these kinds. It is often a vague matter which of two precisely specified heads is hairier than the other. In some cases, it is indeterminate whether one of the heads instantiates a particular 'hairy-making feature' (e.g. some hairs are semi-detached and it is indeterminate whether they count); in other cases, it is indeterminate how various factors weigh up (e.g. the number of

[3] I assume that betterness, in the relevant sense, is a relation between states of affairs. By a 'precise' state of affairs I mean something like an exactly specified microphysical state; it should not admit any borderline cases of instantiation. The point, as I discuss below, is to rule out cases in which 'better than X' is vague just because X is vague.

I will also consider some more complicated unary predicates derived from 'better than', but 'better than X' indicates the main idea. In general, one should presumably say that a binary relation is vague if and only if it is vague as a unary predicate of ordered pairs. If R admits sorites series of pairs, it may not be true that, for some x, the unary predicate $R(-, x)$ admits sorites too. For example, consider the relation on natural numbers such that $R(x, y)$ is true if and only if x is small and $y = x$. This is intuitively a vague relation, and indeed admits a sorites series $(1,1), (2,2), (3,3), \ldots$. But, for any fixed x, the unary predicates $R(x, -)$ and $R(-, x)$ do not.

[4] See DeRose (2008) for one careful discussion of the semantics of gradable adjectives.

hairs, the length of the hairs, and the way they are distributed). Similarly, 'taller than' is vague in at least the first of these ways. So it is simply irrelevant that 'better than' is a binary predicate, or that *some* paradigmatic cases of vagueness arise through a comparison to a vague standard.[5]

Here is the second preliminary point. Part of the indeterminacy thesis is that certain arguments with the logical form of sorites arguments are genuine sorites arguments, i.e. they fail in a characteristically vagueness-related way. But this thesis does not at all suppose or suggest that *every* argument of the sorites form is a genuine sorites. For example, as far as logical form goes, a sorites argument is essentially a case of mathematical induction. In that sense, the usual proof that the sum of the first n odd numbers equals n^2 has the same form as a sorites argument. But this similarity is no reason to think that the proof is unsound.[6] There are three main factors that distinguish an arithmetical case like this from the normative case at hand. Noting them here will help to set up the dialectic.

First, we are quite generally disinclined to think that arithmetic might be vague.[7] In contrast, moral discourse is almost always carried out in vague terms: things that people deem to matter, like *personhood* or *happiness*, are bound to admit borderline cases. There are, besides, many different morally relevant respects in which one state of affairs might be better than another. It is at least plausibly vague how these respects weigh up. Thus

[5] Temkin (2012, §9.2.3), following Ryan Wasserman, considers spectrum-like arguments for 'hairier than' and other non-normative predicates. To the extent that those arguments are paradoxical, the discussion in this paper can be adapted straightforwardly to them.

[6] I thank Theron Pummer for pressing this point.

[7] I set aside here the idea that there might be some sort of indeterminacy in the more arcane corners of mathematics – for example, concerning the truth of the axiom of choice. This purported indeterminacy does not generate sorites series.

many have thought moral vagueness to be pervasive, and have put forward sorites arguments as evidence of this fact.[8] The indeterminacy thesis need not overcome a general presumption against vagueness.

Second, although it might at first be surprising that the sum of the first n odd numbers is n^2, this is because it appears *unlikely* – who would have guessed! – rather than *repugnant*. There is no countervailing intuition that we cannot rather quickly set aside. In contrast, the Repugnant Conclusion strikes many as false, and not merely unlikely.[9] The spectrum arguments, and not the arithmetical induction, have the air of paradox. And it is only *paradoxical* sorites-like arguments that indicate vagueness. To put it another way, the indeterminacy thesis is supported by inference to the best explanation. But if a sorites-like argument is not paradoxical, then there is nothing to explain, and the inference fails.

Finally, the axioms of arithmetic are more fundamental than any hypotheses about the sum of odd numbers, to the extent that the axioms are sometimes considered definitive of their subject matter. In contrast, the negation of RC is roughly on a par with the kind of premisses from which RC is purportedly derived. So there is more room than in the arithmetical case to think of a spectrum argument as a *reductio* of its premisses.

As these comments should make clear, the indeterminacy thesis will

[8] See Shafer-Landau (1995) and Constantinescu (2014), Dougherty (2014), Schoenfield (2015), and Dunaway (2016) for recent examples.

[9] It is worth mentioning here that although I will focus on RC, some of the spectrum arguments lead to even more striking conclusions. For example, the*Very Repugnant Conclusion* (Arrhenius, 2003): any world in which every life is excellent would be worse than some world in which there are vastly many more lives that are full of unmitigated suffering, and all the other lives are only just worth living. The indeterminacy thesis applies to these spectrum arguments as well.

have only indirect interest to those who accept the Repugnant Conclusion independently of the spectrum arguments, and for whom, consequently, there is no paradox. However, as it turns out, the philosophers who have most emphasised the spectrum arguments have also tended to see the denial of RC (or at least some related condition) as non-negotiable, and have therefore found these arguments profoundly worrying. While Parfit is steadfast in his hope to find an adequate 'Theory X' that avoids RC, Arrhenius sees the spectrum arguments as 'impossibility theorems', threatening to undermine the basic methodology of moral philosophy. Temkin and Rachels have a more specific, but perhaps equally radical, diagnosis: they take the spectrum arguments to show that 'better than' is not a transitive relation. This itself threatens to leave axiology and practical reason in disarray.

In such a context, the indeterminacy thesis offers an extremely conservative way forward. It affirms (or is at least compatible with) transitivity, and revises the inductive premisses of the spectrum arguments only to the extent that the right theory of vagueness revises the inductive premisses of sorites arguments.[10] I will have much to say about this in section 4. The bottom line, though, is that (on most views!) we constantly and successfully reason with vague concepts. The vagueness of 'better than' is no threat to the general cogency of axiological thinking.

Here is an outline of the paper. My first goal is to develop the indeterminacy thesis and establish a positive case for it. In section 2, I introduce

[10] Of course, some approaches to vagueness affirm the premisses of the sorites argument, but call into question its validity. I will also consider this sort of move, but it is convenient mainly to speak in terms of revisions of the premisses.

the basic spectrum argument and discuss its main premisses. In section 3, I explain the formal relationship between spectrum arguments and sorites arguments. In section 4, I explain why, given this formal relationship, the indeterminacy thesis is hard to dismiss. I also contrast indeterminacy with cognates like parity and imprecise equality.

The thesis of indeterminacy does not involve a commitment to any particular axiology, or, indeed, to any particular theory of vagueness. My second goal is to illustrate how the view works under some further assumptions. First, in section 5, I specialise to theories of vagueness like supervaluationism that adhere fairly closely to classical logic. Then, in section 6, I specialise even further by providing a toy model.

I next use this discussion to address some objections. The most explicit and best-developed objection to the indeterminacy thesis is the one in chapter 9 of Temkin (2012). He claims that there is a basic structural disanalogy between spectrum arguments and sorites arguments. I explain why this argument misses its mark in section 7. Then, in section 8, I consider a more elaborate spectrum argument based on an assumption of 'non-elitism'. I explain how the indeterminacy thesis applies to this kind of spectrum argument, and suggest two ways it can rebuff the charge of elitism.

Finally, in section 9, I recap and re-evaluate how my case for the indeterminacy thesis shifts the dialectical balance.

2 The Basic Spectrum Argument

Let us look at the basic spectrum argument (Figure 1).

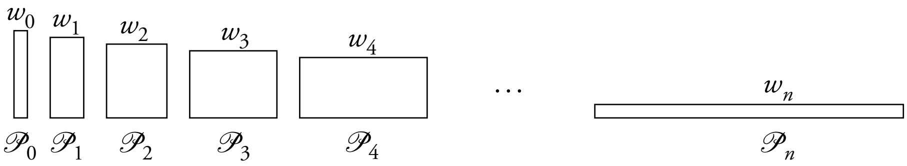

Figure 1: The populations $\mathscr{P}_i$ at welfare levels w_i.

As usual in this literature, each rectangle represents a population; the height represents the quality of life in the population, and the width represents the number of lives. We first consider (on the left) a population $\mathscr{P}_0$ of high-quality, henceforth 'blissful' lives, corresponding to a welfare level w_0. We construct a sequence of better and better populations, $\mathscr{P}_1$, $\mathscr{P}_2$, and so on. Given $\mathscr{P}_i$, we construct $\mathscr{P}_{i+1}$ by decreasing the quality of life a small amount, from w_i to w_{i+1} – the difference corresponding, perhaps, to a single grain of chocolate – while adding many more new people at the lower level. The idea is that we lose some value insofar as the quality of life falls, but we gain even more, because we gain so many new good lives.

Thus the argument relies on a principle which, following Arrhenius, I call

> **The Quantity Condition**
>
> Suppose that w_i, w_{i+1} are positive welfare levels, such that w_{i+1} is lower than w_i, but the difference between them is small. Then, for any number $N > 0$,
>
> (QC) Any sufficiently large population $\mathscr{P}_{i+1}$ of lives at level w_{i+1} would be better than a population $\mathscr{P}_i$ of N lives at level w.

By applying principle QC sufficiently many times, we end up with a population $\mathscr{P}_n$, on the right, in which all the lives are 'drab', or only just worth living. We make things better and better and better, so we reach the repugnant conclusion: for any initial population $\mathscr{P}_0$ of blissful lives, at level w_0, there is a better population $\mathscr{P}_n$ of drab lives, at level w_n.

Before comparing this argument to a sorites argument, let me clear up the status of its premisses. There are two important implicit premisses. First, we have relied on some sort of finiteness or boundedness condition: we have assumed that it is possible to get from the blissful level w_0 to the drab level w_n by a finite number of 'small' decrements. This is open to question, but I leave that discussion for other work (Chapter I of this thesis). Second, we have assumed that the relation 'better than' is transitive. We have said that each population is better than the one before, but to get that the last population is better than the first, we need transitivity. Some people think that transitivity is a logical feature of grammatical comparatives like 'better than'.[11] On the other hand, Temkin and Rachels have taken spectrum arguments like this one to provide a *reductio* of transitivity. Whether or not the appeal to grammar is sufficient, abandoning transitivity strikes many of us as a desperate move. It is the kind of desperate move that the indeterminacy thesis allows us to avoid.

Finally, the main explicit premiss, the Quantity Condition, is also controversial. It is controversial how, if at all, the size of a population contributes to its value. For example, average utilitarianism has its *prima facie* attractions. But it denies that any population at level w_{i+1} can be better than one at level

[11] See Broome (2004, §4.1) for a forceful statement of this view.

w_i, if w_{i+1} is lower than w_i. So average utilitarianism denies QC. Similarly, many people are attracted to some version of the idea that adding good lives to a population is morally neutral.[12] Suppose we elevate that to an axiological principle: adding good lives to a population does not make it better. Since $\mathscr{P}_i$ is surely better than a same-size population $\mathscr{P}$ at level w_{i+1}, and adding good lives to $\mathscr{P}$ to get $\mathscr{P}_{i+1}$ does not make it better, it seems that $\mathscr{P}_{i+1}$ cannot be better than $\mathscr{P}$. (There are many subtleties here; I wish merely to point out the basic difficulty.)

There are two things to say about this. First, there are some more complicated spectrum arguments which arguably circumvent the controversy. I would like, until section 8, to bracket those complications and focus on this simplest case. Second, isomorphic arguments are sometimes made in less controversial settings. For example, Temkin and Rachels present an argument that any pain, no matter how long and torturous, is better than a mild but sufficiently long headache. The argument proceeds on the basis of the following analogue of QC:

> Suppose that I_i, I_{i+1} are levels of pain intensity, and the difference between them is small. Suppose that $\mathscr{P}_i$ is a pain of constant intensity I_i. Then a sufficiently long pain of constant intensity I_{i+1} would be worse than $\mathscr{P}_{i+1}$.

This is much less controversial than QC. At least, a pain of fixed intensity is undoubtedly worse the longer it goes on. The main points of this paper

[12] In Narveson's famous formulation, 'We are in favor of making people happy, but neutral about making happy people' (1973, p. 80).

apply, *mutatis mutandis*, to the more complicated arguments in population ethics as well as to the argument about pain.

3 Sorites Arguments

Now I will explain the formal connection between the basic spectrum argument and sorites arguments. It will help to have some standard examples of sorites arguments in mind. A sorites argument about tallness might go like this. We have a sequence of one thousand people, p_1, p_2, and so on, each one millimeter shorter than the last. Here are the two premisses (I present the second as a schema, a choice that will become relevant in section 5).

> (A) The first person, being 2m tall, is tall.
>
> (B) If p_i is tall, then p_{i+1} is also tall.

One millimeter never makes the difference between tall and not tall. But then, by repeated modus ponens, we may conclude

> (C) The last person, p_{1000}, is tall, despite being only 1m in height.

More generally, one can prove in this way that everyone of whatever height is tall. The conclusion is absurd, and almost no one thinks that we should accept it on the basis of this argument.

Let me give a second example of a sorites argument.[13] This second example provides a typical example of axiological vagueness. We consider an embryo at different stages in its development, going one second at a time.

> It would be better to destroy the one-day-old embryo than to cut off my thumb.
>
> If it would be better to destroy the embryo at one time than to cut off my thumb, it would still be better to destroy the embryo one second later than to cut off my thumb.

But then

> It would always be better to destroy the embryo, even as a 39-week-old foetus, than to cut off my thumb.

This seems to be a completely standard sorites argument for 'better to x than to cut off my thumb'. The existence of sorites arguments in a normative context should not, after all, be surprising. If anything, the orthodoxy is that few things are completely precise outside of mathematics and fundamental physics.

Now, the general resemblance between the spectrum argument and the sorites arguments is obvious: they all involve a long sequence of steps. However, it is perhaps less obvious what exactly the formal relationship is between them. Let me explain that now.

[13] Roughly the same example is given in Schoenfield (2015); she works with permissibility rather than betterness.

In fact, there are at least two ways of drawing the connection. As we will see, the second way makes the analogy somewhat clearer, but let me start with the way that has been most considered in the literature. On this first way of doing things, the putatively vague predicate, the analogue of 'tall', is 'better than $\mathscr{P}_0$'. So a population is F (let us say) if and only if it is better than $\mathscr{P}_0$. We have by construction

(A1) $\mathscr{P}_1$ is better than $\mathscr{P}_0$.

(To make (A1) even more convincing, we could replace it by the tautological '$\mathscr{P}_0$ is at least as good as $\mathscr{P}_0$', although that would not, strictly speaking, be of the form 'x is F'.) We also have

(B1) If $\mathscr{P}_i$ is better than $\mathscr{P}_0$, then $\mathscr{P}_{i+1}$ is better than $\mathscr{P}_0$.

Why? By construction, $\mathscr{P}_{i+1}$ is better than $\mathscr{P}_i$; so, by transitivity, if $\mathscr{P}_i$ is better than $\mathscr{P}_0$, then $\mathscr{P}_{i+1}$ is also better than $\mathscr{P}_0$.

Finally, iterating modus ponens, we deduce

(C1) The last population, $\mathscr{P}_n$, is better than $\mathscr{P}_0$.

That is one way to put the spectrum argument in the form of a sorites argument. However, there is a different way of doing it, which is I think more fundamental. On this second way, instead of considering the sequence of populations, we consider the corresponding sequence of welfare levels. The predicate in question is 'G' defined by the following biconditional.[14]

[14] Qizilbash (2005) explicitly considers this predicate, and the view he sketches resembles the toy model I describe in section 6.

> **(Predicate *G*)** A welfare level w is G if and only if a population of sufficiently many lives at level w would be better than $\mathscr{P}_0$.

Recall that w_1 through w_n are the welfare levels of the lives in populations $\mathscr{P}_1$ through $\mathscr{P}_n$. First of all, we have

> (A2) The first welfare level, w_1, is G.

(We could, again, replace (A2) by a more tautologous claim: a population of sufficiently many lives at level w_0 would be at least as good as $\mathscr{P}_0$.) Then we have the inductive premiss,

> (B2) If w_i is G, then w_{i+1} is also G.

Why is that? According to QC, given any population at level w_i, a sufficiently large population at level w_{i+1} would be better. So, given a population at level w_i that is better than $\mathscr{P}_0$, a sufficiently large population at level w_{i+1} would be better still.

Therefore, we can deduce

> (C2) The last welfare level, w_n, is G.

That is just to say that sufficiently many drab lives would be better than the blissful population $\mathscr{P}_0$ – whence the Repugnant Conclusion.

The two versions of the argument, (A1,B1,C1) and (A2,B2,C2), are very closely related. But the first form of the argument slightly obscures the logic, and is less clearly analogous to a standard sorites. Let me explain why.

Suppose that one tried to deny some instance of premiss (B1). This would be confused: the populations $\mathscr{P}_i$ are *constructed* to make (B1) true.

The only thing that we can do is to deny that the construction succeeds. In contrast, one can straightforwardly deny (B2). So the second form of the argument makes the logic clearer.

It is also more clearly analogous to the standard sorites. Consider again the case of tallness. We have an underlying parameter, the height, which is precise for all intents and purposes.[15] This parameter changes gradually from one case to the next, and whether or not someone is tall depends only on this parameter; tallness is a matter of height. That is also what is happening in the second form of the spectrum argument. We have a single precise parameter – namely, the welfare level – that varies gradually from one case to the next, and whether the predicate G applies is, again, simply a matter of that parameter. But the first form of the spectrum argument is a little bit different. The cases $\mathscr{P}_1$ through $\mathscr{P}_n$ differ along two dimensions. As before, there is gradual variation in one parameter, the welfare level, but there is also potentially massive variation in the size of the populations. And whether or not a population is better than $\mathscr{P}_0$ depends on both of these parameters, not just on a single parameter that varies gradually. This is a way in which the argument (A1,B1,C1) is different from a standard sorites argument.[16]

[15] This is not to deny that my height is vague. It is presumably vague which atoms are part of me, and therefore my spatial extent is vague. The claim is rather that whether or not I am tall supervenes on my height, and yet the vagueness of 'tall' is not simply a matter of the vagueness of height. We can assume that each person in the sequence has a height that is definitely within half a millimeter of the specified value.

[16] Despite this concession, I think that the parameter-dependence of the two sequences is more similar than it may appear. I will come back to this in section 7.

4 The Indeterminacy Thesis

The central claim of the indeterminacy thesis, as applied to the spectrum argument of section 2, is that (A2,B2,C2) is a sorites argument. The basic argument for the indeterminacy thesis is inference to the best explanation. If it walks like a duck, and it quacks like a duck, then it surely is a duck. So too, insofar as (A2,B2,C2)

(a) has the same *form* as a sorites argument, and

(b) is *paradoxical* in the same way as a sorites argument,

it surely *is* a sorites argument. Before explaining in more detail why the indeterminacy thesis is, at least, a *good* explanation, let me elaborate a bit on (a) and (b), and address some initial objections.

The basic sense of (a) is that (A2,B2,C2) has the same logical form as the standard sorites argument (A,B,C). In light of the discussion at the end of section 3, I can add that the cases considered in the argument represent gradual variations in a single precise parameter. That is not quite a matter of *logical* form, but it is part of the form more broadly. As for (b), the main thing I have in mind is that the premisses of the argument have a great deal of intuitive support, and the conclusion is extremely counter-intuitive. At least, many people have thought so, and in this paper I am simply taking such intuitions as read.

Still, it is natural to wonder whether there isn't some deeper sense in which (a) and (b) fail, or whether there isn't some disanalogy of another kind.

Let me mention a common initial thought along these lines. Many of my respondents have pointed out that to accept the conclusion of the tallness sorites would involve a fundamental conceptual confusion. One would have failed to understand what it *means* to be tall. In contrast, even the most hardened opponent of RC should not claim that those who accept it simply fail to understand the word 'better'. Now, I agree that this is a difference between the two cases, but the question is whether this difference undermines the indeterminacy thesis. Betterness is conceptually and epistemically subtler than tallness, but what does that show? The challenge for opponents of the indeterminacy thesis is to find a difference *and* explain why it matters. The minimum aim of this paper is to convince the reader that the challenge has not yet been met, and that, moreover, there are reasons to think that the challenge is a hard one.

Another initial objection is that the conditions (a) and (b) refer specifically to the argument (A2,B2,C2). So even if (A2,B2,C2) is a sorites argument, this does not show that the *original* spectrum argument (whether understood in the form (A1,B1,C1), or in some other way) is a sorites argument. I may have identified a sorites argument, but I have not said anything about the argument in which everyone else is interested. In responding to this objection, let me set aside the thought that considerations parallel to (a) and (b) apply to (A1,B1,C1), and perhaps to other arguments in the neighbourhood. After all, I conceded in section 3 that (A1,B1,C1) might in some ways be dissimilar to a standard sorites. My main response is, rather, that if (A2,B2,C2) fails because of vagueness, then, as a matter of logic, (A1,B1,C1), and all other arguments in the neighbourhood, *also* fail because of vagueness, whether or not they are sorites arguments in the strictest sense. I will develop this response in section 5, when I specialise to a particular logic of vagueness.

Let me now elaborate on why the indeterminacy thesis is a good explanation of (a) and (b), and hence a good working hypothesis. I claim it is both conservative and charitable in important ways.

First, it is theoretically conservative. It identifies the data given by the spectrum argument as an instance of a widespread phenomenon, something we were committed to explaining anyway. We do not need any new theoretical or conceptual resources; we have completely general reasons for thinking that betterness is vague and admits sorites sequences. And absent any presumption against vagueness, we might come to the indeterminacy thesis in the following way. Suppose that we deny RC. We then believe that some welfare levels are G, and that some welfare levels are not G; in particular, we accept (A2) and deny (C2). Moreover, we have quite general grounds for thinking that G, like most things outside fundamental physics, is vague. But, if in doubt, there is a standard way to check: look for a sorites sequence, a sequence $w_1, w_2, \dots, w_n$ of welfare levels that intuitively satisfy the inductive premiss (B2). This is, at least at a first pass, what it *means* for G to be vague.[17] Is there such a sequence? Yes: that is exactly what the spectrum argument shows. It shows that G is vague. Of course, there is something of a problem, because (A2) and (B2) entail (C2). But we know the name of this problem: it is the sorites paradox, and nothing else.

[17] See, for example, Bueno and Colyvan (2012) for a defence of the view that 'A predicate is vague just in case it can be employed to generate a sorites argument' (p. 29). Whether or not this is exactly right, the existence of a sorites series must be strong prima facie evidence of vagueness.

The second point is that the indeterminacy thesis promises to leave untouched most of our axiological thinking. Again, vagueness is widespread; we habitually work with vague concepts. What I take the sorites paradox to show, among other things, is that reasoning with vague predicates is a delicate matter. But it is not as if vague predicates suffer from irretrievable incoherence. We don't need to give them up altogether, and we couldn't if we wanted to. By the same token, according to the indeterminacy thesis, the spectrum arguments show that axiology is a delicate matter, but it is not rotten through and through. I note here that in this paper I am purely concerned with what is better than what, rather than with what one ought to do (even if I occasionally slip into the language of choice). So I will only incidentally consider an important question – what value vagueness means for moral choice.[18]

Third, the indeterminacy thesis is charitable to our intuitions. Theories of vagueness typically point to some defect in the inductive premiss (B). But, in doing so, they also aim to tell us what is right about it: why it seems compelling, and to what extent it might be reliable in ordinary circumstances. According to the indeterminacy thesis, the very same considerations will explain why the premiss (B) in the spectrum argument seems compelling and in some sense reliable.

I take these to be three theoretical virtues of the indeterminacy thesis. The third one is particularly significant: it means that it is hard to argue against the indeterminacy thesis based on intuitions, since the indeterminacy

[18] See Williams (2014a); Dunaway (2016); Moss (2016) for recent work on this subject. I plan to address this issue in future work.

thesis promises to explain away those intuitions. There is a final feature of the indeterminacy thesis which has a similar significance. The final feature is that there is very little consensus about what the correct theory of vagueness might be. Therefore the exact implications of the indeterminacy thesis are very hard to pin down, and therefore the indeterminacy thesis is very hard to refute. The situation is even more difficult because most treatments of vagueness stray to some extent, and in different ways, from classical logic. So not only is it hard to know what the implications are, it is hard to know what the right notion of 'implication' is! (I will consider some examples in the next section.) In addition, one need not take the view that vagueness is a monolithic phenomenon. One could defensibly claim that the vagueness of tallness and the vagueness of betterness arise in different ways, have different logics, or have different cognitive roles. So it is not enough to argue that the theory of vagueness that is appropriate to tallness is inappropriate to betterness.[19]

To be clear: this under-specification is not a theoretical virtue of the indeterminacy thesis. I would by no means put it forth as a reason for *endorsing* the thesis. It is, nonetheless, a reason why the indeterminacy thesis is an effective spoiler: the dialectical force of the spectrum arguments is diminished to the extent that the indeterminacy thesis *might* be true, and it is hard to rule it out.

[19] Embracing pluralism in this way does *not* undercut the first point above, insofar as we are already committed to giving a theory of the vagueness of betterness. It is not too surprising that normative vagueness might be sui generis, if one accepts that normative facts are sui generis. See Williams (2012) for a discussion of pluralism about indeterminacy, and especially Williams (2014b, §1) for the importance of distinguishing theories of indeterminacy by cognitive role.

Parity and Imprecise Equality

Let me illustrate these considerations by distinguishing the indeterminacy thesis from some others in the neighbourhood. Parfit (2016) has been developing a view which he calls 'the Imprecise Lexical View' which has at least a family resemblance to the kind of model I have in mind. The salient difference is that, for him, the central notion is not vagueness but what he calls 'imprecise equality'.The indeterminacy thesis claims that we cannot guarantee that $\mathscr{P}_{i+1}$ is better than $\mathscr{P}_i$; it may be indeterminate whether this is true. On Parfit's view, it may (instead) turn out that $\mathscr{P}_{i+1}$ is imprecisely equally as good as $\mathscr{P}_i$. Another potentially distinct view would appeal to 'parity' rather than vagueness: sometimes, it would claim, $\mathscr{P}_{i+1}$ is merely on a par with $\mathscr{P}_i$.

And so the question arises: supposing that the diagnosis I am pushing is generally right, in that the spectrum arguments point beyond a determinate and precise trichotomy of value relations. Is what's going on actually vagueness, or is it something else, like parity?[20]

First, some general comments. There is much disagreement in the literature about whether various cases are cases of parity or of vagueness. And, to my mind, the arguments are not very strong either way. For example, in the case I considered earlier of destroying an embryo versus cutting off my thumb, I am sympathetic to the idea that there is more than vagueness in play. Perhaps, for some broad range of cases, destroying the embryo is on par with, or imprecisely as bad as, cutting off my thumb. But, on the other

[20] I focus on parity because it is more widely studied than imprecise equality; as far as I can tell, though, the following comments could apply to either one.

hand, perhaps such cases are adequately analysed in terms of vagueness. The distinctions, are especially difficult to make if you think that there is such a thing as metaphysical vagueness. In Chang's work on parity (Chang, 2002), it seems that parity is supposed to be some deep structural feature of value, whereas vagueness is something lightweight and conventional and merely linguistic. If you think that there is metaphysical vagueness, then this kind of distinction can't really be the operative one.[21]

Nonetheless, let me reiterate why I think the indeterminacy thesis has the upper hand over the analogous parity thesis. The trump card of the indeterminacy thesis is that it promises an account of why the inductive premiss (B2) seems true. That is why the thesis is hard to dismiss; it explains away the countervailing intuitions. I will have some more to say about how this works in section 5, but the basic idea is that the indeterminacy thesis does not outright *deny* any instance of QC – it claims that some instances are borderline true. Parity-based views do not have this kind of structure. They outright deny some instances of QC, introducing an alternative hypothesis – namely, that a sufficiently large population $\mathscr{P}_{i+1}$ at level w_{i+1} would be on a par with $\mathscr{P}_i$. Moreover, cases of parity are fundamentally symmetrical. If $\mathscr{P}_{i+1}$ is on a par with $\mathscr{P}_i$, there is no reason why $\mathscr{P}_{i+1}$ should seem better than $\mathscr{P}_i$, rather than vice versa; there is no reason why the inductive premiss (B2) would appear true instead of false. This does not accurately reflect the

[21] This is one way in which axiological vagueness might differ from the vagueness of tallness (cf. footnote 19). Despite recent work on the subject (see Barnes and Williams (2011); Wasserman (2015); Wilson (2016) for three very different approaches), there is widespread skepticism about metaphysical vagueness. Schoenfield (2015) argues that moral vagueness *must* be metaphysical vagueness, given moral realism. In general, it seems clear that questions about axiological vagueness lead quickly into metaethics (Schiffer, 2002; Constantinescu, 2014; Dougherty, 2014).

psychology of the spectrum arguments.

The first theoretical virtue of the indeterminacy thesis is also relevant here. Not everyone accepts the possibility of parity, but everyone should accept the possibility of vagueness. And even if we have independent reasons to accept the possibility of parity, it is not clear why it should play a role here. Vagueness can do the work, just as it does in every other sorites argument.

5 Classical Versions of the Indeterminacy Thesis

The basic stance of the indeterminacy thesis is: tell me what to say about sorites arguments, and I'll tell you what to say about spectrum arguments. It is not my business here to decide what one *should*, in fact, say about sorites arguments. However, in this section, I will briefly sketch some typical approaches to the sorites. The purpose is to give at least a sense of what the final picture might look like, and to ground further discussion of the thesis.

Classical versus non-classical

Theories of vagueness can, in the first instance, be divided into those that respect classical logic, and those that do not. In the first camp I include epistemicism and supervaluationism, while truth-functional degree theories fall in the second.[22] 'Respecting' classical logic means, for one thing, that

[22] It is a vexed question what 'the' logic of supervaluationism is; see Varzi (2007) for an overview. In this paper, I adopt a 'local' perspective, according to which the truth predicate is disquotational. The logic of truth preservation is then a classical modal logic (KT or stronger, often assumed to be S5) with respect to the operator 'It is definitely true that' (or 'Definitely'

classically valid inferences preserve definite truth.[23] In particular, the theorems of classical predicate logic like excluded middle come out definitely true. Views in this first group must accept the existence of cutoffs:

Tallness Cutoffs

There exists a person p_i such that p_i is tall and p_{i+1} is not.

This is the most obvious cost of these classical theories; I will say more about it soon. In contrast, non-classical theories may deny, or anyway not entail Tallness Cutoffs. The ways to do this are potentially diverse, but, as Bacon (2015, §1.1.1) has argued, one typically has to give up a wide swathe of classical logic including (for example) the axiom of conjuctive syllogism:[24]

(CS) $(P \rightarrow Q) \wedge (Q \rightarrow R) \rightarrow (P \rightarrow R)$.

Now, of course, departing from classical logic to this extent is a cost to the theory. But, as far as the indeterminacy thesis is concerned, it is a *sunk* cost. Giving up such classical theorems as (CS) is already justified (if at all) by standard cases of vagueness. It is not something the indeterminacy thesis needs to justify anew. To put it another way: if one finds such revisions

for short). The more traditional 'global' approach to supervaluationism (see e.g. Keefe (2000)), identifies truth with supertruth. The truth predicate is then not disquotational, and the obvious notion of validity invalidates classical metarules like conditional proof. But the difference between these approaches is not very important for me here, and everything I say could be rephrased in 'global' terms.

[23] So e.g. we have modal axiom K. I tend to speak of 'definite' truth rather than 'determinate' truth, because it is less of a mouthful. As usual, a sentence or proposition is *indeterminate* or *borderline* if it is neither definitely true nor definitely false. Also, see Smith (2013, especially §5.2) for an account of vagueness that claims to preserve classical validity without affirming classical tautologies as definitely true.

[24] Example: in traditional fuzzy logic, suppose that P, Q, R have degrees of truth $1, 1/2, 0$ respectively. Then (CS) has degree of truth $1/2$. (This works whether we define '→' in terms of negation and disjunction, or use the Łukasiewicz conditional.)

of logic objectionable, that is a reason to reject non-classical theories of vagueness; it is not a reason to reject the indeterminacy thesis.

In most of what follows I will concentrate on classical theories of vagueness. This is, in part, because I find them more plausible, and in part because they are more tractable: it is fairly easy to say what the indeterminacy thesis looks like, and to do so in fairly specific terms, even if there is some variety among classical views (e.g. epistemicism versus supervaluationism). There is also the sociological fact that supervaluationism is a common default theory, even if it is not universally endorsed.

A final reason for focusing on classical theories of vagueness is that they make certain debates much clearer. I have mentioned that one debate surrounds the role of parity; another surrounds transitivity. But it is not entirely clear what these debates are about, outside of a classical framework. On classical views, it is obvious what it means to affirm trichotomy (and hence exclude parity): given populations $\mathscr{P}$ and $\mathscr{Q}$,

> **Trichotomy**
>
> $\mathscr{P}$ is better than $\mathscr{Q}$, or $\mathscr{P}$ is worse than $\mathscr{Q}$, or $\mathscr{P}$ is exactly as good as $\mathscr{Q}$.

Classical views can endorse this disjunction while allowing that there may be cases in which none of the disjuncts definitely applies. They can thus affirm trichotomy while allowing that betterness is vague. In contrast, non-classical theories typically do not affirm the disjunction unless one of the disjuncts is definitely true. They reject trichotomy in roughly the same way they reject bivalence. The best they can ensure is that the disjunction is not definitely false.

Similar considerations apply to transitivity. Classical theories can affirm the conditional:

> **Transitivity**
>
> If $\mathscr{P}$ is better than $\mathscr{Q}$ and $\mathscr{Q}$ is better than $\mathscr{R}$, then $\mathscr{P}$ is better than $\mathscr{R}$.

In non-classical theories of vagueness, the conditional may often fail to be definitely true, so it is unclear how such a theory can affirm transitivity. At least heuristically, transitivity fails in the same way that (CS) fails.

Applications to the Spectrum Argument

Now let me discuss classical versions of the indeterminacy thesis in more detail. As I have explained, the upfront cost of classical theories is that they endorse Tallness Cutoffs and its analogues. The corresponding version of the indeterminacy thesis transposes this to the setting of the spectrum argument, endorsing

> ***G*-ness Cutoffs**
>
> There exists a welfare level w_i such that w_i is G and w_{i+1} is not.

This is counterintuitive. However, Tallness Cuttoffs is also counterintuitive. Unless we can point to some relevant difference between the two cases, the countering intuitions are objections to classical views of vagueness, not objections to the indeterminacy thesis per se.

Notice that Tallness Cutoffs is just the claim that the inductive premiss (B) has some false instance. So the task of explaining why (B) seems compelling is closely related to the task of explaining why Tallness Cutoffs seems false. The claim of the indeterminacy thesis is that the same explanations work for (B2) and G-ness Cutoffs respectively. There are two types of considerations usually adduced.

The scope error

The first type of consideration is that (B) has no definitely false instances. This lulls us into thinking that it has no false instances.[25] To put it another way, Tallness Cutoffs, though true, is easily confounded with a claim which is similar but false:

> **Narrow-Scope Tallness Cutoffs**
>
> There exists a person p_i such that, *definitely*, p_i is tall and p_{i+1} is not.

This amounts to a scope error for the operator 'definitely': definitely, Tallness Cutoffs is true, but we think it is false because we mistakenly move 'definitely' within the scope of the existential quantifier.

What would be really counterintuitive, by these lights, would be to have an *assertable* (or knowable, or distinctly conceivable) instance of 'p_i is tall and p_{i+1} is not'. But, typically, the assertability of S tracks the definite truth of S. Since this sentence schema has no definitely true instances, it has no

[25] See for example Keefe (2000, pp. 185-6).

assertable instances.[26] It is, moreover, plausible that all of the instances are *far* from definitely true, and thus far from assertable; I will take this up below.

Now, if this is right, we can certainly use the same maneuver in the case of the spectrum argument. The claim will be that the inductive premiss (B2) has no definitely false instances, even though some of its instances are borderline. Insofar as G-ness Cutoffs seems false, it is because we conflate it with something else: the claim that there exists a *determinate* cutoff.[27]

Here is a possible objection. We do not reject G-ness Cutoffs because we conflate it with something else; we reject it because (B2) follows from Transitivity and QC. To answer this objection, we can just change the level of explanation. Assuming that Transitivity is definitely true, (B2) is logically equivalent to QC. Why do we accept QC? All the same considerations can apply. QC has no definitely false instances, and this lulls us into thinking it has no false instances. Or, if you prefer, the negation of QC has no definitely true instances, hence no assertable instances. We are mistaken about the status of QC in a way characteristic of vagueness – and *as a result* we are mistaken about the status of (B2), in the very same characteristic way.

Degrees of Truth

There is a second strategy, complementary to the first. (A version of this strategy is also typically evoked by non-classical theories of vagueness.) The

[26] At least at a first pass, supervaluationists and epistemicists take the same line here; epistemicists, of course, emphasise an epistemic reading of 'definitely'.

[27] 'There exists a determinate cutoff' is just shorthand for 'There is a welfare level w_i such that, definitely, w_i is G and w_{i+1} is not'. Thus Narrow-Scope Tallness Cutoffs claims that there exists a determinate cutoff for tallness.

idea is that there are gradations between definite truth and definite falsehood; I will call the levels 'degrees of truth', although, as Dorothy Edgington (1996) observed, 'degrees of closeness to definite truth' might be more apt, since the theories I am considering affirm bivalence. Thus if x is definitely not tall, then 'x is tall' has the lowest possible degree of truth (let us call it 0); this should increase continuously with the height of x, attaining its maximum (let us call it 1) when x is definitely tall.

One can always make sense of something akin to degrees of truth. One can rank propositions by the relation of definite material implication. That is, T is at least as highly ranked as S if, definitely, S implies T. In supervaluationist terms, T is true on every precisification on which S is true. This relation is a preorder, i.e. it is reflexive and transitive. A theory of degrees of truth might refine this logically defined ranking in some way (but everything I say can be understood in terms of the logical ranking). For example, supposing that there is a probability measure on the set of precisifications, the degree of truth of a sentence may be the measure of the set of precisfications on which it is true.[28] In any case, all definitely true propositions have the same (highest) rank, and all definitely false propositions have the same (lowest) rank. In the tallness sorites (and similarly in every standard example), the rank of 'p_i is tall' decreases steadily as p_i becomes shorter; it is true on a steadily shrinking set of precisifications. Moreover, each instance of the inductive premiss (B) is *almost definitely true* – that is, true to a high degree – since there are qualitatively few precisifications that make the antecedent

[28] This kind of view is explored, for example, by Kamp (1975), Lewis (1980), Edgington (1996), and Williams (2014a). The following discussion is especially informed by Edgington's work.

true and the consequent false.

In application to the sorites, the key idea is that a proposition that is almost definitely true is, as it were, almost as good as one that is definitely true. Cashing out 'almost as good' is clearly a central problem for such a view. As far as the sorites goes, the natural idea is that a proposition that is almost definitely true, like the inductive premiss (B), is liable to seem true. Similarly, a proposition that is almost definitely false (like the claim in a particular case that p_i is tall and p_{i+1} is not) is liable to seem false, being so far from definitely true.

Insofar as this strategy works, it can be applied to the spectrum argument. The indeterminacy thesis can hold that all the instances of (B2) are, if not definitely true, then almost definitely true, and so, as far as that goes, liable to seem true. It is not hard to believe that our moral intuitions might have trouble discerning cases of definite truth from cases of almost definite truth.[29] The indeterminacy thesis can give a similar account of QC, since QC and (B2) must have the same degree of truth.[30] Thus every instance of QC will be almost definitely true, and liable to seem true.

As Keefe (2015) has observed, it is not entirely clear that this strategy

[29] It may help to keep in mind here a point I will develop more fully in section 7. The degree to which $\mathscr{P}_{i+1}$ is better than $\mathscr{P}_i$, in the ordinary sense of the phrase, is conceptually distinct from the degree to which it is true that $\mathscr{P}_{i+1}$ is better than $\mathscr{P}_i$. One has to carefully distinguish the thought that $\mathscr{P}_{i+1}$ is *much* better than $\mathscr{P}_i$ from the more technical thought that $\mathscr{P}_{i+1}$ is *definitely* better than $\mathscr{P}_i$. In particular, the intuition that $\mathscr{P}_{i+1}$ is *much* better than $\mathscr{P}_i$ does not hinder the strategy under discussion, as long as it is almost definitely true that $\mathscr{P}_{i+1}$ is much better than $\mathscr{P}_i$. Thanks to Gustaf Arrhenius for pressing me on this point.

[30] I assume here that Transitivity is definitely true, and that classically valid inferences preserve degrees of truth. More precisely, assume that if p and q jointly entail r, and q is definitely true, then p and q have the same degree of truth. This works for the logically and probabilistically defined degrees of truth mentioned above.

does work, or how much it adds. After all, Tallness Cutoffs is definitely true. So the degree-theoretic considerations I have sketched so far cannot by themselves explain why Tallness Cutoffs seems false. We have to appeal to some version of the earlier strategy: Tallness Cutoffs seems false because the existential quantification is not made true by any determinate instance. We can now add that each instance is very far from being definitely true, and is liable to seem false.

Thus degrees of truth may play some role in explaining why the inductive premisses are plausible. More interesting, I think, is their potential role in explaining why the inductive premisses are *reliable*. Let me explain the relevant sense of reliability. A natural thought is that a well-informed, rational person will believe S if and only if S is definitely true. But when S is merely borderline, well-informed, rational people will still have some sort of quasi-doxastic attitude, short of belief, towards S. This quasi-doxastic attitude should play a role in determining behaviour, contributing towards the kind of hedging that people display about borderline cases. (It may not completely explain this hedging, since, in the first place, people are actually badly informed and irrational, and, in the second place, pragmatic and other factors will usually be in play.) The idea is that the appropriate attitude towards S is a function of the degree of truth of S, and, moreover, a continuous function. If S is close enough to definite truth, the appropriate attitude towards S will be close to full belief.

The simplest version of this picture is to understand the degree of truth of S as the ideal credence in S, conditional on the precise facts. Thus degrees of truth are analogous to objective chances, considered as ideal credences conditional on 'admissible' information. If S has degree of truth 0.95, then it would be a mistake to rely on S in roughly the same way that it would be a mistake to rely on a 20-sided die not landing with 20 facing up. Credences plug into a broadly Bayesian framework, and into a theory of rational choice, in the usual kind of way.[31]

There are potential alternatives to this credential view of degrees of truth. The degree could correspond to a kind of hedging different from that associated with partial belief. For example, Williams (2014a) promotes a different view, on which the degree of truth is, roughly, the propensity for the agent to act as if S is true. So, for example, if the sentence 'John is tall' gets weight 2/3, then, if you are forced to classify John as tall or not tall, you will plump for 'tall' two thirds of the time. This suggests a different way in which S's having a high degree of truth leads rational, well-informed people to act as if S is true.

One sense, then, in which propositions with a high degree of truth might be 'reliable' is that they might warrant attitudes and behaviour similar to those warranted by propositions that are definitely true. Another, related sense is that they might be *inferentially* reliable.[32] For example, the conclusion of a classically valid single-premiss deduction is at least as true as the premiss; multi-premiss deductions preserve degrees of truth in a more complicated

[31] This kind of view is advocated by Bacon (2015), Dunaway (2016), and (as far as I understand the end-point) Moss (2016). (A natural thought is that, on these ideal-credence views, expected utility theory will effectively resolve any axiological vagueness, but the situation is not straightforward. I will take up this issue in other work.) Similar moves can be made by non-classical theories, with the caveat that the degrees of belief in question must not satisfy the probability calculus. See Field (2000); Schiffer (2000); Smith (2013). Bacon (2015) contains an especially helpful survey of such alternatives.

[32] This is especially emphasised by Edgington (1996). Williams (2011) discusses the logic of these degreed theories starting from the thought that logic norms belief.

sense. If a deduction has few premisses, and they, being close to definitely true, warrant something close to full belief, then the conclusion will also be close to definitely true, and warrant something close to full belief. Serious problems only arise when an inference involves *many* premisses that are close to definitely true. The conclusion of such a inference may be definitely false. That is what happens in sorites arguments, and, on the indeterminacy thesis, that is what happens in the spectrum argument.

According to this sort of theory, the *right* attitude towards the premisses (B2) and QC is something close to full belief. Moreover, these premisses are inferentially reliable in the sense just explained. This illustrates the sense in which the indeterminacy thesis is conservative: it holds that our intuition in favour of QC does not lead us far astray.

Summing up

Let me summarise what the indeterminacy thesis will look like, given a classical theory of vagueness. In doing so, I will address the worry raised in section 4 that the indeterminacy thesis might be right about some forms of the spectrum argument but not about others.

In this section I have mainly focused on the argument (A2,B2,C2). That is the version of the spectrum argument of most direct concern to the indeterminacy thesis, since it is the version most clearly resembling a sorites. As in any other sorites, the first premiss will be true and the conclusion false. Since the argument is clasically valid, the inductive premiss, (B2), must have at least one false instance. But it need not have any definitely false instances. Every instance may be very close to definitely true.

The inductive premiss (B2) is based originally on the principle QC and Transitivity. Let us assume that transitivity is definitely true. Then an instance of (B2) holds if and only if the corresponding instance of QC holds. If one is indeterminate, then so is the other. More precisely, they must have the same degree of truth. Some instance of QC is false, but every instance is almost definitely true. QC, like (B2), is liable to seem true.

Now let us turn to the argument (A1,B1,C1), and indeed to my original sketch of the spectrum argument. Here we use QC to construct, given $\mathscr{P}_i$, a better population $\mathscr{P}_{i+1}$. Since some instance of QC is false, this construction might fail. On the other hand, since every instance of QC is almost definitely true, the construction almost definitely succeeds. The upshot is that we can always find $\mathscr{P}_{i+1}$ that is almost definitely better than $\mathscr{P}_i$.[33] By the same token, we can guarantee that the inductive premiss (B1) is, in every instance, almost definitely true. The indeterminacy view, applied initially to (A2,B2,C2), thus leads to a vagueness-based account of why other forms of the spectrum argument appear paradoxical, whether or not they are sorites arguments in any strict sense.

[33] More carefully: let S_N be the sentence 'A population of N people at level w_{i+1} would be better than $\mathscr{P}_i$.' Make the plausible assumption that if S_N is true, then so is S_{N+1}. This means that the degree of truth of S_N is monotonic in N, and that the degree of truth of QC is the supremum. (Supervaluationistically, QC is true on exactly those precisifications on which at least one S_N is true.) We can thus find N such that the degree of truth of S_N is as close as we like to the degree of truth of QC. This is presumably enough to ensure that some S_N counts as almost definitely true.

6 The Toy Model

I have explained what the indeterminacy thesis claims about the premisses of the spectrum argument, assuming a classical theory of vagueness. But can these claims be made good? Perhaps there is no reasonable population axiology on which the inductive premiss is indeterminate. Or, even if there is, the indeterminacy might not work in the right way to block the repugnant conclusion and to see off various objections. I cannot dissolve the worry completely, if that would involve presenting an axiology entirely beyond reproach. What I can do is present, as a starting point, a toy model in which everything works as advertised, and which is at least *prima facie* plausible. Since such a model exists, there cannot be a simple formal objection to the indeterminacy thesis. A bit more generally: even if the toy model turns out to be quite wrong as a model of *betterness*, it does seem to be a pretty reasonable model of *a* vague but transitive binary relation. And the vagueness of this relation leads to paradoxical arguments of both types (A1,B1,C1) and (A2,B2,C2). It seems that vagueness *can* do the explanatory work that we require.

First a qualitative description.[34] The model will be sufficientarian in flavour. (I will discuss the interpretation more carefully in section 8.) It claims an important distinction between lives that are 'satisfactory', or above a level of sufficiency, and those that are not. The level of sufficiency is somewhere

[34] The model I am describing is similar to views proposed by Qizilbash (2005) and by Knapp (2007). I do not think they had any very specific models in mind, but what follows is a natural way of spelling out their ideas, given a classical theory of vagueness and certain general assumptions (e.g. separability and trichotomy).

between the levels of drab and blissful lives. In the drab population $\mathscr{P}_n$, there are no satisfactory lives, while in $\mathscr{P}_0$ there are many; that will be enough to make the latter better than the former. As far as the spectrum argument goes, the principle QC fails if the level of sufficiency lies between w_i and w_{i+1}, since at that point $\mathscr{P}_{i+1}$ has no satisfactory lives, while $\mathscr{P}_i$ has many. That is therefore the point at which the inductive premisses (B1) and (B2) fail. This is the distinctively sufficientarian way to avoid RC, and to render the spectrum argument unsound.

The role of vagueness is just to make the sufficientarianism easier to stomach. The most obvious objection to sufficientarianism is that a small change in welfare from above to below the level of sufficiency cannot plausibly make such a big difference. This objection loses some of its force when we realise that 'satisfactory' is vague, and that there is a similar problem for all vague predicates. (Indeed, if we were to adopt a theory of vagueness that rejected Tallness Cutoffs, we could analogously reject the existence of a cut-off for 'satisfactory'; it would be misguided to talk about the level of sufficiency at all.)

Now to the details. I will suppose that each life has a welfare level given by a real number. For simplicity I will focus only on positive levels, although it is easy enough to extend the model symmetrically to include negative ones. The level of sufficiency will be some real number α. Let us suppose it is between 10 and 100, but otherwise indeterminate. So welfare levels equal to α or above are 'satisfactory', and those below are not.

The overall value of a population is given by the number of satisfactory lives it contains, with ties broken by 'total welfare' (i.e. the sum of the real numbers representing welfare levels).[35]

Let us suppose that the blissful lives in population $\mathscr{P}_0$ are at welfare level 101, while the drab lives in $\mathscr{P}_n$ are at level 1. Then the blissful lives are definitely satisfactory, and the drab lives are definitely not satisfactory. It is obvious from what I have said that no population at level 1 can be better than $\mathscr{P}_0$. The Repugnant Conclusion is definitely false.

More generally, consider the spectrum argument. QC, as I earlier suggested, fails just when the level of sufficiency α lies between w_i and w_{i+1}. Suppose we take $w_1 = 100, w_2 = 99, w_3 = 98$, and so on. Then it is never definitely true that α lies between w_i and w_{i+1}, so it is never definitely true that QC fails. This means that the inductive premisses (B1) and (B2) are never definitely false.

We can supplement this with a story about degrees of truth. The simplest thing to say is that the degree of truth of a sentence S is the proportion of candidates for α that would make S true. For example, suppose S is the sentence 'Welfare level w is satisfactory.' There are three cases. If w is at least 100, then S is true regardless of where the threshold α lies: it is true to degree 1. If w is below 10, then S is false regardless of where the threshold lies: it is true to degree 0. In other cases, S is indeterminate: it has degree of truth $(w - 10)/90$, since S would be true if and only if the threshold were between 10 and w.

With this supplement, we see that QC and the inductive premiss (B2)

[35] The point of breaking ties in this way is that losses to one person can be compensated by equal gains to another, as long as neither or both of them crosses the level of sufficiency. More nuanced tie-breaking rules are possible, but this will do for my current purposes.

are almost definitely true. At any rate, their degrees of truth are at least 89/90. This is because they are false only if α is between w_i and w_{i+1}; they are false for at most 1/90 of the candidates for α. As for (B1), suppose that we make $\mathscr{P}_{i+1}$ twice as large as $\mathscr{P}_i$. Then, again, (B1) is false only if α is between w_i and w_{i+1}. So (B1) is almost definitely true.

7 Structural Objections

As I mentioned in the beginning, a lot of people have noticed something like the indeterminacy thesis, but then they have set it aside very quickly, without sustained argument. The typical thought is that there is an obvious structural disanalogy between the spectrum argument and sorites arguments. For example, Derek Parfit writes

> It may be objected that my argument is like what are called *Sorites Arguments*, which are known to lead to false conclusions.…A Sorites Argument appeals to a series of steps, each of which is assumed to *make no difference*. My argument would be like this if it claimed that [$\mathscr{P}_1$] is *not worse* than [$\mathscr{P}_0$], [$\mathscr{P}_2$] is not worse than [$\mathscr{P}_1$], [$\mathscr{P}_3$] is not worse than [$\mathscr{P}_2$], and so on. But the argument claims that [$\mathscr{P}_1$] is better than [$\mathscr{P}_0$], [$\mathscr{P}_2$] is better than [$\mathscr{P}_1$], [$\mathscr{P}_3$] is better than [$\mathscr{P}_2$], and so on. The objections to Sorites Arguments are therefore irrelevant.[36]

[36] Parfit (2004, fn. 13). He is really discussing not the spectrum argument I have presented, but his 'second paradox', which is closely related to the form of spectrum argument I will consider in section 8.

My response to this is, again, to call attention to the argument (A2,B2,C2). This argument appeals to a series of steps, each of which is assumed to make no difference to G-ness. The objections to Sorites Arguments are therefore potentially relevant. But if they are directly relevant to (A2,B2,C2), then they are indirectly relevant to the spectrum argument in any form, in the way I have discussed in section 5 and modelled concretely in section 6.

The subtlest and most interesting objection of this structural kind is the one developed in chapter 9 of Temkin's *Rethinking the Good*. To be sure, my basic response to his objection is the one just given: (A2,B2,C2) has *exactly* the structure of a sorites argument. Still, what Temkin says is initially persuasive, and deserves a more direct response, which will in any case illuminate some features of the view on offer.

Temkin focuses on the first form of the argument, (A1,B1,C1), in which the different cases are different populations, and the predicate is 'better than $\mathscr{P}_0$'. (In fact, he is writing about the pain spectrum I mentioned in section 2, but the considerations are exactly parallel.) Here is the disanalogy Temkin sees between the tallness sorites and the spectrum argument, in my reconstruction.

> **Tallness Disanalogy**
>
> The standard sorites sequence moves from a tall case through *less and less tall* cases to a non-tall case. It might be a bit mysterious where and how exactly the transition occurs from tall to not tall; but at least we are moving in the right direction. In the spectrum argument, however, we move from a better-than-$\mathscr{P}_0$

> case through *better and better* cases to a not-better-than-$\mathscr{P}_0$ case. This would be like moving from a tall case through *taller and taller* cases to a non-tall case. We are going in completely the wrong direction. (If we *could* go through taller and taller cases to a non-tall, hence overall *less* tall case, then 'taller than' would not be transitive. By the same token, we should conclude that 'better than' is intransitive.)

The first and main point is that Temkin's objection clearly does not apply to the second form of the spectrum argument, (A2,B2,C2). In that argument, the precise parameter on which G-ness depends – the analogue of height – is the welfare level. Intuitively, high welfare levels are G, and low welfare levels are not G. There can be no doubt that this form of the spectrum argument moves in the right direction, from high welfare levels (ones that are G) through lower welfare levels (ones that are less plausibly G), to low welfare levels (ones that are definitely not G). So Temkin has not identified a disanalogy between (A2,B2,C2) and the tallness sorites. So it is hard to see how the considerations he adduces can tell against the indeterminacy thesis.

The second point is that the indeterminacy thesis easily rebuffs the argument for intransitivity implicit in the Tallness Disanalogy. That argument is based on the claim that $\mathscr{P}_{i+1}$ is better than $\mathscr{P}_i$; the cases get 'better and better'. However, remember, a central claim of the indeterminacy thesis is that *it is sometimes indeterminate* whether $\mathscr{P}_{i+1}$ is better than $\mathscr{P}_i$. As long as we accept this indeterminacy, there is no argument in the offing that 'better than' is intransitive. Temkin can insist on the intuition that $\mathscr{P}_{i+1}$ is better than $\mathscr{P}_i$; but that is precisely the kind of intuition that the indeterminacy thesis promises to explain away.

Despite all this, the Tallness Disanalogy is a source of legitimate puzzlement. Each person in the tallness sorites is (definitely!) shorter than the one before. Shouldn't the indeterminacy thesis hold that (A1,B1,C1) is *just like that*, with each population worse than the one before? If not, how can we possibly get from definitely better than $\mathscr{P}_0$ to definitely worse than $\mathscr{P}_0$? Consider, after all, the normative sorites I discussed in section 3: the act of destroying the embryo gets worse (or, anyway, no better) from second to second until in the end it is definitely worse than cutting off my thumb. But the indeterminacy thesis cannot say that each population is no better than the one before: that would require definitely false instances of QC, which the thesis is designed to avoid. *Isn't* this an important disanalogy, an important sense in which the waterfowl fails to quack?

Now, I've already conceded that (A1,B1,C1) may be a little different from a standard sorites, and shown that, nonetheless, a paradoxical argument of this type can arise from a vague binary relation – but I also admit that it is initially hard to see what is going on with this argument, and to see in a detailed way *why* the Tallness Disanology is not an important one. It would be valuable to put things in a clearer light.

The key, I suggest, is to distinguish two senses in which the sequence of people in the tallness sorites moves 'in the right direction' from people who are definitely tall to people who are definitely not tall. One sense, emphasised by the Tallness Disanalogy, is that the people get shorter and shorter. But remember that we *also* have a natural gradation between definite truth and definite falsity, given by the relation of definite material implication. Edgington's locution of *closeness* to definite truth serves well here: some people are closer than others to being definitely tall.[37] Two things are happening in the sorites sequence: the people gradually get shorter, but they also gradually get further from being definitely tall.

Now, when it comes to understanding how classical theories of vagueness treat the tallness sorites, I claim that it is really the *latter* fact that does all the work. How so? First of all, we are trying to get from a person who is definitely tall – tall on every precisification – to a person who is definitely not tall – tall on no precisification.[38] So insofar as one is worried about 'moving in the right direction', what seems to be crucial is that subsequent people are tall on *fewer* precisifications. But that is just another way of saying that they are further from being definitely tall. Second, the characteristic claim that the inductive premiss (B) is not always definitely true is exactly the claim that p_{i+1} may be strictly further than p_i from being definitely tall. Finally, the thought that each instance of the inductive premiss is 'close' to being definitely true is just the thought that p_{i+1} is 'not much' further than p_i from being definitely tall. That is: p_{i+1} is tall on fewer precisifications, but not 'many' fewer. In this sense, p_i and p_{i+1} are similar with regards to tallness. Of course, tallness should supervene on height, but the diagnosis of the sorites argument is all about how the cases in the sequence descend the

[37] Explicitly, when I say that X is closer than Y to being definitely tall, I mean: it is definitely true that if Y is tall then X is tall, and (in order for 'closer' to be a strict comparative) it is not definitely true that if X is tall then Y is tall. As I discussed in section 5, this logically-defined preorder could be supplemented by a more refined theory of degrees.

[38] I use the language of precisifications as a supervaluationist would (and I always mean *admissible* precisifications). But at least at a first pass, any clasical theory of vagueness can make use this idea: precisifications are ways of resolving vagueness that are not definitely incorrect. (For subtleties surrounding this move, see Bacon (2015, ch. 3).)

gradations between definite truth and definite falsity.

It is easy to overlook the point just made because 'shorter' and 'further from definitely tall' *almost* coincide. Consider Adric, who is 200cm in height; he is definitely tall. Turlough, who is 199cm in height, is definitely tall. That is an example in which Turlough is shorter than Adric, but Turlough is no further than Adric from being definitely tall: they are both definitely tall, full stop. On the other hand, the two comparatives coincide in the loose sense that 'further from definitely tall' implies 'shorter', and 'shorter' implies 'not closer to being definitely tall'. Because of this, it is not strictly speaking true that each person in the sorites sequence is further from definitely tall than the one before. Strictly speaking, the first few people are definitely tall, so equally close to being definitely tall. After some indeterminate point, the people become further and further from being definitely tall. Then the last few people are equally far from being definitely tall, since they are all definitely *not* tall. Still, the two comparatives loosely coincide, and it also seems acceptable to say, in a loose phrase, that the people gradually get further from being definitely tall.

However – and here is the key point – 'worse' and 'further from being definitely better-than-$\mathcal{P}_0$' *need not coincide*, even in the loose sense just explained.[39] For suppose that it is indeterminate whether $\mathcal{Q}$ is better than $\mathcal{P}_0$. And suppose that $\mathcal{R}$ contains the same people as $\mathcal{P}_0$, but everyone is very slightly better off. Then $\mathcal{R}$ is definitely better than $\mathcal{P}_0$. But because $\mathcal{R}$ and $\mathcal{P}_0$ are very similar, it may still be indeterminate whether $\mathcal{Q}$ is better

[39] Note the hyphenation to help parse some unwieldy phrases: we are considering the binary relation *X is further than Y from being definitely better-than-*$\mathcal{P}_0$.

or worse than $\mathscr{R}$. In such a scenario, $\mathscr{Q}$ is further than $\mathscr{R}$ from being definitely better-than-$\mathscr{P}_0$, but it is indeterminate whether $\mathscr{Q}$ is worse than $\mathscr{R}$. This is not, of course, to say that the two comparatives are completely unrelated. Consonant with transitivity, we can expect that 'further from definitely better-than-$\mathscr{P}_0$' implies 'not definitely better'. But, the example shows, it does not imply 'worse'.

Note that this line of thought does not rely on the indeterminacy thesis as a view about spectrum arguments. It just relies on the general hypothesis that 'better than $\mathscr{P}_0$' could be vague. Here is a concrete and familiar example from outside the setting of population ethics. Suppose that P_0 is the life of a rock star, R is the life of a slightly happier rock star, and Q is the life of a philosopher. Then (in some such case) Q is further than R from being definitely better-than-P_0, but it is indeterminate whether Q is worse than R. And in this respect, there is nothing special about 'better than'. We could instead consider the predicate 'redder than'. Suppose that P_0 is an orange-ish patch of colour, R is a slightly redder orange-ish patch of colour, and Q is a reddish-purple patch of colour. Then R may be closer than Q to being definitely redder-than-P_0, even though it is indeterminate whether R is redder than Q.

Where does this leave us? Let me return to the spectrum argument. We cannot say that the populations are gradually getting worse – that is the point made by the Tallness Disanalogy. But the argument (A1,B1,C1) can *still* be analogous to the tallness sorites in the sense that the populations gradually get further from being definitely F, i.e. further from being definitely better-than-$\mathscr{P}_0$.[40] If what I've said is right, *that* analogy is sufficient for vagueness to play the same explanatory role in (A1,B1,C1) as it does in the tallness sorites argument. It provides the important sense in which the cases gradually change from definite F-ness to definite not-F-ness. It is not directly relevant whether the cases are gradually getting worse. Getting worse is only *one* way to get further from definite F-ness. This is my detailed explanation of why it is that Tallness Disanology does not undermine the indeterminacy thesis.

8 Elitism

I will now consider a different spectrum argument from the one given section 2. Instead of QC, it is based on a principle of 'non-elitism'.[41] The indeterminacy thesis in this setting will be the same, in outline, as before. However, the new spectrum argument is more worrying for population ethics in general, just because the non-elitism principle is *prima facie* more compelling than QC: no one wants to be called an elitist, or even a borderline elitist. Thus it is worth looking at the argument from non-elitism in more detail.

Consider a sequence of populations $\mathscr{Q}_0$, $\mathscr{Q}_1$, …, $\mathscr{Q}_n$ constructed in the following way (Figure 2). Let z to be a welfare level below w_n, but still positive. The inital population $\mathscr{Q}_0$ is just $\mathscr{P}_0$ with the addition of a vast number of lives at level z. (The number needs to be large enough for the

[40] This is indeed what happens in the toy model. If $\mathscr{P}_{i+1}$ is always (say) twice as large as $\mathscr{P}_i$, the degree to which it is true that $\mathscr{P}_i$ is better than $\mathscr{P}_0$ is just the degree to which it is true that w_i is satisfactory. This gradually decreases as w_i decreases from 100 down to 10.

[41] One can use the principles discussed below to give an argument for QC, but the dialectic is a bit easier to understand if we use them to argue directly for RC.

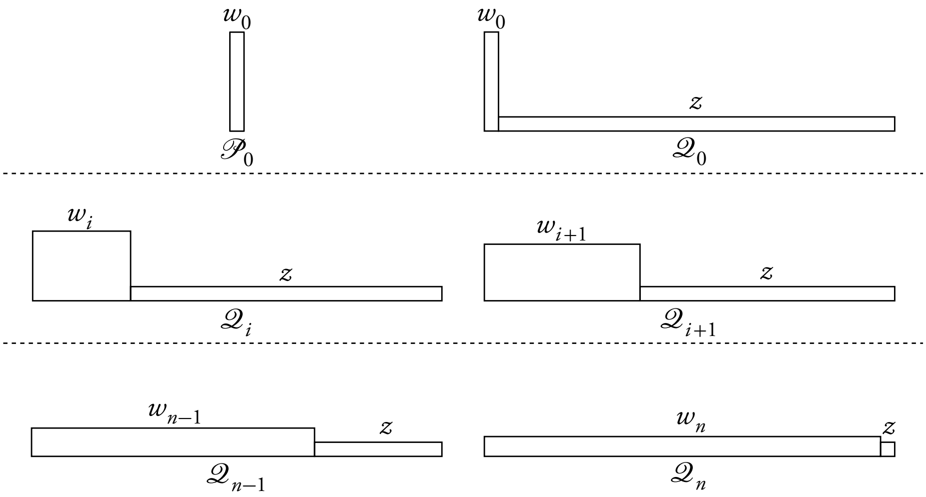

Figure 2: The populations $\mathscr{Q}_i$, with some lives at level w_i and some at z.

subsequent argument to go through.) Given $\mathscr{Q}_i$, we construct $\mathscr{Q}_{i+1}$ by slightly lowering the welfare level of the best-off people from w_i to w_{i+1}, and raising the welfare level of some of the worst-off people from z to w_{i+1}. The non-elitism principle claims that this will result in a net improvement, as long as enough people benefit. (What counts as 'enough' should not depend on the number of lives initially at level z.) Here is a formulation parallel to that of the Quantity Condition in section 2.

Non-Elitism

Suppose that w_i, w_{i+1} are positive welfare levels, such that w_{i+1} is lower than w_i, but the difference between them is small.

Then, for any number $N > 0$,

(NE) Any population $\mathscr{Q}_{i+1}$ with sufficiently many lives at level w_{i+1} and the rest at level z would be better than a popula-

> tion $\mathscr{Q}_i$ of the same size with N lives at level w_i and the rest at level z.

Using this, we can construct some $\mathscr{Q}_{i+1}$ that is better than $\mathscr{Q}_i$. As long as enough people were included in $\mathscr{Q}_0$, this process can continue until we reach $\mathscr{Q}_n$, a vast population of drab lives, at or below w_n. By construction, each $\mathscr{Q}$-population is better than the one before, so, by Transitivity, $\mathscr{Q}_n$ is better than $\mathscr{Q}_0$.

Now, to obtain the Repugnant Conclusion, we need only claim that $\mathscr{Q}_0$ is better than $\mathscr{P}_0$. This would follow, for example, from

> **The Mere Addition Principle**
>
> If $\mathscr{Q}_0$ can be obtained from $\mathscr{P}_0$ by adding lives that are worth living, and leaving all the other welfare levels alone, then $\mathscr{Q}_0$ is better than $\mathscr{P}_0$.

It is certainly possible to deny the Mere Addition Principle.[42] It is even possible to claim that it is indeterminate whether $\mathscr{Q}_0$ is better than $\mathscr{P}_0$. But that is not the response characteristic of the indeterminacy thesis. The characteristic response is rather to call into question the definite truth of the inductive step. In a sorites argument, it is the inductive step which is suspect.

Before discussing the indeterminacy thesis in detail, note the following heuristic justification for NE. If the benefiting group is large enough, they will gain many times more welfare in total than is lost in total by the people

[42] This is a relatively popular move; for example, it is part of the critical level utilitarianism espoused by Broome (2004) and others. On the other hand, more complicated spectrum arguments can be produced in which the Mere Addition Principle is replaced by intuitively weaker ones: cf. the use of the 'dominance addition' and 'non-sadism' principles in Arrhenius (2013).

who are harmed. However, this heuristic would seem to imply immediately that $\mathscr{Q}_n$ is better than $\mathscr{Q}_0$. In comparing $\mathscr{Q}_n$ with $\mathscr{Q}_0$ it is natural to wonder whether such a *large* loss of welfare to each of the many people in $\mathscr{P}_0$ can really be compensated by minuscule gains to however many others. So an important part of NE is that the best-off people lose only a little each time. Since it is the welfare level of the best-off people that changes gradually in each step of the spectrum argument, this is the parameter that is analogous to height in the tallness sorites.

To formulate the indeterminacy thesis in this context we need to put the spectrum argument in the form of a sorites. That is easy enough, following the model of section 3. The only complication is that the population $\mathscr{Q}_0$ is not well specified, requiring enough lives at level z to make the argument go through. But here is one way to do things. The vague predicate, the analogue of 'tall', will be 'H', defined by the following biconditional.

> A welfare level w is H if and only if any population with sufficiently many lives at level w, and the rest at level z, would be better than a population $\mathscr{Q}_0$ of the same size, consisting of $\mathscr{P}_0$ and additional lives at level z.

We then have

(A3) The first welfare level, w_1, is H.

(B3) If w_i is H, then w_{i+1} is also H.

(These follow from NE and Transitivity.) Therefore, (C3) The last welfare level, w_n, is H.

The Repugnant Conclusion then follows from (C3) and the Mere Addition Principle.

Like the earlier argument (A2,B2,C2), this one has exactly the form of a sorites argument. Assuming classical logic, the indeterminacy thesis for this spectrum argument must deny that every instance of (B3) is true. Assuming Transitivity, it denies that every instance of NE is true. This may be unpalatable, but vagueness provides mitigation: there are no determinate counterexamples to NE; every instance of NE is almost definitely true.

I have already said what I can here about the nature and extent of this mitigation, in sections 4 and 5. In the rest of this section, I will instead tilt the scales a little more in favour of the indeterminacy thesis by arguing that borderline violations of NE need not be symptomatic of *elitism* at all.

The point can be made most clearly with reference to the toy model of section 6. There NE fails when w_i and w_{i+1} straddle the level of sufficiency, α. There are two possible stories about the nature of this threshold.

The first possibility is that the threshold is *personal*: it is significant for the value of lives to the people who live them. There are a few ways to cash out this idea. Consider again the population $\mathscr{Q}_i$, with some lives at w_i and many others at z; compare it to a population $\mathscr{P}_{i+1}$ of the same size, with every life at w_{i+1}. In terms of chance, one might ask: would it be better for an individual to have a life chosen uniformly at random from $\mathscr{Q}_i$ than a life chosen uniformly at random from $\mathscr{P}_{i+1}$? Alternatively, one could use the 'serial lives' heuristic of Lewis (1946). Would it be better for an individual to live out the lives from $\mathscr{Q}_i$ in sequence than those from $\mathscr{P}_{i+1}$? If the answer to either question is affirmative, it gives a sense in which the threshold between w_i and w_{i+1} is significant for personal good. Indeed, according to these heuristics, it is inapt to think of the *evaluative* difference between w_i and w_{i+1} as being small. (We can still imagine that the *physical* and perhaps even the *phenomenal* difference between them is small.) According to the chance heuristic, for example, a welfare gain from w_{i+1} to w_i is so important for each individual that an arbitrarily small chance of it cannot be outweighed by a correspondingly large chance of a welfare loss from w_{i+1} to z.

Now, of course this is implausible. In imagining lives at levels w_i and w_{i+1}, we might specify that the only difference between them is some objectively small pleasure, like a grain of chocolate, or a 10-second headache. It is implausible that such a small difference in the underlying physical or experiential facts can make such a significant evaluative difference.The old point was that vagueness can mitigate this implausibility. Once we have specified w_i and w_{i+1}, the supposition that the threshold lies between them is almost definitely false; on the credential view of degrees of truth, we should have a very low credence in it. We should certainly not expect the location of the threshold to be detectable in our attitudes and judgments, even if those same attitudes and judgments suggest to us, via the spectrum argument, that the threshold must exist.

The *new* point is that, if the threshold is a personal threshold in this sense, then the charge of elitism does not stick. Even if there were a *determinate* counterexample to NE, it would not be an example of elitism. For the situation is not aptly described as one in which even a small harm to a few well-off people cannot be outweighed by benefits to many worse-off others. It is, rather, a situation in which an evaluatively *large* difference cannot be surpassed by aggregating evaluatively *small* ones, even when the small differences pertain to a single person's wellbeing. This is a doctrine not of elitism but of the kind of value superiority most famously espoused by Mill.[43]

The second interpretation of the toy model is that the threshold between w_i and w_{i+1} is of merely impersonal significance. This is the interpretation on which the toy model is really sufficientarian. Containing a life at w_i rather that w_{i+1} makes a big difference to the value of a population, even though, for an individual, the evaluative difference between these lives is small. Whatever one thinks about sufficientarianism – and about impersonal goods more generally – it would again be wrong to call this view elitist. It is not a matter of giving greater weight to the interests of the well off. There are two ways to see this, formal and substantive. Formally, suppose we are considering giving a fixed-size benefit either to someone who is definitely above the threshold, or to someone who is definitely below it. Then it is at least as good to give it to the person below the threshold, and sometimes it is better. This is not elitist; on the contrary, it illustrates the sense in which sufficientarianism is a cousin of egalitarianism. And, substantively, in the cases where it *is* better to give the benefit to the better-off person, this is not

[43] Mill (1863, chapter 2):

> If one of [two pleasures] is, by those who are competently acquainted with both, placed so far above the other that they. . . would not resign it for any quantity of the other pleasure which their nature is capable of, we are justified in ascribing to the preferred enjoyment a superiority in quality, so far outweighing quantity as to render it, in comparison, of small account.

See also Arrhenius and Rabinowicz (2005) for a formal analysis of this kind of value superiority.

because the person is better off, but because it raises them above the threshold. The mode of explanation is sufficientarian rather than elitist.

These two interpretations of the toy model show that, more generally, the indeterminacy view cannot be impugned by a generic charge of elitism. We have to look at specific axiologies in more detail to evaluate both the rationale for violations of NE and the role of vagueness in making such violations easier to maintain.

9 Conclusion

The argument I have given for the indeterminacy thesis rests on the two claims with which I started section 4. First, the argument (A2,B2,C2) has the same form as a sorites argument; second, it is paradoxical in the same way as a sorites argument. I later extended the thesis to cover the argument (A3,B3,C3) from non-elitism.

There are two main gaps in what I have said. First, although I considered (and, I hope, decisively refuted) several objections to the *first* claim about the form or structure of spectrum arguments, I did not say much in defence of the *second* claim, concerning their paradoxical nature. Of course, the main sense in which they are paradoxical is clear: to many people, at least, the premisses seem right and the conclusion wrong. This paper is most directly addressed to such people. But even such people may wonder whether the psychology of spectrum arguments is *quite* the same as the psychology of the sorites, and whether, given any subtle differences there may be, the core explanation can remain the same. The reason I have not said much on this score is, I am afraid, simply that it seems hard going. What exactly *are* the psychological differences, and, crucially, why do they *matter*? The standard explanations from classical theories of vagueness seem to adapt seamlessly to spectrum arguments. If they are sufficient in one case, why not in the other? This is an important question for understanding not only spectrum arguments but sorites arguments as well.

Second, granting the two claims, I have indicated why the indeterminacy thesis provides a *good* explanation of them, but I have not argued in any detail that it provides the *best* explanation. There is a reason for this. The indeterminacy thesis is not committed to any particular axiological principles. For example, in terms of the discussion of section 8, the thesis does not compete with sufficientarianism or value superiority; rather, it is an attempt to make such views more palatable. It must therefore be evaluated within a broader axiological theory. It would be futile to argue that the indeterminacy thesis is anything more than very good *pro tanto* without taking on (or surveying at length) more substantive commitments.

Still, the arguments I have presented shift the balance towards certain kinds of axiological views. These views endorse transitivity and reject RC. They also endorse trichotomy, or at least downplay the role of parity in the spectrum arguments. At the same time, they are views on which the inductive premisses like (B1,B2,B3) are *often* right. Unlike the average and person-affecting views I mentioned in section 2, they do not hold that these premisses fundamentally misfire. Rather, assuming classical logic, they posit thresholds at which the inductive premisses fail. The *prima facie* implausibility of such thresholds is exactly the sort of thing that classical theories of vagueness seek to explain away.

Of course, rejecting transitivity would allow us to hold that the premisses of the spectrum argument are definitely true, while rejecting the Repugnant Conclusion. But the indeterminacy thesis fares only slightly worse by that metric, and is far less revisionary in other ways. If intransitivity is on the table, then indeterminacy should be too. Similarly, one could of course accept RC, or deny the inductive premisses wholesale, without recourse to vagueness. But these responses are unattractive precisely to the extent that the spectrum arguments are paradoxical.

III | The Veil of Ignorance and the Risk of Non-Existence[*]

ABSTRACT. Roughly speaking, the 'veil of ignorance principle' identifies the moral point of view with the point of view of rational self-interest in the face of self-locating uncertainty. I sketch a positive argument for an axiological version of this principle, and explore its implications for population ethics.

1 Introduction

Moral evaluations are often supposed to be *impartial*. One sense in which they might be impartial is that they might correspond to the judgment of someone behind a 'veil of ignorance', ignorant of who he is in each alternative under evaluation. The judge surveys the world as if in third-person, seeing every detail but not recognizing which life is his own. Such a judge necessarily puts aside self-interest as usually understood: not knowing which life is his, he cannot judge on the basis of his own well-being. There are, however, at

[*] This chapter largely develops and interprets ideas that first appeared in 'Utilitarianism With and Without Expected Utility', joint work with David McCarthy and Kalle Mikkola (2016) (henceforth ***MMT***). I will make some specific connections to that work below, but I would like to acknowledge here the inseparable influence of McCarthy, in particular, on many points throughout, and especially in section 3. (On the other hand, he may not agree with everything I say!)

least two points of view open to him.First, he might be inspired to adopt an impersonal 'point of view of the universe', engaging in a distinctively moral form of evaluation, excluding self-interest altogether. Second, he might evaluate on the basis of 'veiled' self-interest – self-interest tempered by self-locating ignorance. In doing so, he would have to take every person's interests into account, on the self-interested basis that he might be any one of them. What I will call *the Veil of Ignorance Principle* – to be made precise below – is the idea that these two forms of evaluation coincide: the point of view of the universe is the point of view of veiled self-interest.

There is something deeply attractive about this principle, but, taken literally as referring to a self-interested judge, it faces a number of interpretive difficulties. The judge is stipulated to be ignorant of his own identity, but what other evidence is available to him? Does he maintain his own tastes and values? His own attitude toward risk? Or, if not, to what criteria does he appeal? I will be particularly interested in a problem arising from population ethics: if the alternatives under evaluation contain different people, then who is the judge supposed to be? Does he know that he exists? Besides such interpretive issues, there is the question of justifying the principle. Is there anything to be said beyond its prima facie plausibility?

In this paper I will present an argument for a precise version of the Veil of Ignorance Principle. It is *axiological* in that it concerns what is good for an individual, rather than what an individual prefers. It thus leaves open whether the former should somehow be analysed in terms of the latter; but, crucially, it sidesteps the worries mentioned above about the judge's idiosyncracies. It is *welfarist* in the sense that it assumes that the evaluatively salient features of each possible world are determined by the welfare levels of each person in that world (that is, by how good each world is for each person). There might be other features that are salient in principle, but I am assuming that they can be held fixed, or that we can ignore them for some other reason. Setting aside the *sui generis* difficulties of infinite populations (see Bostrom (2011)), I will assume that there are only finitely many possible individuals involved in any given context of evaluation. Enumerating all these relevant individuals in some way, I can then fully specify a world by a 'welfare distribution', a list of welfare levels $(x_1, x_2, \ldots, x_n)$. Here is the precise Veil of Ignorance Principle I will consider:[2]

The VoIP. Welfare distribution $(x_1, x_2, \ldots, x_n)$ is at least as good overall as $(y_1, y_2, \ldots, y_n)$ if and only if it would be at least as good for an individual to have equal chances of $x_1, x_2, \ldots, x_n$ rather than equal chances of $y_1, y_2, \ldots, y_n$.

In examples I will typically display welfare distributions L or M in tables like the following, in which each row gives the welfare level of the named person:

L	
Ann	x
Bob	y

M	
Ann	a
Bob	b

The VoIP says that L is better overall than M just in case it would be better for an individual to have a half-chance of x and a half-chance of y rather than a half-chance of a and a half-chance of b. I will call a probability

[2] At the end of section 3, I will introduce a more general version that applies to cases in which the welfare distributions are uncertain.

measure over welfare levels (e.g. a half-chance of x and a half-chance of y) a *prospect*, and a probability measure over welfare distributions a *lottery*. With one exception I will come to momentarily, I am simply going to assume that it makes sense to ask whether one prospect is better for an individual than another, and whether one lottery is better overall than another. This assumption was defended by John Broome in his *Weighing Goods* (1991, §6.1), and I agree with what he says. In fact, the project of this paper is closely related to his project in that book, in ways I will explain in section 2.

The VoIP, as I intend it here, makes sense even when different individuals exist in each alternative. We can formally allow that one of the 'welfare levels' that may occur in a welfare distribution is *non-existence*, traditionally denoted by Ω. This is not to claim that non-existence is a welfare level in anything but name. The situation is just that Ω can appear in a welfare distribution, and if it occurs in the ith place then it signifies that the ith possible person does not exist. For example, suppose that in the second alternative Bob does not exist:

M	
Ann	a
Bob	Ω

Then the question raised by the VoIP is whether it is better for an individual to have a half-chance of x and a half-chance of y, or a half-chance of a and a half-chance of not existing. This is again a comparison between two prospects in terms of their value to an individual. But it is the one kind of case where I will allow that it is controversial whether the comparison makes sense. Quite generally, the VoIP claims that the evaluation of welfare distributions boils down to what I will call the *Risky Existential Question*.

The Risky Existential Question. What should we say about the value for an individual S of a prospect in which there is a chance that S does not exist?

This generalises the better-known 'existential question' (Arrhenius and Rabinowicz, 2015): what should we say about the value for S of *certain* non-existence? There are two typical answers to this latter question. *Comparativists* claim that non-existence is comparable to other welfare levels, amounting to a 'neutral' level of wellbeing. *Non-comparativists*, in contrast, claim that non-existence is not comparable to existence at any level of well-being. It is not clear how the second (in particular) of these answers should generalise to cases of uncertainty.

Here is a prospectus. In section 2, I review some previous discussions of the veil of ignorance, as a way of introducing more fully the themes of this paper. The main message is that these previous discussions have been hampered and obscured by framing the veil of ignorance in terms of preferences instead of betterness. The present, purely axiological Veil of Ignorance Principle sidesteps all these difficulties.

Next, in section 3, I present an argument for the VoIP, based on three basic dominance principles. The fact that there is such an argument is, of course, the main positive feature of the VoIP. This argument and the discussion of section 2 should together revive the veil of ignorance as a way of understanding a number of issues in distributive and population ethics.

In that vein, starting in section 4, I explore and evaluate answers to the Risky Existential Question and their consequences for population ethics. Although there are a huge variety of potential answers, three basic stories suggest themselves.

Comparativism. Non-existence is comparable to other welfare levels; for example, it is better for an individual to have a very good life than never to exist at all.

In conjunction with the VoIP, comparativism leads to variations on critical-level utilitarianism (CLU). In fact, it yields an efficient argument for CLU, significantly weakening the premisses of Harsanyi's famous aggregation theorem, and extending it to the variable-population setting. (I will introduce Harsanyi's theorem in section 2.)

Strong Non-Comparativism. Any prospect in which the individual *S* is certain to exist is incomparable for *S* to any prospect in which *S* is not certain to exist.

In conjunction with the VoIP, this leads to extremely widespread incomparability: two alternatives can be compared only if they have the same expected population size.

Conditionalism: A prospect is exactly as good for an individual *S* as it is good for *S* *conditional* on *S*'s existence.

I think this is the most intuitively satisfying answer to the Risky Existential Question. In conjunction with the VoIP, however, it leads to problematic variations on average utilitarianism.

There are a number of lessons one might draw from this, depending on one's attitude towards the Risky Existential Question. First, one can find here an efficient argument for critical-level utilitarianism – either because one is antecedently inclined towards comparativism, or because comparativism leads to the only attractive view overall. Actually, as I explain in section 5, the argument for CLU also implicates the many non-comparativists who play down the significance of non-comparativism for overall value. A little roughly, these 'mixed' non-comparativists hold that non-existence is comparable to other welfare levels when it comes to overall value, even if not when it comes to value for the individual in question.

Alternatively, one can find here a constructive argument for average utilitarianism – about the only such argument I know. As I explain in section 2, an inference from the veil of ignorance to average utilitarianism is traditional, but three features are novel. First, I provide an argument for the VoIP itself. Second, I tie the inference to a particular theory of prudential value (namely, conditionalism). Finally, the particular form of average utilitarianism comes along with an unusual, and in some ways attractive, treatment of uncertainty, which I describe in section 5.

Less sanguinely, however, I think this raises a dilemma for those inclined to reject comparativism for individual value and (in contrast to the mixed non-comparativists) overall value as well.[3] On the one hand, average utilitarianism

[3] Such people endorse the view that 'We are in favour of making people happy, but neutral about making happy people' (Narveson, 1973, p. 80), at least at the level of axiology. The view that we are neutral about making *unhappy* people is much less popular, leading to the so-called 'asymmetry' between happy and unhappy lives (see e.g. Roberts (2011b)). From the perspective of this paper, accepting the axiological claim that bad lives are worse than non-existence puts one in the comparativist camp (or at least among the mixed-noncomparativists, if the claim is about overall value). That leads to further issues which,

has become unpopular, and for good reason. So it seems we must reject the combination of the VoIP and the otherwise intuitive conditionalism. On the other hand, the VoIP and strong non-comparativism together lead to such widespread incomparability as to render axiology impotent. Thus the pure non-comparativist must either give a plausible alternative to the dominance principles used to support the VoIP, or else answer the Risky Existential Question in a way that does not lead to an implausible or impotent axiology. To be sure, I do not claim that this is a *fatal* dilemma for non-comparativists: they have many theoretical resources to invoke. But it does illustrate how the Veil of Ignorance Principle provides a fresh way of looking at a number of issues in population ethics and the ethics of distribution that have otherwise grown stale.

2 A Motivating History of the Veil

Vickrey and Harsanyi

Although the phrase 'veil of ignorance' was coined by Rawls (1971), the kind of principle I am considering was articulated earlier (and independently) in the work of William Vickrey (1945) and John Harsanyi (1953). Although their treatments of the veil were slightly different, the general thrust of what I will say applies to both; I will focus on Vickrey for concreteness.[4]

Vickrey was writing at a time when expected utility theory had gained new prominence, with the publication of von Neumann's and Morgenstern's

however, I will not dwell on in this paper, in part because I think the asymmetry is not best understood in this purely axiological way.

[4] See Mongin (2001) for a careful comparison of the two treatments.

Theory of Games and Economic Behavior in 1944. Assuming that the preferences of an individual S satisfy some natural structural conditions (which I will consider in section 4), von Neumann and Morgenstern proved that there must be an 'individual utility function' U_S assigning real numbers to outcomes and satisfying the following rule:

Expected Utility for Preferences. S prefers an uncertain outcome A to an uncertain outcome B if and only if the expected value of U_S on A is greater than its expected value on B.

So, for example, S is indifferent between an option that guarantees utility 5 and an alternative that gives a half-chance of utility 4 and a half-chance of utility 6. Moreover, U_S is all but uniquely determined by S's preferences. One can fix the scale of utility by stipulating arbitrarily that one outcome has utility 0 and that another, preferred outcome has utility 1, but beyond that there is no freedom.

It is tempting to speculate that the 'utility' of expected utility theory coincides with the 'utility' of utilitarianism, so that the *average* or perhaps *total* utility of an outcome is normatively significant. Vickrey suggested that the veil of ignorance might vindicate this speculation:

> If utility is defined as that quantity the mathematical expectation of which is maximized by an individual making choices involving risk, then to maximize the aggregate of such utility over the population is equivalent to choosing that distribution of income which such an individual would select were he asked which of various variants of the economy he would like to become a

> member of, assuming that once he selects a given economy with a given distribution of income he has an equal chance of landing in the shoes of each member of it. Unreal as this hypothetical choice may be, it at least shows that there exists a reasonable conceptual relation between the methods used to determine utility and the uses proposed to be made of it. (Vickrey, 1945, p. 329)

Although in places Vickrey appears to identify 'aggregate' with *total* utility, this passage must concern *averages*: the individual's expected utility, given the choice of an 'economy', will be the average utility within that economy. However, when the same people exist in every alternative, there is no effective difference between average and total utilitarianism. In order to separate distinct issues, I will for now assume that the population is constant in this way, returning to the variable-population case in my discussion of Rawls below.

In fact, Vickrey's 'reasonable conceptual relation' is deeply problematic. For even if the individual behind the veil would seek to maximize average utility, it does not follow that *we* have any reason to do so; the formal observation that an average can be interpreted as an expectation has no clear normative significance.[5] And even the description of the veil raises conceptual and interpretive difficulties. Reference is made to what 'an individual would select', but *which* individual? In the first place, it is implausible that any

[5] On this point see Barry (1989, pp. 334–5); Broome (1991, p. 56–57). A distinct debate concerns to what extent the 'utilitarianism' that supposedly emerges from the veil argument agrees with utilitarianism as traditionally conceived (see Greaves (MSb) for a recent overview).

realistic individual in fact has preferences precisely consistent with Expected Utility for Preferences. In the second place, it is implausible that any two expected-utility maximizers necessarily have the same preferences in the counterfactual circumstances ('were he asked…'). In particular, it is not clear why the utilities appearing in the average can be identified with the utilities recognised by the members of the economy themselves. Suppose Ann is one such member. Ann (let us accept for the sake of argument) implicitly assigns utilities to various outcomes, and seeks to maximize the expected utility. The judge behind the veil also (we have to assume) assigns utilities to various scenarios in which he finds himself 'in the shoes' of Ann. *A priori*, the latter utilities are the ones that enter into the average, not those corresponding to Ann's own preferences.[6]

Later discussions of the veil of ignorance recognised difficulties along these lines. For example, Vickrey himself starts his 1960 discussion with

> the heroic assumption that, abstracting from differences in age, sex, or family status, each individual has preference patterns

[6] The claim that these utilities coincide is a version of the 'principle of acceptance' used by Harsanyi and others to explain interpersonal comparisons of well-being in terms of preference satisfaction (see Greaves and Lederman (MS) for references and a critical discussion), as well as by Weymark (1991, p. 293) in his formalisation of Harsanyi's veil as the 'Impartial Observer Theorem'. Note that there are at least two issues here. First, there is the issue of whether the judge's preferences over possibilities for each individual agree with that individual's own preferences. Even if they do, there is the further question of whether the judge's preferences get the interpersonal comparisons right. Remember that each individual's utility function can be normalised in different ways; the problem is that the notion of average utility is only sensible if we choose compatible normalisations for the individual's utility functions. Naively, 'compatible' means that, for any two individuals S and T, $U_S(X) = U_T(Y)$ if and only if S prefers X just as strongly as T prefers Y (in a monadic sense of 'prefers'). The question is whether the judge is then indifferent between X-in-S's-shoes and Y-in-T's-shoes, so that he assigns the same utility to these two outcomes. Issues along these lines have been raised in particular by Sen (1977) and Weymark (1991).

> exactly similar to those of every other individual.... (Vickrey, 1960, p. 524)

But our axiological version of the veil of ignorance completely sidesteps these problems. To get a preference-based version of the veil, we just need to posit a connection between what is better and what is preferred. We could make the heroic assumption that each person actually prefers what is best for themselves, or, much more plausibly, say that anyone would *ideally* prefer what is best for themselves (this being an ideal of *self-interested* rationality). We then obtain the following principle.

The VoIP for Ideal Preferences. Welfare distribution $(x_1, x_2, \ldots, x_n)$ is better overall than $(y_1, y_2, \ldots, y_n)$ if and only if some (hence any) individual would ideally prefer to have equal chances of $x_1, x_2, \ldots, x_n$ rather than equal chances of $y_1, y_2, \ldots, y_n$.

We can then posit that these ideal preferences satisfy expected utility theory, generating an ideal individual utility function with respect to which betterness overall (or 'ideal moral preference') tracks average utility. (Recall here my temporary assumption that the same people exist in every alternative.) To reach such a conclusion, however, we do not have to mention preferences at all. We can simply posit that the individual betterness relation itself satisfies expected utility theory: there is a numerical representation of welfare levels satisfying the following condition.

Expected Utility for Individual Betterness. Prospect A is better for an individual than prospect B if and only if it has greater expected welfare.

Vickrey himself seems to have something like the VoIP for *actual* rather than *ideal* preferences in mind. The underlying premiss that people actually prefer what is best for themselves faces the same difficulties I raised against Vickrey's account of the veil: actual preferences, even self-regarding ones, are very frequently idiosyncratic and irrational. I think the origin of this premiss is a preference-satisfaction theory of wellbeing, which holds, in its naivest form, that if S prefers A to B then, *for that very reason*, A is better for S than B. The idiosyncracy and irrationality of actual preferences show why at least this naive form of preference-satisfactionism is so hard to defend.

The diversity of actual and indeed 'reasonable' preferences might seem to pose a problem even for the axiological veil of ignorance. In particular, consider the view that a wide variety of attitudes towards *risk* are fully rational.[7] It may be possible that, even though they are both fully rational, Ann prefers welfare y to half-chances of x and z, while Bob prefers the latter to the former. Ann is more risk-averse than Bob, in this instance. This pattern is possible even if the ideal preferences of Ann and Bob both satisfy expected utility theory, but it may arise also if some deviations from expected utility theory are compatible with full rationality. One *might* conclude from this that y is better for Ann than half-chances of x and z, while the opposite is true for Bob. It would then be nonsense to ask (as the VoIP does) whether y is better for an individual than half chances of x and z, without specifying *which* individual.

But this conclusion is the wrong one to draw. After all, it is not that there is some idiosyncratic attitude towards risk that is right for Ann and wrong for

[7] See Barry (1977, p. 318) for this as an objection to the veil of ignorance.

Bob independent of the circumstances in which they find themselves. Rather, they may have some prerogative, in light of those circumstances, to settle what the betterness facts leave unsettled. It may be rationally permissible, but not required, for Ann and Bob to have the preferences they do, and this is presumably because the betterness facts are themselves indeterminate in some way. I am using a very broad notion of 'indeterminacy' here: all that matters is that fully rational preferences are not fully constrained. It may be, for example, that y is *on a par* with half chances of x and z. The axiological VoIP is perfectly compatible with this kind of indeterminacy.

Rawls and Kavka

Rawls's understanding and proposed application of the veil of ignorance was much more elaborate than Vickrey's, and he argued that it led ultimately to his own non-utilitarian principles of justice. But in the intramural dispute over average versus total, he agreed that the veil differentially supported the average view (Rawls, 1971, §27). In evaluating a possibility from behind the veil one should act as if uncertain who one is, but certain that one does exist in that world. Thus, in agreement with average and in contradiction to total utilitarianism, a world with one person at welfare level 100 is better than a world with 100 people each at welfare level 99: one is certain to have higher welfare in the former. Note that this particular judgment can also be explained in terms of Rawls's favoured maximin rule: one distribution is better than another if the worst-off person in the first is better off than the worst-off person in the second.

To get to maximin from the veil of ignorance, one must appeal to a principle like

Pessimism. A prospect is only as good as its worst possible outcome.

The application of Pessimism and the maximin rule itself have been roundly criticised by Harsanyi (see Harsanyi and Rawls (1975)). Someone with pluralist tendencies might think that it is rationally *permissible* to have extremely pessimistic preferences, but at issue here is the stronger claim that such preferences are rationally required. Pessimism says that a large chance of a large gain never compensates for a small chance of a small loss. This is clearly too extreme. However, Rawls's position is that the judge is in so deep state of cluelessness (and perhaps there are some other features of the case) that he cannot assign probabilities at all to each outcome; there is no question of large chances versus small ones. We should thus not understand Pessimism to apply to 'prospects' in my technical sense of the word – probability measures over welfare levels – but to situations that involve this deeper cluelessness.

If the question is, as a general matter, how to evaluate prospects involving some sort of deep cluelessness, then I will agree with Rawls that the equi-probability assumption and the application of expected utility theory are open to question, even if Pessimism itself seems too extreme. However, as things stand, this question is largely irrelevant. To the extent that the veil of ignorance is merely a recipe for reconstructing moral evaluations, then the rules for evaluating prospects behind the veil are by and large open to stipulation. The application of Pessimism can only really be judged on its output, i.e. on the utter implausibility of maximin. Only when we have a serious independent argument for some particular version of the veil of ignorance principle –

linking moral evaluation to some particular kind of ignorance – does it become relevant to ask how one should evaluate prospects of the requisite kind.

Our VoIP gives equal chances to each welfare level in the distribution. In that sense, our argument for the VoIP vindicates Harsanyi and Vickrey over Rawls. It is neutral in another way: it officially leaves open the question of how prospects are to be evaluated, given these equal chances. Pessimism is still on the table, in that sense, and would imply maximin.[8] However, the question that arises is the conventional one about how to evaluate prospects that involve probabilities, not prospects that involve deep cluelessness. And Pessimism is a completely implausible answer to *that* question.

In Rawls's version of the veil – and those of Vickrey and Harsanyi – the judge is certain to exist in the prospect under evaluation. Gregory Kavka noted the significance of this fact:

> …Rawls' conception of the original position involves a (possibly justifiable) bias in favor of those already existing in the sense that it favors the interest of existing persons over the interests which would exist if certain persons who might or might not exist were brought into existence. (Kavka, 1975, p. 240)

This passage contains two distinct thoughts that ought to be disentangled. One thought, less relevant here, is that Rawls's veil improperly focuses on the *present* generation, those 'already existing'. As Kavka later argues, it is at least

[8] There are different ways to extend Pessimism to answer the Risky Existential Question; different ways of doing it lead to different versions of maximin.

sometimes more appropriate to consider all generations together, and that is what I have in mind throughout. The second, more relevant thought is that Rawl's veil discounts the interests of merely possible people, those 'who might or might not exist'. The simplest way to explicate this second thought is to suppose, with Kavka, that non-existence carries with it the utility value of zero. Then, if one evaluates a world solely in terms of the welfare of the people who exist in that world, one may not properly take into account the *zero* welfare of the people who do not exist.

At a superficial level, it is easy to adjust the veil of ignorance to support total rather than average utilitarianism, and this is what Kavka does. When comparing alternatives A and B, we should consider ourselves to have equal chances of being any individual that exists in *A or B*.[9] Thus if B contains some individuals who do not exist in A, we should, in evaluating A, take into account some probability of not existing, and therefore having zero utility. More generally, if we allowed that non-existence carries with it some *possibly non-zero* utility, the veil of ignorance so formulated would support critical-level utilitarianism.

Quite besides the controversial move of associating a utility with non-existence, this modification of the veil of ignorance raises the question of whether one can be less than certain of one's own existence. Kavka considers this point and offers several suggestions. The simplest is to give up the idea that the judge behind the veil is evaluating on the basis of *self*-interest. Rather, he can make a choice for the sake of a client individual who might or might not exist in each alternative, and whose identity is unknown. In the same

[9] See Tännsjö (2002, p. 343) for a slight variation on this move.

way do we make choices for the sake of our possible descendents.

This seems about right to me. But at bottom we can get rid of the judge altogether, by (again) focusing on the purely axiological VoIP. Presumably the guardian's judgment is based on a concern for what is *good* for his client; insofar as he strays from this criterion, he cannot be said to be choosing for the client's sake. More generally, insofar as the figure of the judge or guardian makes sense, the axiological VoIP delivers the desired result; but we need not worry if the figure does not always make sense. Of course, thinking about a guardian's judgments may still give us a useful way to think about what is good for the client. Perhaps some would even wish to *analyse* the client's good in these terms.[10] The point remains that we can reframe the veil of ignorance in axiological terms, and then leave it to the others to defend a 'guardian angel' view of individual value.

Harsanyi and Broome

Now let us return to John Harsanyi. He had invented the veil of ignorance independently of Vickrey, with essentially the same purpose in mind (Harsanyi, 1953). But he is justly more famous for the aggregation theorem (Harsanyi, 1955, 1977), which, like the veil of ignorance, gives an account of why the 'utility' of utilitarians might coincide with the 'utility' of expected utility maximization. The underlying logic of the theorem is quite different from the veil-based argument, and, officially, it deals only with a single population that exists in every alternative. Since the theorem provides a useful point of

[10] For a discussion and references of this analytic move, see Arrhenius and Rabinowicz (2015, §22.6)).

comparison, let me sketch how it goes.[11]

Like Vickrey, Harsanyi dealt with preferences rather than with value. However, as John Broome argues in *Weighing Goods*, the aggregation theorem can and should be recast axiologically. (Thus we do for the veil of ignorance what Broome did for the aggregation theorem.) In those terms, the theorem has three main premisses. First, it assumes that both individual and overall value satisfy expected utility theory. This means that there are 'social' and 'individual' utility functions, so that each lottery is ranked by expected social utility, and each prospect by expected individual utility. (As before, there is a little freedom in how to normalise these utility functions.) The second assumption is the Strong Pareto principle:

Strong Pareto. Let L and M be lotteries.

1. If L is at least as good as M for every individual, then L is at least as good as M overall.
2. If L is at least as good as M for every individual, and better than M for some individuals, then L is better than M overall.

(Broome calls this 'the principle of personal good', to distinguish it from its traditional formulation in terms of preferences.) The third assumption is *Anonymity*, the principle that if L and M differ only by a permutation of individuals, then L must be just as good as M. The conclusion is that we can normalise the social utility function so that it assigns to each welfare distribution its average individual utility.

[11] My discussion here draws particularly on *MMT* §5.3 and §5.4.

Harsanyi's proof does not derive a form of the veil of ignorance principle and *then* note that expected utility theory is the standard account of prudential value, thus deriving the utilitarian aggregation rule. Rather, he derives the utilitarian aggregation rule directly, which may optionally be seen as justifying the veil of ignorance principle post hoc. Indeed, the original version of Harsanyi's theorem catered to a widespread skepticism among economists about the possibility of interpersonal comparisons of wellbeing and/or preference strength. (Harsanyi himself thought that this skepticism was misguided.) Thus his main result was not the simple utilitarian rule but just the claim that overall value had to be some weighted average of utility. The undetermined weights reflect the undetermined interpersonal comparisons. Harsanyi went on to derive the utilitarian rule almost as an afterthought, only at this stage using Anonymity to set all weights equal. This last step certainly presupposes interpersonal comparisons.[12] But the main result forgoes such comparisons, and, without them, the veil of ignorance principle makes little sense. It fundamentally involves including the welfare levels of different people within a single person's prudential calculus.[13]

[12] To see why Anonymity requires interpersonal comparisons, consider a world w containing only Ann and Bob. Permutation invariance means that w is just as good as a world w' in which Ann's life is just as good as Bob's life is in w and Bob's life is just as good as Ann's life is in w.

[13] It may be added that the project of aggregation makes little sense either. If there really are no interpersonal comparisons, then there is no sense in which one person's loss can outweigh another person's gain, and we should not expect to say much except in cases of unanimity. (This is one possible lesson of Arrow's impossibility theorem.) At best we should expect a theory of overall value with widespread incomparability. This creates difficulties for Harsanyi's basic theorem – the one without Anonymity – since it assumes that the overall betterness relation (or, in Harsanyi's terms, the relation of moral preference) is complete. For a further discussion of this well-known issue see e.g. Mongin (1994, pp. 349–350) and *MMT* §5.3. In light of this, it may be best to interpret that basic theorem in the following terms: if there really are no interpersonal comparisons, then one outcome is at least as good as another if and only if *every* weighted average of individual utility functions gives it a

In principle, Harsanyi's theorem applies to the variable population context, as long as we are willing to assign an individual utility value to non-existence. The theorem then yields *critical level utilitarianism*.[14] Remarkably, Harsanyi himself was unwilling to go this route. Instead, he vehemently endorsed average utilitarianism, claiming for it 'incomparably superior results' (Harsanyi, 1977, fn. 12). As this phrase suggests, he was impressed by standard intuitions against total utilitarianism, including considerations similar to the Repugnant Conclusion, and the general sense that increasing population size is morally neutral. (These concerns are particularly clear in his correspondence quoted in Ng (1983).) But of course he also appealed, without real argument, to the veil of ignorance. This reinforces the point that Harsanyi's aggregation theorem originally followed a different line of thought, unrelated to the veil.

However, the present paper shows that the two lines of thought are closely related after all. The premisses we use to argue for the VoIP are much weaker than the ones used in Broome's reconstruction of Harsanyi's theorem, and the basic argument is much less technical. If, in addition, we assume Expected Utility for Individual Value, the utilitarian conclusion of Harsanyi's theorem follows right away.[15] Thus the veil of ignorance may well provide the best way of proving and understanding Harsanyi's theorem, overcoming its reputation as the latter's poor relation.

higher value. But this is just the Pareto preorder (i.e. the one generated by the Strong Pareto condition in the text).

[14] Broome (2004) gives a different argument for critical level utilitarianism based on Harsanyi's theorem, which does not rely on a 'utility of non-existence'. The relationship between these arguments is rather subtle; I will say a bit more in section 5.

[15] I return to this point in §4 below.

Summary

At this point we lack two things. First, we lack any real justification for any version of the veil of ignorance principle that does not already presuppose that individual and overall value are matters for expected utility theory. This is especially important because it is much less clear than usual that expected utility theory is appropriate in cases of possible non-existence. Indeed, the second thing we lack is a well-justified story about how individual value depends on the probability of non-existence. In other words, we lack a well-justified answer to the Risky Existential Question.

3 An Argument for the Veil

In this section, I explain how three plausible principles entail the VoIP. In fact, they are equivalent to a generalised version of it. What is the status of these principles? They derive from some well-known examples in distributive ethics. They can be understood as rejecting certain kinds of egalitarian and prioritarian concerns. Thus the argument for the VoIP most *directly* implicates those broadly in the utilitarian camp – for example, people who accept utilitarianism in fixed-population cases, and are wondering how to extend their theory to variable-population cases. However, the three premisses are much more widely acceptable than any particular utilitarian theory, and in particular do not require any sort of numerical representation of welfare.[16]

[16] The results I discuss in this section are essentially the main Theorems 1.3.1 and 2.3.1 in *MMT*. There the three principles are called *Posterior Anonymity*, *Anteriority*, and *Reduction to Prospects*, respectively. Their relation to egalitarianism is briefly discussed in §5.1 of that work. For the connection to prioritarianism see e.g. McCarthy (MS)

Moreover, egalitarians and others should still be interested in the argument, for two reasons. First, they should go along with it at least in cases when equality, priority, and so on, are not at stake. Second, we can understand the three principles as characterising a kind of impartial but 'beneficent' or 'personal' value, setting aside the further question of how to combine this personal value with impersonal values such as equality and priority. The phrase 'personal good' is sometimes used the way in which I here use 'individual good', i.e. concerning what is good for one individual. For example, that is how Broome uses the phrase. But there is a different way of thinking on which overall as opposed to individual evaluations can still be 'personal', concerning perhaps 'plural facts about personal good' (Bader, 2014, p. 5). This idea is hard to pin down, but the three principles may be viewed as one way of doing so: after all, the VoIP gives a particularly concrete way in which overall value might be reducible to individual value. I will have more to say about this 'personal good' interpretation of the VoIP as we go, and especially when I turn to non-comparativism in section 5.

(1) Against Fairness

Here is the first principle. Consider these two lotteries.

L	
Ann	x
Bob	y

L_1	H	T
Ann	x	y
Bob	y	x

The notation is meant to suggest that, under L, it is certain that Ann gets welfare level x and Bob gets welfare level y, whereas, under L_1, there is a half chance (corresponding to a coin landing Heads) that Ann gets x and Bob gets y, and a half chance (corresponding to Tails) that Ann gets y and Bob gets x.

Here is one way to reason about these two lotteries. Under both L and L_1, it is certain that *one person* will get x and that *the other* will get y. But it does not morally matter *which* person gets x and *which* gets y: Ann and Bob have equal status. So L and L_1 are equally good. (Note here that x and y are lifetime welfare levels. This is important because if, on the contrary, x and y represented one-time benefits, then it might matter who got which. For example, we might want to award the larger benefit to the person who was antecedently worse off.)

More generally, we can distinguish between a welfare distribution, which assigns a specific welfare level to each specific individual, and an 'anonymised distribution', which specifies how many people get each welfare level, but not which specific individual gets each one. Since a lottery assigns a probability to each welfare distribution, and each welfare distribution determines an anonymised distribution, each lottery assigns a probability to each anonymised distribution. The example suggests the following general principle of impartiality:

Principle 1. If the same anonymised distributions get the same chances in L as in L_1, then L is just as good as L_1.

This principle can be questioned. Diamond (1967) suggested that L_1 is, at least sometimes, *fairer* than L. When distributing goods it is better to distribute them randomly, as in L_1, than in some predetermined manner, as in L. With respect to fairness, or, as this type of fairness is sometimes known, ex-ante equality, L_1 may be better than L. Therefore, since other things appear to be equal, L_1 may be better than L overall.

This argument has some plausibility. But several responses are available. The first kind of response is that the kind of fairness in question is not ultimately good *for* anyone, since that would already be reflected in the welfare levels. So even if fairness is broadly relevant, there is still a person-affecting respect in which L_1 and L_2 are equally good. This is one place where it would make sense to focus on 'personal' value, and understand Principle 1 in those terms. Going a little further, it is not clear that this kind of fairness is axiological at all; it may be best understood as a non-axiological ideal of justice.

The second kind of response (really a spelling-out of the first) appeals to a separate principle of *stochastic dominance*. To consider the case at hand, suppose we accept Diamond's suggestion that L_1 is better than L. Suppose we modify L_1 by imposing a small cost on each of Ann and Bob. In other words, suppose that we can find welfare levels x^- and y^- very slightly worse than x and y respectively; the modified lottery will be

L_1^-	H	T
Ann	x^-	y^-
Bob	y^-	x^-

If the cost is small enough (i.e. if x^- and y^- are close enough to x and y), then the resulting lottery L_1^- is presumably so similar to L_1 that it is still better, or at least no worse, than L.[17] On the other hand, the outcome of L_1^- on Heads must be worse than L, since everyone is worse off. And the outcome of L_1^- on Tails is only as good as the outcome on Heads, since they only differ by a permutation of individuals. In summary, Diamond's judgment suggests that L_1^- is no worse than L despite the fact that it is certain to have a worse outcome. (More broadly: despite the fact that L 'stochastically dominates' L_1^-.) This is counter-intuitive.[18]

(2) Against Equality

Consider the following two lotteries.

L_1	H	T
Ann	x	y
Bob	y	x

L_2	H	T
Ann	x	y
Bob	x	y

Under both L_1 and L_2, Ann has a half chance of heads and a half chance of tails: she faces the same prospect either way. So too with Bob. From that point of view, L_1 is just as good as L_2. More generally, we might accept

Principle 2. If each person faces the same prospect in L_1 as in L_2, then L_1 is just as good as L_2.

[17] As long as x and y do not correspond to non-existence, it is rather plausible there exist slightly worse welfare levels, as this argument requires. If x and y both equal Ω, there is no problem either, since $L = L_1$. If (say) $x = \Omega$ and $y \neq \Omega$, then the argument seems convincing if we take y^- slightly worse than y, and $x^- = \Omega$.

[18] In section 5, I will consider a version of average utilitarianism that does not in general imply that L is better than M if L stochastically dominates M. But it does agree with the limited principle needed here, that L is better than M if it is certain to be better for every individual up to permutation.

This is a very weak 'person-affecting' condition. In order for L_1 to be different in value from L_2, it must be different for some person – not necessarily different in value for that person, but at least it must make a difference to the prospect she faces.[19]

Nonetheless, as Myerson (1981) originally pointed out, under L_2, but not L_1, there is certain to be perfect equality. With respect to equality, L_2 may be better than L_1, and thus, since all else seems to be equal, L_2 may be better than L_1 overall.

As in the case of fairness, a response is possible even if we accept that equality is broadly relevant. Since this kind of equality does not affect anyone's prospects, we can focus on personal value. We can also explicate the intuition in favour of Principle 2 through a dominance argument. Suppose, in the case at hand, we accept Myerson's judgment that L_2 is better than L_1. Then (as before) we should be able to impose a small cost on each of Ann and Bob to create a lottery L_2^- that is still better than L_1:

L_2^-	H	T
Ann	x^-	y^-
Bob	x^-	y^-

However, the prospect faced by Ann under L_2^- is certainly worse than her prospect under L_1: she is worse off on Heads, and worse off on Tails. Meanwhile, Bob's prospect in each lottery is the same as Ann's. So his prospect under L_2^- must be worse than his prospect under L_1. (In short, each person's

[19] It has sometimes been claimed that two lotteries can be incommensurable in value for an individual, even though the individual faces the same prospect in each one. I will return to this point briefly in section 5.

prospect under L_1 stochastically dominates his or her prospect under L_2^-.) We thus find that L_2^- is no worse than L_1 despite the fact that each individual faces a worse prospect. This is counter-intuitive.[20]

(3) Against Priority

Finally, consider these two cases.

L_2	H	T
Ann	x	y
Bob	x	y

M_2	H	T
Ann	a	b
Bob	a	b

When we look at each of these alternatives, we see that every individual is in the same position; and they are certain to remain in the same position as each other. There is no issue of fairness, there is no issue of equality, and there are no tradeoffs to be made between the interests of different individuals. There is, I will say, perfect unanimity. It seems that the only question here is whether the prospect faced by each and every individual is better for that individual in L_2 than it is in M_2. That is, the only issue here is whether a half-chance of x and a half-chance of y would be better for an individual than a half-chance of a and a half-chance of b. In general:

[20] This argument is related to the objection to egalitarianism known as 'levelling down'. In levelling down, every person is brought down to the level of the worst-off. One form of the objection claims that levelling down cannot be an improvement in any respect, i.e. not even pro tanto (Parfit, 1997, p. 211). A more dialectically effective form of the objection holds that levelling down cannot be better all things considered. Presumably the general principle is that making everyone worse off cannot make things better all things considered. That is the principle I have appealed to here.

Principle 3. In cases of perfect unanimity, L_2 is at least as good as M_2 if and only if it is at least as good for each (hence every) individual.

This is a version of the *Pareto principle*. It is much weaker than Strong Pareto in one sense: it only makes a claim about cases of perfect unanimity. It is slightly stronger than Strong Pareto in another sense. For suppose that having half-chances of x and y is *incomparable* to having half-chances of a and b. In that case, Strong Pareto makes no claim about the relative merits of L_2 and M_2. In contrast, Principle 3 claims that L_2 and M_2 are incomparable. It is hard to see why one would accept the Strong Pareto principle in cases of perfect unanimity, but deny the claim about incomparability.[21]

There are two popular ways of denying Principle 3. One has to do with the Risky Existential Question, and I will discuss it in section 5. The second has to do with prioritarianism. Prioritarians typically think that overall value should be *more risk-averse* than individual value. One way of getting this intuition appears to me misguided. It may seem that it is permissible to accept more risk when we are choosing for ourselves than when we are choosing for other people.[22] But this may just illustrate the moral prerogative we have to make suboptimal choices for ourselves. Prioritarians who wish to deny Principle 3 should instead appeal to their general notion of priority. Consider (to make things more concrete) the following lotteries involving

[21] Cf. discussion of the extended Pareto principle 'P_3' in *MMT* (§3.2 and §5.2). The application of Pareto-like considerations in the presence of incomparability is subtle when there is *not* perfect unanimity. Suppose welfare levels x and y are incomparable. Then the welfare distributions (x, y) and (y, x) are incomparable for each individual, but it does not seem right to claim that they are themselves incomparable. Rather, by Anonymity (or Principle 1), they are equally good. I will mention some related issues in section 5 below.

[22] On this thought on the context of prioritarianism, see Parfit (2012, p. 423); for comments on risk-aversion from behind the veil, see Rawls (1971, pp. 143–144).

Ann only.

G_1	H	T
Ann	2	0

G_2	H	T
Ann	1	1

Suppose that, as far as Ann goes, we should compare these lotteries based on expected welfare. They both have the same expected welfare, 1, so they are equally good for Ann. Nonetheless, prioritarians tend to favour G_2 over G_1, because they think the welfare difference between 0 and 1 (on Tails) is more important to overall value than the welfare difference between 1 and 2 (on Heads).[23]

Whatever other critiques there may be of axiological prioritarianism, the dominance objection here is obvious. If G_2 is really better overall than G_1, then we should be able to impose some small cost on Ann. For example:

G_2^-	H	T
Ann	$1-\varepsilon$	$1-\varepsilon$

For some suitably small cost ε, G_2^- must still be no worse than G_1, even though it is worse for the only person whose welfare is at stake.[24] In these one-person cases it is particularly clear what it means to focus on 'personal' value.

[23] Prioritarians don't have to accept this; they could claim that facing a bad prospect (rather than the possibility of a bad outcome) is the condition for priority. But such an 'ex ante' prioritarian should accept the judgment that G_1 is just as good as G_2, and similarly should accept Principle 3.

[24] The attentive reader may notice that this dominance argument is not quite strong enough to establish Principle 3; I will discuss the remaining loop-hole in section 5.

The result

Claim: Principles 1–3 jointly entail the VoIP.[25] I have already secretly given the proof, in the case of two individuals! To recapitulate:

L			L_2	H	T
Ann	x	$\simeq$	Ann	x	y
Bob	y		Bob	x	y

M			M_2	H	T
Ann	a	$\simeq$	Ann	a	b
Bob	b		Bob	a	b

Given the first two principles, L is just as good as L_2, and M is just as good as M_2. And according to Principle 3, L_2 is better than M_2 if and only if half-chances of x and y are better for an individual than half-chances of a and b. The general proof follows the same pattern, and is given in detail and great generality in *MMT* (Theorem 2.3.1).

In fact, we prove there that the three principles are *almost* jointly equivalent to the VoIP. They are equivalent to a stronger form of the VoIP that applies to lotteries (given the kind of background assumptions mentioned in footnote 25). Suppose that, in lottery L involving n people, the ith individual faces prospect $L(i)$. Roughly speaking, we want to say that evaluating L overall is equivalent to evaluating the prospect faced by an individual who

[25] The fine print: for the argument to go through, we need certain *domain conditions* – conditions to the effect that all relevant lotteries exist – and certain technical assumptions to make sensible the machinery of probability theory. The way I have phrased Principle 1 also presupposes that all lotteries in question have countable support, but it has a natural extension to a more general case, which I will omit here.

has equal chances of *subsequently* facing one of $L(1)$ through $L(n)$. As a single prospect – that is, as a probability distribution over welfare levels – this is represented by the mixture $\frac{1}{n}(L(1) + L(2) + \cdots + L(n))$.

The VoIP for Lotteries. Lottery L is at least as good as lottery M overall, if and only if it would be at least as good for an individual to face the prospect $\frac{1}{n}(L(1) + L(2) + \cdots + L(n))$ rather than the prospect $\frac{1}{n}(M(1) + M(2) + \cdots + M(n))$.

4 Comparativism

Having given the argument for the VoIP, I now turn to consider the Risky Existential Question, and, in particular, how the VoIP links it to issues in population ethics. In this section, I explain how comparativism leads to generalisations of total utilitarianism, and (in light of that) explore different ways in which comparativists can avoid the Repugnant Conclusion, one of the key objections to total utilitarian axiology.

Problem-Cases for the Risky Existential Question

First, though, let me step back a little, and introduce the issues more fully. The VoIP has some immediate consequences for population axiology, independent of the Risky Existential Question. For example, suppose welfare distribution P is constructed from Q by doubling the population size and, in particular, doubling the number of people at each welfare level. Suppose that P' is constructed from Q' in the same way. It follows immediately from the VoIP that P is better than P' if and only if Q is better than Q'. More generally, according to the VoIP, the betterness relation is invariant under population scaling: the value relation between two distributions remains the same if we scale up the population size in each one by the same factor.[26] This rules out, for example, *variable value* theories, which give judgments similar to those of total utilitarianism for small populations, and judgments similar to average utilitarianism when the small populations are scaled up.[27]

To go further, however, we need an answer to the Risky Existential Question – and, quite independently of the VoIP, it is natural to wonder whether, and when, the Risky Existential Question admits a non-trivial answer. Consider

Certainty. Only one person, Jill, has any chance of existing. In one alternative, Certain Existence, it is certain that she exists, and that she will have a very good life. In another alternative, Certain Non-Existence, it is certain that she will not exist. Which of these is better for Jill?

There are two standard ways of answering this question. Some people are happy to say that Certain Existence is better for Jill than Certain Non-Existence; they are *existence comparativists*.[28] But others – *non-comparativists* – deny it. It is a category mistake to ask whether the number two is redder, or less red, or just as red as the sun; being an abstract object, it does not register on the scale of redness. So too, non-comparativists hold that asking whether

[26] This is *MMT* Proposition 5.1.1.

[27] According to a concrete version of the variable value view described by Ng (1989), a welfare distribution of n people with average welfare w has aggregate value $(1-(9/10)^n)w$. Suppose that Q has one person at level 10, while P has 1000 people at level 10; and Q' has two people at level 9, while P' has 2000 people at level 9. Then the theory ranks P above P' but Q' above Q. Similar theories were proposed by Hurka (1983) and Sider (1991).

[28] I am simply ignoring for simplicity the possibility of being an existence comparativist and holding that a 'very good life' is not better than non-existence.

Certain Existence is better for Jill than Certain Non-Existence involves a kind of category mistake. Non-existence does not register on the scale of Jill's wellbeing, and is thus (in particular) neither better, nor worse, nor just as good for Jill, compared to existing with a very good life.

But what about a *chance* of non-existence? I will consider non-comparativists starting from section 5. Here I consider comparativists, for whom there is an obvious answer. In evaluating prospects with chances of non-existence, one must simply hedge between the value of non-existence and the values of different states of existence, in whatever way one normally hedges when evaluating chancy scenarios. In particular, the paradigmatic comparativist view will hold that individual value satisfies Expected Utility for Individual Value (see section 2). They will thus grant a numerical representation of welfare levels, including a numerical value for non-existence. This gives antecedently plausible results in many examples.

Vitamin Z. You are in the very early stages of pregnancy. Taking vitamin Z now will improve the life of the child, Jill, if she is born. It will also very slightly increase the risk of early miscarriage.

Which of the following would be better for Jill?[29]

(A) Take the vitamin; there's a 95% chance that Jill comes to exist, and if so, she has a very happy life.

(B) Don't: there's a 95.01% chance that Jill comes to exist, but if so, she has a mediocre life.

[29] Note that the example is set up explicitly to avoid non-identity problems. But even if we think that the identity of the child is at issue, or is somehow indeterminate, in a simple case like this we can presumably ask what is better for the child *de dicto.*

I and many others think it is obvious that if we are to choose solely for Jill's sake, then we are bound to choose (A). Indeed, the choice is not completely unrealistic. We routinely make early-pregnancy interventions like this, and, often, I think, make them for the sake of the child. Moreover, it seems natural to say that (A) is better for Jill, or *at least* (and I will come back to this in section 5) that we should prefer (A) for Jill's sake.

The comparativist has an easy time here. Since non-existence is comparable for Jill to having a mediocre life, there is no real mystery about why (A) is also comparable to (B). It is even easy to see why a comparativist would, in particular, judge (A) to be better than (B). The slight increase in the chance of existence on (B) as compared to (A) can be traded off against the very likely loss of welfare. More concretely, let's suppose that the 'very good' life has welfare 10, while the 'mediocre' life has welfare 1, and non-existence has welfare 0. Then (A) has expected utility 9.5, while (B) has expected utility 0.9501.

Intuitions are not universally in favour of comparativism, however. Besides the basic intuition behind non-comparativism in risk-free cases, there are variations on Vitamin Z like the one that follows.

Pick or Flip? Suppose you have a frozen embryo, and are considering whether to incubate it. You are sure that, if you incubate it, it will develop into a person, Jill, and have a happy life. The timing and circumstances of the incubation, the identity of the person it would create, and the quality of the life she would have, are not at issue. But you can either

(Pick) decide directly to incubate the embryo, or

(Flip) flip a coin.

Which of these would be better *for the child*?

Pick	H	T
Jill	10	10

Flip	H	T
Jill	10	Ω

Pick$^-$	H	T
Jill	6	6

Here the comparativist will say that Pick is better for Jill than Flip (continuing to assume that Ω corresponds to welfare 0). And even Pick$^-$, in which she is sure to have slightly lower welfare if she exists, must be better for Jill than Flip. A small possible loss of welfare on Heads (from 10 to 6) is outweighed by a larger possible gain of welfare on Tails (from 0 to 6). Once one is in a comparativist mind-set, this seems fairly reasonable. But I and many others have a specific counter-intuition here, that Pick is exactly as good for Jill as Flip is, and that Pick$^-$ is worse for Jill than Flip.

Comparativism and the VoIP

With that background in mind, let me consider the implications of comparativism with regard to the VoIP. I will continue to suppose that – as the paradigmatic comparativist view – individual value satisfies expected utility theory. If we then combine comparativism with the VoIP, we obtain *critical level utilitarianism*.[30] To see this in our basic two-person example, note that having half chances of x and y is exactly as good as getting $\frac{1}{2}x + \frac{1}{2}y$ for sure. (Here I freely use the numerical representation of welfare levels given by expected utility theory.) So we find that L is better than M just in case

[30] Cf. Example 2.5.1 and Proposition 3.3.1 in *MMT*.

$\frac{1}{2}x + \frac{1}{2}y > \frac{1}{2}a + \frac{1}{2}b$, or equivalently, just in case $x + y > a + b$. In general, a welfare distribution $(x_1, \ldots, x_n)$ will be better than $(y_1, \ldots, y_n)$ just in case $x_1 + x_2 + \cdots + x_n > y_1 + y_2 + \cdots + y_n$. This is a comparison of total utility, but notice that the sum includes contributions from non-existent people, i.e. some of the summands may be the numerical value of Ω. This numerical value may not be zero; it is the so-called critical level.[31]

We can take all this as an argument for critical level utilitarianism, starting from the three basic principles and an expected-utility approach to individual value. Indeed, this argument is closely related to Broome's reconstruction of Harsanyi's aggregation theorem, which I described in section 2. Recall that the aggregation theorem officially applies only in fixed-population cases, but, given comparativism, we can treat Ω just like any other welfare level. Harsanyi's argument then assumes expected utility theory for individual and overall value, the Strong Pareto principle, and Anonymity. Our argument here depends on much weaker principles: expected utility theory for individual value only, and Principles 1–3.[32] However, Harsanyi and Broome both

[31] In general, I take the critical level to be the point on the welfare scale above which the addition of lives contributes positively to overall value. Given a numerical representation of the welfare scale, we can interpret the critical level as a number, here the numerical value of Ω. We will only get an ordinary total utility representation if we stipulate that Ω is represented by utility value 0. Conceptually, though, what makes this theory 'critical level' rather than ordinary total utilitarianism is *not* merely that Ω may have non-zero numerical value. After all, the numerical value of Ω is open to stipulation insofar as the axioms of expected utility theory do not uniquely determine the utility function. Rather, the point is that it may have non-zero value *even if* we normalise the utility function in such a way that good lives get positive numerical values and bad ones get negative utility values. To put it another way, the critical level may not be the level of a neutral life. Having said that, it seems strange that good lives might be worse than Ω or that bad lives might be better. If one follows the comparativist in attributing individual value to non-existence, it certainly seems natural to identify that value with the value of a neutral life.

[32] In this footnote, I spell out the weakening of the premisses in a bit more detail. I have generally assumed that individual betterness is *ex post* in the sense that the value of a lottery for an individual only depends on the probability distribution over welfare levels faced by

reject comparativism, and so would consider this argument to be unsound.[33]

Variations on CLU, and the Repugnant Conclusion

Although critical level utilitarianism is the paradigmatic comparativist view, interesting variations are available. They are interesting particularly as responses to the Repugnant Conclusion, which I consider below.

One of the advantages of our approach over that of Broome and Harsanyi is that the VoIP itself does not rely on any element of expected utility theory. All it relies on are the three principles explained in section 3. The VoIP tells us how to aggregate welfare whatever the correct story about individual value might be. Since expected utility theory, in the cast of von Neumann and Morgenstern, relies on three main axioms, this opens up three main directions for generalising critical level utilitarianism.

The three main axioms are *completeness*, *continuity*, and *strong independence*.

that individual. Similarly say that the overall betterness relation is *ex post* if the evaluation of a lottery only depends on the probability distribution over social welfare levels, where a *social welfare level* is an equivalence class of welfare distributions under the relation of overall indifference. Expected utility theory, as well as just about every concrete alternative to expected utility theory, requires that the ordering in question is *ex post*. Now, Principle 1 is much weaker than the conjunction of Anonymity and the overall *ex post* condition; Principle 2 is much weaker than the first part of Strong Pareto and the individual *ex post* condition; and Principle 3 is much weaker than the conjunction of Strong Pareto and the Completeness axiom of EUT for overall value. More conceptually, in the fixed-population case, Principles 1–3 are compatible with any individual betterness relation that satisfies the *ex post* condition, and they are also compatible with any overall betterness relation on welfare distributions that satisfies Anonymity. Further comments in relation to Harsanyi can be found in *MMT* (§5.4 and *passim*).

[33] In the terminology I will adopt in section 5, Broome seems to be an *impersonal mixed non-comparativist*, who is thereby led to a version of critical-level utilitarianism. Harsanyi was closest to being a *conditionalist* (as in section 5 below), and was thereby led to average utilitarianism, although it is doubtful that he would have been happy with the kind of average utilitarianism that follows from conditionalism and the VoIP. I consider Broome's own way of extending the aggregation theorem to variable populations in section 5.

Completeness. For any prospects x, y, x is at least as good as y or y is at least as good as x (or both).

Continuity. For any prospects x, y, z such that x is at least as good as y and y is at least as good as z, there is a probability α such that the mixture $\alpha x + (1-\alpha)z$ is exactly as good as y.

Strong Independence. For any prospects x, y, z, and any probability $\alpha \in (0, 1)$, x is at least as good as y if and only if $\alpha x + (1-\alpha)z$ is at least as good as $\alpha y + (1-\alpha)z$.

Dropping the first two axioms still retains the 'expected utility' flavour. Formally, there is a utility function with values in a 'preordered vector space' such that prospects are ranked by expected utility. Correspondingly, the aggregate value of a welfare distribution will be given by total utility, understood in this way.[34]

A more serious deviation from expected utility theory would deny Strong Independence (SI). Violations of SI are commonplace in people's actual preferences, as the Allais paradox shows. But some have suggested that violations are indeed rationally permissible (e.g. Buchak (2013)). The rule Pessimism from section 2 would be an extreme case. In principle, permitting violations of SI is entirely compatible with the VoIP. However, these non-expected utility theories may seem to pose a problem. The reason is that they

[34] This is strictly true so long as we are considering finitely supported prospects; using more general probability measures introduces technical complications, since the notion of expected value will not be so easily defined. See *MMT* §3.4 for the most general results along these lines. §3.3–§3.5 of that work give an in-depth discussion of violations of completeness and continuity in the context of VoIP, while §4 and Example 1.5.4 are devoted to violations of SI.

tend to be ecumenical, allowing for a wide variety of permissible preferences, corresponding to a wide variety of ways of violating SI. They also usually allow that *satisfying* SI is permissible. It is therefore not clear that the VoIP can presuppose a single way of comparing prospects for 'an individual'. Alternatively, in Principle 3, what I called 'perfect unanimity' may be far from perfect, since the same prospect may have different values for different individuals.

I have two responses to this line of thought. First, let us accept for the sake of argument that rational preferences need not satisfy SI. Still, this may not tell us much about the structure of prudential value. For one thing, it depends what 'rational' means. One way in which preferences can be 'rational' is that they are internally coherent, or in some sense 'interpretable'. I think it is quite plausible that preferences can violate SI without being rational in this minimal sense. For example, Buchak (2013) explains how to interpret certain violations of SI in terms of a specific kind of risk aversion, and perhaps that is enough. But this tells us very little about axiology, since the criterion for this weak rationality is purely internal. On the other hand, there is a thicker notion of rationality according to which rational preferences match up with external axiological facts – at least at a first pass, one prefers X to Y if and only if X is better than Y. If preferences that are rational in *this* sense can violate SI, that might tell us something about axiology. But it is a much stronger claim that they can.

Second (and I rehearsed this response in section 2), even with the thicker notion of rationality, if there are many rationally permissible sets of preferences, we need not conclude that the axiology of prospects must be relativised to particular individuals. It more plausibly means that axiology does not determine a unique system of rational preferences. So, overall, the denial of SI as a requirement on rational preferences does not pose any problem for the VoIP.

While some people object to comparativism *per se*, total utilitarianism – the simplest comparativist view – is often criticised because it leads to the *Repugnant Conclusion*. Let me conclude this discussion by explaining how the above considerations relate to the usual discussions in population ethics.[35]

Recall that the Repugnant Conclusion claims that even a large population of blissful lives would be worse than some sufficiently large population of lives 'barely worth living'. We can use the VoIP to translate this into a claim about individual value:

Repugnance. A sufficiently small chance of a blissful life (and non-existence otherwise) is worse for an individual than a life that is barely worth living.

Critical level utilitarians can avoid Repugnance in the following way. They can claim that the critical level is above the level of lives that are barely worth living. This ensures that a prospect that hedges between a blissful life and non-existence is always better than a life that is barely worth living. However, it smacks of inconsistency to hold that a life that is 'worth living' is nonetheless worse than non-existence.[36]

[35] We give a related discussion in *MMT* (§2.6).

[36] A version of CLU that I consider in section 5 escapes this apparent inconsistency.

Denying either of Completeness or Continuity gives another way to avoid Repugnance. By denying Completeness, we can hold that having a small chance of a blissful life, and non-existence otherwise, is incomparable to having a life that is barely worth living. If we still maintain Continuity, this requires that lives barely worth living are incomparable to non-existence. At the level of population ethics, the 'incomplete critical-level utilitarianism' of Blackorby et al. (2005) is a version of this view. This picture is different in principle from what I have called non-comparativism, since we could still hold that the 'blissful' lives are better than non-existence. Moreover, the *kind* of incomparability is arguably different: on the present view, comparing blissful lives with non-existence does not involve a category mistake.[37]

Alternatively, if we accept Completeness but deny Continuity, we can hold that a life that is barely worth living is better than non-existence, but still an arbitrarily small chance of a blissful life (and non-existence otherwise) is better than a life that is barely worth living. At the population level, larger and larger populations of of lives barely worth living are better and better, but never better than any population of blissful lives. This is a 'lexical' utilitarian view like the one described in chapter I of this thesis.

Denying Strong Independence does not particularly help to avoid Repugnance. Most plausible violations of Strong Independence involve *risk aversion*. Compared to SI, they under-emphasise the value of a chance of blissful life, and over-emphasise the value of a chance of non-existence (the worst available outcome). Taken to an extreme, they approximate Pessimism. Thus they

[37] This sort of consideration leads Chang to distinguish parity ('incomparability' in my loose sense) from incomparability properly speaking. See especially the chaining argument (Chang, 2002, p. 673).

tend to even more strongly favour certain but mediocre existence over a small chance of blissful existence. It is true that (on the other hand) risk-loving violations of SI may mitigate Repugnance, by emphasising the posibility of a good outcome and de-emphasising the possibility of non-existence. However, since in Repugnance the chance of a blissful life is arbitrarily small, we can only avoid Repugnance altogether if we take on the extreme version of a risk-loving view, to the effect that a prospect is just as good as its best possible outcome.

5 Non-Comparativism

It is much less obvious, even in outline, how non-comparativists should answer the Risky Existential Question. Here I will consider three figures. First, I will consider what I call the *strong non-comparativist*, who denies the intuitive result in Vitamin Z. Then I consider the *conditionalist*, who endorses the intuitive result in Vitamin Z and the (more controversially)intuitive result in Pick or Flip. Finally, I consider non-comparativists (including strong non-comparativists and conditionalists) for whom the Risky Existential Question is of diminished importance: they claim that Certain Existence is better than Certain Non-Existence even though it is not better for Jill. These I call 'mixed' non-comparativists.

Strong non-comparativism

If we accept that non-existence is incomparable to any state of existence, then it is quite natural to suppose that a prospect that contains some chance of non-existence must be incomparable to any state of certain existence, and more generally that two prospects are incomparable if they contain different chances of non-existence.[38] For example, consider Pick or Flip. The outcomes on Heads are exactly as good for Jill on Heads. But then, by Strong Independence, the value relation between the two prospects is the same as the value relation between the two outcomes on Tails. According to pure non-comparativists, 10 is incomparable to Ω, and therefore Pick is incomparable to Flip. As this reasoning suggests, the present *strong non-comparativist* view is compatible with Strong Independence and hence with the generalisation of expected utility theory obtained by rejecting Completeness.

This looks initially like a principled way to go, but I find it troubling, because it results in *very* widespread incomparability. At the level of prospects, it results in incomparability where the Risky Existential Question has intuitively positive answers (for example, in Vitamin Z). At the level of lotteries, two lotteries come out to be incomparable whenever they differ, however slightly, in expected population size.Axiology is impotent to guide action, if incomparability is widespread in either, let alone both, of these ways.

The problem may be even worse than just mentioned. I have assumed in

[38] Given the VoIP, this more general claim follows from the preceding one, at least when we are dealing with rational probabilities. This is because individual betterness must satisfy a condition we call *Omega Independence* (see *MMT* §2.2). According to this condition, the value relation between any two prospects P and Q does not change if we mix both of them with the same rational probability of non-existence.

this whole discussion that we can speak of the individual betterness relation as a relation on prospects, where a prospect is a probability measure over welfare levels, including Ω. Call this view *Prospectism* (a term which is used in a different but related way by Hare (2010)). A non-comparativist may have good reason to reject Prospectism. Consider the following example.

State Permutation. A coin will be flipped, and Jill will exist on one of the outcomes but not the other. However, you can determine ahead of time on *which* of the outcomes she will exist, and what her welfare level will be. Which of the following alternatives is better for Jill?

On-H	H	T
Jill	6	Ω

On-H$_+$	H	T
Jill	10	Ω

On-T	H	T
Jill	Ω	10

I think that everyone should accept that On-H$_+$ is better (for Jill and overall) than On-H. Pre-theoretically, I also find it hard to believe that On-T is not just as good as On-H$_+$, and better than On-H. But strong non-comparativists might be led to think otherwise. After all, On-T is not better for Jill on Heads, nor is it better for Jill on Tails. We can be certain that the outcome of On-T will not be better for Jill than the outcome of On-T. So it is hard to see how On-T could be better for Jill than On-H. The difficulty here is particularly one for *strong* non-comparativists. They are motivated by state-wise reasoning to say that to say that Pick is incomparable to Flip, but similar state-wise reasoning suggests that On-T is incomparable to On-H.

This line of thought has been levied against the use of expected utility theory whenever there is the possibility of incomparability (Temkin, 2012;

Schoenfield, 2014; Bales et al., 2014). On the face of it, we can replace Ω in the above reasoning by any outcome x that is incomparable to both 6 and 10. However, I am using the word 'incomparable' in a broad way, and for some sorts of incomparability, there are plausible ways of explaining why the above state-wise reasoning is inappropriate.

Suppose, for example, that the incomparability is a matter of *weighing indeterminacy*. That is, outcome x involves certain features, and it is (broadly speaking) indeterminate how these features weigh up. On some plausible ways of weighing things, x is better than 10, while on others it is worse than 6; but there is no determinately right way to do the weighing, and that is the sense in which x is incomparable to 6 and 10. It may help to think about what this might mean for rational preferences. Let us suppose that when two things are incomparable, it is permissible to prefer either one to the other. However, that doesn't mean that it is permissible to prefer 6 to x and simultaneously prefer x to 10. For one thing, it is not rational to have such intransitive preferences. But there is another explanation that does not appeal explicitly to transitivity. The basic permission is not about preferences, but about weighing. When there is weighing indeterminacy, it is rationally permissible to weigh things up however one likes, at least within some range. But there is no way of weighing things up so that 6 comes out better than x yet x comes out better than 10. So there is no rational permission to have such intransitive preferences. Similarly, there is no way of weighing things up so that On-T comes out anything but better than On-H. So there is no rational permission to prefer On-H to On-T, and this reflects an axiological fact that On-T is better than On-H.

However, strong non-comparativism does not fit well with this kind of story. The reason Ω is incomparable to 6 and to 10 is not supposed to be that there are different factors that weigh in different directions. Non-existence doesn't even register on the scale of Jill's welfare, and there is no stand to take about *where* it registers.[39]

Thus non-comparativists who are attracted to strong non-comparativism may well go further. In denying the intuitive judgments about State Permutation, they deny Prospectism. In the end they are bound to reject the VoIP, or to think that it is incoherently stated. But their view is very difficult to maintain. First, I do not know of any worked-out, principled alternative to expected utility theory that denies Prospectism in the way strong non-comparativists would like. Second, even if there is such a theory, it is almost bound to yield enormous incomparability. On my initial version of strong non-comparativism, there was only necessarily incomparability when two situations gave different probabilities to existence. Now we should expect incomparability unless existence is guaranteed on the very same states of affairs, as seen in the incomparability of On-H and On-T. Such a theory of individual value would not appear to have much use for population ethics, even if it were true.[40]

[39] Rabinowicz (MS) gives a structurally similar defense of Prospectism, using fitting attitude theory. In related work, he uses fitting attitudes to distinguish parity from incomparability-strictly-speaking (Rabinowicz, 2012). He proposes that a variety of preferential attitudes are fitting in cases of parity, whereas incomparability requires a 'preferential gap'.

[40] As something close to an example, consider the 'deferentialism' sketched by Hare (2010). Hare presents deferentialism as providing a link between incomplete preferences and choice. But we can reinterpret it as providing a link between incomplete value and choice (and I have an easier time comprehending the theory in these terms). In the current setting, deferentialism implies that either of two prospects can be permissibly chosen for the sake of the individual if there is some state of nature of non-zero probability in which

In the next two subsections, I consider two moves open to non-comparativists who are bothered by widespread incomparability. They can give up the state-wise reasoning that leads to strong non-comparativism. Or they can deny Principle 3, thus denying the importance of non-comparativism to overall evaluation. (In the terms I introduced at the beginning of section 3, this is one way of denying that overall betterness is a matter of *personal* betterness.)

Conditionalism

Here is an initially attractive alternative to strong non-comparativism. Rank prospects by their value *conditional on* existence – that is, rule out non-existence and rescale the probabilities accordingly.[41] Call this rule *conditionalism*. It gives what I claimed were the intuitively correct answers in Pick or Flip and Vitamin Z, and also in State Permutation. That is a strong point in its favour.

Before analysing conditionalism, let me note a variation on it which may now spring to mind. The alternative is to rank prospects by their value conditional on existence *and then weighted* by the probability of existence. Concretely, we might take expected utility conditional on existence, and then multiplied by the probability. For example, the value of Flip for Jill would be the utility 10 she gets conditional on existence, times the probability

the individual exists on one option but not the other. This widespread permissibility is analogous to the widespread incomparability described in the main text, and problematic to the same extent.

[41] Views of this sort are considered by Harsanyi in his correspondence quoted in Ng (1983), and by Voorhoeve and Fleurbaey (2016).

1/2 of existence; hence, Flip would have value 5. This 'weighted' version of conditionalism is in most cases formally the same as supposing that Ω has utility value 0, and then calculating expected utility. The one case in which it differs is when Ω has probability 1, and so the operation of conditionalising on existence does not make sense. Weighted conditionalism can thus coherently maintain that Ω itself is incomparable to any outcome on which the individual exists. I take it the attraction of weighted conditionalism is that it gives a plausible-sounding way to reconstruct the verdicts of comparativism (and thence total utilitarianism) without actually giving Ω a utility value.

However, I think weighted conditionalism is not as attractive as it initially seems. Compare once more Flip and Pick$^-$.

Flip	H	T
Jill	10	Ω

Pick$^-$	H	T
Jill	6	6

According to plain conditionalism, Pick$^-$ is worse for Jill than Flip, and this seems about right. According to Weighted conditionalism, however, Pick$^-$ is better for Jill than Flip, having value 6 instead of $10 \times 1/2 = 5$. Not only is this, I think, counterintuitive, but, more importantly, the weighted conditionalist cannot justify this judgment using expected utility theory. Observe that the outcome on Heads is much worse for Jill under Pick$^-$ than under Flip. That seems to be one respect in which Pick$^-$ is worse than Flip. If the outcome on Tails were much *better* for Jill under Pick$^-$ than under Flip, that would be a countervailing respect in which Pick$^-$ would be better than Flip. So we might reason: yes, on Heads, Flip is better for Jill, but, on Tails, Pick$^-$ is better to an even greater degree; thus Pick$^-$ is better overall.

The weighted conditionalist *cannot* argue this way, because he maintains that 6 is not better than Ω. But then it is simply not clear to me *why* Pick$^-$ is better for Jill than Flip.

Now let me return to the basic, unweighted version of conditionalism. The paradigmatic version of this theory is the one that evaluates prospects by *expected utility* conditional on existence. Given VoIP, this corresponds to evaluating welfare distributions in the manner of Vickrey: evaluate as if for the sake of an individual whose identity is uncertain, but who is certain to exist. The paradigmatic moral theory is therefore average utilitarianism.[42] In general, conditionalists will avoid Repugnance, just as average utilitarians avoid the Repugnant Conclusion.

Beyond that, however, average utilitarianism faces very serious problems (see e.g. Hurka (1982a,b)). The main point I wish to make is that this produces a dilemma. On the one hand, the three principles leading to the VoIP seem plausible, as does conditionalism. But average utilitarianism seems unacceptible. What should we give up? (And *how*?)

In negotiating that dilemma, it may be worth pointing out that the kind of average utilitarianism recommended by the VoIP has some curious features, which may make it more acceptable than the ordinary kind. The best-known problems for average utilitarianism involve situations in which the outcome is certain. But further problems become clear when we ask what average utilitarians should say about chancy situations. Consider, for example, the following lotteries involving 100 people.

[42] Cf. *MMT* Example 2.5.2.

L_{100}	H	T
Ann	Ω	-20
Bob$_1$	10	Ω
Bob$_2$	10	Ω
......	...	...
Bob$_{99}$	10	Ω

M_{100}	H	T
Ann	Ω	10
Bob$_1$	-20	Ω
Bob$_2$	-20	Ω
......	...	...
Bob$_{99}$	-20	Ω

The most obvious way for an average utilitarian to compare these two lotteries is by *expected average utility*.[43] By this criterion, L_{100} and M_{100} are equally good. This judgment is extremely implausible, since M_{100} is worse (on any account I have considered) for all the Bobs, and better only for Ann.

One might instead think to rank lotteries by *average expected utility*, where the expectations are calculated conditional on existence. This yields the more plausible result that L_{100} is better than M_{100}. On the other hand, consider the following two lotteries, in which H, T$_1$,...,T$_{99}$ are 100 equiprobable outcomes:

L'_{100}	H	T$_1$	$\cdots$	T$_{99}$
Ann	Ω	10	$\cdots$	10
Bob	-20	Ω	$\cdots$	Ω

M'_{100}	H	T$_1$	$\cdots$	T$_{99}$
Ann	Ω	-20	$\cdots$	-20
Bob	10	Ω	$\cdots$	Ω

These lotteries have the same average expected utility, despite the fact that L'_{100} is almost certain to create a world with a single well-off person and M'_{100} is almost certain to create a world with a single badly-off person. This is again implausible.

[43] Since average utility does not make sense for empty worlds (i.e. worlds in which no one exists), it is not clear what 'expected average utility' means when there is a chance that no one exists. But let us suppose that this is dealt with in some way, e.g. by stipulating a value for empty worlds, or by conditionalising on the proposition that the world is not empty.

The form of average utilitarianism recommended by conditionalism and the VoIP does *not* rank lotteries by expected average utility, nor by average expected utility. The former of these theories violates Principle 2, and the latter violates Principle 1. Instead, we obtain a theory that ranks lotteries by *expected total utility divided by expected population size.* Call this theory *veiled average utilitarianism* (VAU). VAU vindicates the intuitive judgments in the just-considered cases; it ranks L_{100} above M_{100} and L'_{100} above M'_{100}.

Now, one entirely reasonable reaction to all this is that the failure of expected utility theory (and especially the failure of Strong Independence) counts strongly against VAU. But there are two counterbalancing considerations. First, the VoIP, and the three principles justifying it, articulate a way in which overall value reduces to individual value. Insofar as we have a non-EUT theory for individual value (e.g. conditionalism), then of course we must have a non-EUT theory for overall value. Moreover, insofar as the non-EUT theory for individual value is a principled one, we should count the theory for overall value as principled as well. This is just to say that the committed non-comparativist should be *ready and willing* to bite some bullets. The second point is that the problems with expected average value are severe enough that someone generally sympathetic to average utilitarianism may prefer veiled average utilitarianism overall. One way in which VAU's violation of SI might not be *so* bad is that it still licenses a kind of case-by-case reasoning similar to that given by expected utility theory. While a value function on lotteries that satisfies EUT satisfies

$$V(\alpha L + (1-\alpha)M) = \alpha V(L) + (1-\alpha)V(M)$$

the value function of VAU satisfies

$$V(\alpha L + (1-\alpha)M) = \frac{1}{N_L + N_M} \alpha N_L V(L) + (1-\alpha) N_M V(M),$$

where N_L and N_M are the expected number of people under L and M. Thus the value of a mixture of lotteries can be easily calculated in terms of some simple statistics of those lotteries. This is not true of standard non-expected utility theories.

It must nonetheless be clear that these observations about veiled average utilitarianism have more the character of consolation than of defence.

Mixed Non-Comparativists

Consider again the example Certainty (p. 131). This is a case of perfect unanimity, so Principle 3 may apply. According to that principle, Certain Existence is better overall than Certain Non-Existence just in case it is better for Jill. Non-comparativists deny that it is better for Jill. The non-comparativists I have considered so far accept Principle 3, at least in this case, and therefore that Certain Existence is not better overall than Certain Non-Existence. But what I call *mixed* (as opposed to *pure*) non-comparativists allow that Certain Existence is better than Certain Non-Existence overall. Thus they reject Principle 3, and must reject the VoIP as written. However, I will argue that these mixed non-comparativists can accept a modification of Principle 3 and of the VoIP, and, with respect to this modification, they fall into the 'comparativist' camp. In particular, we obtain a non-comparativist version of the argument for critical level utilitarianism.

Two Varieties of Mixed Non-Comparativism

To get a handle on the mixed non-comparativist view, let me revisit the dominance argument for Principle 3. I mentioned in footnote 24 that the argument contains a loophole. What it shows – convincingly, in my view – is that, in cases of perfect unanimity, L_2 is at least as good as M_2 *if* it is at least as good for each individual. The converse amounts to the claim that L_2 is incomparable to M_2 if it is incomparable for each individual. But this is less compelling, and anyway is not supported by the dominance argument. There could be something else besides individual betterness that decides between L_2 and M_2 when individual betterness is silent, as it is when we compare Certain Existence and Certain Non-Existence.

With that in mind, we can distinguish two kinds of mixed non-comparativists. Some, whom I will call 'personal', think that we should favour Certain Existence over Certain Non-Existence *for Jill's sake*. For example, they may think that the former is better than the latter because Jill would be 'non-comparatively benefited' by having a very good life.[44] The fact that Certain Existence is *good* for Jill counts in its favour, and there is no Jill-related fact that counts for or against Certain Non-Existence. It is this goodness for Jill that is decisive even when there is no question of *betterness* for Jill.

Other mixed non-comparativists, whom I will call 'impersonal', think that we should favour Certain Existence over Certain Non-Existence overall,

[44] See Bykvist (2007) for a discussion of non-comparative benefits. He is officially neutral about what the import of non-comparative benefits are, but expresses what I take to be a personal mixed non-comparativist view in his discussion of guardian angels (pp. 353–356).

but, seeing as how the former isn't better for Jill, we cannot do so strictly for Jill's sake. What is decisive here is the *impersonal* value of the life Jill would have on Certain Existence, perhaps correlated with but not simply a matter of that life being good for her. (See the discussion of 'neutral' wellbeing in Broome (2004, p. 142) for a clear example of this view.)

In case the difference between these views appears obscure, here are two ways to pull them apart a little. First, the personal version begins with a distinction between lives that are good and those that are bad for the people living them. But impersonal mixed non-comparativists need not be committed to any such distinction. Fundamentally at issue for them is whether Jill's life in Certain Existence has the right character to make it better *overall* that it be lived than not lived, but this need not be a matter of the life being good rather than bad for Jill. Second, consider the combination of mixed non-comparativism with conditionalism. Since mixed non-comparativists accept that Certain Existence is better overall than Certain Non-Existence, they are very likely to think that Pick is better overall than Flip (p. 134). But conditionalists think that Pick and Flip are equally good for Jill. It is then not true that we should prefer Pick for Jill's sake: as far as she goes, we should be indifferent. Thus conditionalism is most compatible with the impersonal version of mixed non-comparativism. On the other hand, this combination is not very attractive, since it falls prey to the dominance argument used to support Principle 3.

Recall that in section 3 I proposed understanding the three principles, and hence the VoIP, as characterising a kind of 'personal' (but not merely individual) value. From that point of view, mixed non-comparativists can endorse Strong Non-Comparativism and VoIP as stories about personal value, but hold that when it comes to overall moral evaluation, there is a further story to be told. What distinguishes personal mixed non-comparativists is that they still try to tell that further story in terms of what we ought to do for the sake of individuals, for example by appealing to categorical rather than comparative judgments about the personal value of individual lives.[45]

Mixed Non-Comparativism and Critical Level Utilitarianism

Having got the different views on the table, let me now explain how mixed non-comparativists can accept a version of the VoIP, and are thereby pushed towards critical level utilitarianism.

Consider first the personal mixed non-comparativists. They recognise a Jill-regarding, not merely overall sense in which Certain Existence is more choice-worthy than Certain Non-Existence, even if they are reluctant to call this 'betterness for Jill'. For example, Bykvist (2007, p. 355) recognises a 'preventive' value in preventing someone from existing if they would have a bad life, and this value is relevant to deliberations for the sake of that possible person. One can therefore formulate new versions of Principle 3 and the VoIP that invoke not value for an individual but choice-worthiness regarding that individual. These modified versions should be as acceptable to personal mixed non-comparativists as the original versions were acceptable to comparativists.

[45] I should note that while personal mixed non-comparativism can be detected in the work of Bykvist and others, I do not know of anyone who has unambiguously and systematically endorsed such a view. Still, it is worth seeing how the view works out. In contrast, Broome is unambiguously an impersonal mixed non-comparativist.

Before, the VoIP provided an argument from comparativism to critical level utilitarianism, in which the utilities being summed represented betterness for individuals. Now we obtain an argument from personal mixed non-comparativism to a form of critical level utilitarianism in which the utilities being summed represent choice-worthiness regarding individuals. However, betterness for Jill coincides with choice-worthiness regarding Jill as long as Jill is certain to exist. So the numerical representation of choice-worthiness regarding Jill is also a numerical representation of individual value for Jill, excepting only that it assigns a numerical value to her non-existence. In this way, we can interpret the critical level utilitarianism on offer as aggregating individual value. This is so similar to the comparativist picture that the only mystery is why the personal mixed non-comparativist declines to use the phrase 'better for' to mean 'more choice-worthy regarding'.

As for impersonal mixed non-comparativists, there are a few moves available. Perhaps the simplest is to consider *Robinson Crusoe lotteries*: lotteries in which only one person has any chance to exist. Call him 'Crusoe' (sorry, Friday). Given any lottery L, we can consider the corresponding Crusoe lottery L_C in which Crusoe faces the same prospect that Jill faces in L. We can say that L is 'Crusoe-better for Jill' than M if and only if L_C is better overall than M_C. We can more generally talk about the 'Crusoe value' of a lottery or prospect for Jill. So we again obtain a sense of Jill-regarding choice-worthiness, to which now impersonal mixed non-comparativists can subscribe.[46]

[46] In Bader's terms, Crusoe value is a version of *individual but impersonal* value, just as there may be *overall* (or in his terminology, 'general') *but personal* value.

At this point, the discussion runs parallel to the discussion of personal mixed non-comparativism. First, we can formulate modified versions of Principle 3 and the VoIP that invoke Crusoe value instead of individual value. From this Crusoe version of VoIP, and the hypothesis that Crusoe value satisfies expected utility theory, we obtain a version of critical level utilitarianism in which the utilities being summed represent Crusoe value rather than individual value.[47] Moreover, it seems that Crusoe value must coincide with individual value when the individual is certain to exist.[48] So, again, the numerical representation of Crusoe value merely extends the numerical representation of individual value by giving a numerical value to non-existence. Once more, the critical level utilitarianism on offer has a direct interpretation as aggregating individual value.

While I suggested that the personal mixed non-comparativist's view was but a minor variant of comparativism, the impersonal version of the view does have an important distinguishing feature. The impersonal mixed non-comparativist's version of the VoIP no longer involves a reduction of overall value to individual value.[49] From one point of view, according to which it must be personal value that matters morally, the absence of this reduction appears mysterious. On the other hand, it provides an advantage in respond-

[47] Note that if overall value satisfies expected utility theory, then so does Crusoe value, since it is just a matter of the overall value of one-person lotteries.

[48] For example, we get this result from the original version of Principle 3, applied to Robinson Crusoe lotteries in which Robinson is certain to exist. Alternatively, we can derive it from Strong Pareto plus the assumption that individual value completely preorders such lotteries.

[49] Remember why: on the personal version of the view, the value of the critical level can be explained in terms of choice for the sake of particular individuals, or in terms of categorical rather than comparative judgments of individual value. On the impersonal version of the view, the critical level is just a feature of overall value, not to be explained in terms of value for individuals.

ing to the Repugnant Conclusion. In my discussion of comparativism, I complained that it smacks of inconsistency to hold that a life that is 'worth living' is worse than non-existence when it comes to individual value. On the current impersonal view, to say that some individually good lives fall below the critical level only commits one to the view that these lives contribute negatively to *overall* value – a strange view, but not an inconsistent one.

John Broome's position in *Weighing Lives* is naturally interpreted as an impersonal mixed non-comparativist view. Let me conclude this discussion of non-comparativism by relating what he does to the picture just described. The upshot is that our argument for the VoIP, adapted to impersonal mixed non-comparativism, yields a streamlined version of Broome's argument for critical-level utilitarianism.

Broome uses his version of Harsanyi's theorem to derive fixed-population utilitarianism. To extend the result to variable populations without accepting comparativism, Broome appeals to a separability condition. To state a version of this, suppose that we divide all possible people into two groups, $\mathscr{P}$ and $\mathscr{Q}$. For any lottery L, we can consider the 'restriction' $L|_{\mathscr{P}}$ of L to $\mathscr{P}$. This is a lottery in which no one outside $\mathscr{P}$ can exist, and everyone in $\mathscr{P}$ faces the same prospect as in L.

Separability. Suppose that $L|_{\mathscr{P}}$ is just as good as $M|_{\mathscr{P}}$. Then L is at least as good as M just in case $L|_{\mathscr{Q}}$ is at least as good as $M|_{\mathscr{Q}}$.

There are several ways of motivating such a principle. But the main idea is the obvious one, that since the $\mathscr{P}$ people are unaffected by the choice between L and M, we should be able to ignore them. This thought is especially compelling when the $\mathscr{P}$ people are located far away, or in the distant past. But if we do not think that spatiotemporal location is morally relevant, we should accept Separability in general.[50]

How does Broome use Separability to argue for critical-level utilitarianism? Here is a reconstruction, in slightly different terms from the ones he uses. First, Separability implies a modified version of Strong Pareto that involves Crusoe value instead of individual value.[51] Suppose that Anonymity holds and that overall value (hence Crusoe value) satisfies expected utility theory. Then we have the main premisses for Harsanyi's theorem, in terms of Crusoe value instead of individual value. We thereby obtain a version of critical level utilitarianism in which the utility values to be summed represent Crusoe value. But we can again reinterpret this critical level utilitarianism as aggregating individual value, as in footnote 48.

How does this compare to my suggestion above, that the impersonal mixed non-comparativist like Broome can adopt the Crusoe version of the VoIP? First of all, Broome's assumptions of Anonymity and expected utility theory for overall value entail Principle 1, and Separability entails Principle 2. Separability and Completeness for overall value together imply the Crusoe version of Principle 3. Thus a subset of Broome's crucial premisses can be used to justify first the Crusoe version of the VoIP, and thence critical level utilitarianism. On the other hand, one might accept the relevant versions of Principles 1–3 without appealing to the full force of Separability.[52] To echo

[50] A more extended form of this argument is particularly clear in Blackorby et al. (1995), in which the authors appeal to 'independence of the utilities of the dead'.

[51] Specifically: if L is at least as Crusoe-good as M for every individual, it is at least as good overall; if, in addition, it is Crusoe-better for some individual, then it is better overall.

[52] We give a complementary discussion of Separability in §3.2 of *MMT*. Roughly, we

my remarks in section 2, the veil of ignorance turns out to be an efficient way of understanding Broome's argument for utilitarianism, rather than a misguided alternative.

6 Conclusion

My first goal in this paper was to rehabilitate the veil of ignorance by laying out more basic normative principles to which the VoIP is equivalent. These principles reflect an ideal of impartial but personal value, to which ideals of fairness, equality, or priority could in principle be adjoined. The VoIP forcefully raises the Risky Existential Question, in answer to which my second goal has been to provide a rough map of the terrain. Comparativists and the related mixed non-comparativists find here an efficient argument for critical level utilitarianism, greatly simplifying the work of Harsanyi and Broome; conditionalists find a positive argument for average utilitarianism; but pure strong non-comparativists find a theory of overall value laid waste by incomparability.

show that Separability is equivalent to Strong Independence, given the VoIP. This is a powerful observation which, however, I do not exploit here.

List of Tables for Chapter IV

IV | Time-Relative Interests and Population Ethics

I was of three minds,
Like a tree
In which there are three blackbirds.
—Wallace Stevens

ABSTRACT. Discussions in population ethics most often concern distributions of lifetime welfare among *persons*. On the other hand, Parfit's work has made popular the idea that what is normatively relevant is not personal identity but psychological connectedness (and/or continuity). Moreover, it is often thought that psychological connectedness, unlike personal identity, comes in degrees. How can we reimagine population ethics in light of these ideas? That is the broad question I consider here, but it is motivated by a narrower one: how best to systematize and understand certain ethical verdicts associated with Jeff McMahan's theory of time-relative interests.

1 Introduction

Jeff McMahan's *time-relative interests* or *TRI* account (McMahan, 2002) is most often considered to be an account of the badness of death. But in the first instance it is a general view about prudential value. The account is naturally formulated in terms of *person-stages*, or persons at times. I take it that each person-stage has a welfare level, reflecting how well things are going at the time in question. But the prudential deliberations of one person-stage typically take into account the welfare of some *other* person-stages, in the quotidian sense that I now take into account my happiness tomorrow. In these terms, the TRI account answers the following question. Consider person-stages S and T. What weight should considerations about T's welfare have in S's prudential deliberations, or in deliberations on behalf of S? As I will put it: what stake does S have in T's welfare?

The most obvious answer to this question is that S has a full stake in T's welfare just in case T is the same person as S, and none otherwise. What matters is personal identity, interpreted as a relation between person-stages. In contrast, the TRI account claims that S's stake in T's welfare tracks the degree to which S and T are psychologically connected. What matters is psychological connectedness, and this is a matter of degree.[1] So, for example, if a newborn baby is only loosely psychologically connected from one week to the next, then, at any given time, it has only a small stake in its future wellbeing. In contrast, a typical adult has strong psychological connections from one decade to the next, and therefore a big stake in her future wellbeing. Thus a human can at one time, as a newborn, have a weak interest in continuing to live, and a strong interest at another.

The TRI account of prudential value is also supposed to have certain

[1] In formulating the account, McMahan (2002, p. 80) writes generically in terms of 'prudential unity relations', which include psychological as well as potentially other kinds of relations. (He of course develops a view of what these are.) The archetype is Parfit's '*Relation R*: psychological connectedness and/or continuity, with the right kind of cause' (Parfit, 1986, p. 215). I will stick to psychological connectedness for the sake of brevity and concreteness; one can easily adapt the discussion to a more complicated account of the prudential unity relations.

implications for ethics, which I will refer to as *the orthodox verdicts*. For example, suppose we face the choice between saving the life of the newborn or that of the adult. Letting the newborn die would preclude (let us suppose) 60 years of good life; letting the adult die would preclude only 40 years of good life. This may make it seem that we should save the newborn and let the adult die, thus maximizing the number of years of good life that are lived. But these are only the headline figures. The adult has a big stake in his future wellbeing; he would really be *deprived* of those 40 years. The newborn has very little stake in its future wellbeing. Compared to the adult, it would hardly be deprived by death. We ought to save the adult.[2]

Although this story (which I will introduce more fully below) has a certain plausibility, it faces an array of problem-cases (several of them introduced by McMahan himself). These by and large do not raise problems for the TRI account narrowly construed as an account of prudential value. But together they raise doubts about whether this account of prudential value can be incorporated into a coherent moral theory that vindicates the orthodox verdicts. One aim of this paper is to allay those doubts. A bit more generally, McMahan has explicated the TRI account largely case-by-case, leading to a wonderfully nuanced but somewhat indeterminate picture. We are to be 'guided' by time-relative interests, but what does that mean in general?[3] I

[2] I note that there are two plausible ways of talking about deprivation here. One might say that the newborn faces a *greater* deprivation (60 years of welfare instead of 40), but the deprivation *matters less* to him. Or one might say that a loss is only a deprivation *to the extent that it matters* to the person so deprived at the time of evaluation. Then the adult faces greater deprivation. I prefer this second way of talking, but I admit it is not completely standard.

[3] Note that the big picture may well involve considerations that go beyond the broadly welfarist focus of the TRI account. But in this paper I will focus purely on whatever guidance can be had from considerations of welfare and psychological connectedness.

will sketch some relatively systematic ways of understanding the view, and indeed go to the opposite extreme of presenting a simplistic but anyway specific decision rule. I do not claim that this is the best possible way of capturing the orthodox verdicts – though I have not seen better. But it is enough to see off the basic worry of incoherence, and, I hope, to move the literature beyond examples and putative counter-examples back towards a more fundamental mode of theorising.

After describing the TRI account and the problem cases more fully in sections 2 and 3, I develop the basic strategy in section 4. The rough idea is to think of psychological connectedness as a degreed version of personal identity. The claim is that the problems of the TRI account then become simple generalisations of certain well-known problems in population ethics. Although I develop this point more systematically than it has been heretofore, the basic picture is already widely known. However, I use it to motivate a novel 'population ethics first' strategy for developing the TRI account. Schematically, this involves two steps:

Step 1. Settle on a theory of population ethics.

Step 2. Adapt it to deal in degrees of psychological connectedness rather than all-or-nothing personal identity.

In contrast, the more usual strategy is to start from the notion of time-relative interests and to figure out how they should guide ethical judgments. Let me explain this a little more. What I will call 'standard' normative theories are those that deal in distributions of lifetime welfare among persons. In particular, the standard account of prudential value is the one according to which the prudential value of an outcome *A* for a person-stage *S* is the lifetime welfare in *A* of the person to whom *S* belongs. In these terms, the thought behind the usual strategy is something like this:

> On the standard account, prudential value is a matter of lifetime welfare. On the TRI account, it is a matter of time-relative interests. On standard accounts of population ethics, we should be guided by considerations about lifetime welfare. So on the TRI account, we should be guided by considerations about time-relative interests.

From my point of view, the TRI account of prudential value is just what you get when you adapt the standard account of prudential value to deal in degrees of psychological connectedness. We want to adapt standard accounts of population ethics in the same way, but there is no fundamental reason the result must be expressed directly in terms of time-relative interests. This is, again, more a point of strategy than a conceptual one. It would of course be implausible if the end result had no relationship whatsoever to the TRI account of prudential value. But we need not, and perhaps in the first instance *should* not, theorise directly about the form of that relationship.

Neither of the two steps in this programme is trivial. Indeed, finding an adequate theory of population ethics is famously intractable. But we cannot lay *that* problem at the door of the TRI account. The programme suggests that *however* one solves the problems of population ethics, essentially the same considerations will overcome the problems of the TRI account. This may be true even if the solution is to reject some of the trouble-making intuitions.

Indeed, I should emphasise that the issues raised by Step 2 are of very general interest. Many people reject the intuitions in population ethics that correspond to the orthodox verdicts of the TRI account, while remaining sympathetic to the idea that degrees of psychological connectedness are what matter. How, for example, can average utilitarianism or prioritarianism incorporate this idea?

Starting in section 5, I explore some options for carrying out this programme. Perhaps it will not be surprising that I am able to complete neither of the two steps in a fully satisfactory way. As far as the first step goes, I do present a theory of population ethics that gives the desired answers in all the most relevant cases. I explain this theory, which I call *complex necessitarianism*, at the end of section 5, and use it as an example in what follows. What about the second step? How can one adapt one's favourite theory of population ethics to deal in degrees of psychological connectedness? Here I explore two general approaches.

Standard theories of population ethics suppose that each available option gives rise to a set of people, whose lifetime welfare levels form the basis for moral evaluation. The first general approach, explored in section 6, is to reconsider which entities should count as 'people' when interpreting these theories. Insofar as this strategy succeeds, we do not really have to *adapt* our favourite theory of population ethics, so much as to *reinterpret* it in terms of the normatively relevant entities. We would thus obtain a recipe for completing Step 2 of the programme that would not depend on how we complete Step 1.

My first example of this approach posits that personal identity is sometimes *indeterminate*, and the degree of psychological connectedness determines the degree of indeterminacy. Crucially, this picture allows that personal identity is a transitive relation, thus maintaining that each outcome contains a set of persons, even if it is indeterminate which person-stages compose which people. The main problem with this picture is that the facts about personal identity do not properly supervene on the psychological facts, or at least I do not see how to ensure that they do.

A second picture, which I consider more briefly, holds that personal identity is sometimes *partial*. For two person-stages at different times to be stages of the same person is for them to have the same psychological constituents, and this overlapping is a matter of degrees. An advantage of this picture is that the sense in which identity comes in degrees is more closely tied in concept to psychological connectedness. However, it relies on a highly speculative psychological picture, which may simply be untenable.

A third picture would rely on David Lewis's view that posits overlapping four-dimensional persons. However, I will argue that these Lewisian persons are not fit subjects for population ethics, in part *because* of their overlapping nature.

The second general approach appears more promising. I take it up in section 7. There I consider a counterpart-theoretic treatment of personal identity. This 'prudential' counterpart relation is an on/off relation like personal identity, but it is not usually an equivalence relation. It therefore does not group person-stages into persons. So this approach is not directly compatible with standard normative theories. However, it is sometimes possible to adapt these theories in such a way that the prudential counterpart relation functionally replaces the personal identity relation. In the case of prudential value, we recover the TRI account, because of the way in which the prudential counterpart relation supervenes on psychological connectedness. In the case of complex necessitarianism, we obtain an ethical theory that systematically reproduces the orthodox verdicts of the TRI account. Mission accomplished!

Be that as it may, one of the advantages of the viewpoint I adopt in this paper is that it allows our thinking about the TRI account to be informed by the wide range of theoretical and intuitive considerations that arise in population ethics. To illustrate this point, I conclude in section 8 by suggesting an alternative to the orthodox verdicts, still broadly in line with McMahan's development of the view. The alternative I suggest is driven in part by theory and in part by intuition. The main theoretical issue is the justification of the so-called *asymmetry* in population ethics, which plays a big role in the preceding discussion. The point of intuition is that I think it can be permissible to create additional good lives even at some cost to people who exist independently of the choice to do so. In terms of the TRI account, this corresponds to denying that, in the initial motivating case, we *must* save the adult over the newborn; it is merely permissible to do so. Or *if* we must, that is not something to be explained by time-relative interests alone. The resulting theory amounts to a positive proposal for understanding the TRI account and its ethical implications, with the suggested emendation.

For critics of the TRI account, this is at least a fixed target. For advocates, it illustrates a methodology that holds out some hope. As serious as the problems of population ethics may be, we need not think that the TRI account adds further insuperable difficulties to an already difficult project.

One point of terminology. The TRI account is often described in axiological terms. It is, after all, an account of the *badness* of death. When it comes to prudential value, I have no qualms about this emphasis on axiology. But when it comes to the TRI account as a general normative theory, it seems to me that many of the relevant issues are not best understood axiologically. For example, suppose that one option ought morally to be chosen over another, when those are the only two options. Then I will say that the first option is *morally preferable* (or sometimes just *preferable*) to the second. The orthodox verdicts of the TRI account strongly suggest that the relation of moral preferability is intransitive. There are three options A, B, and C, such that, out of A and B, one ought morally to choose A; out of B and C, one ought morally to choose B; but out of A and C, one ought morally to choose C. This pattern cannot be the outcome of a rule to maximize value, and at least in that sense it cannot (*pace* Temkin) be a matter of axiology. For this and other reasons I will often speak in terms of what one ought to do, or what is preferable, rather than what is best.

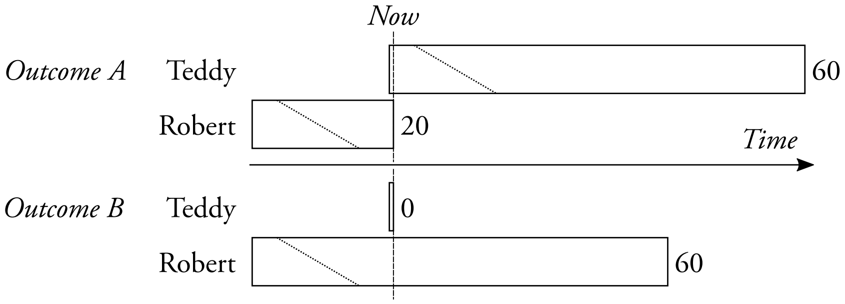

Table 1: Emergency Room. In these diagrams, time moves forward to the right. The vertical line ('Now') represents the time at which the decision is to be made. The height of each rectangle represents welfare over time; the numbers to the right are the ages at time of death. The dotted diagonal lines schematically represent psychological transitions (here between infancy and adulthood).

2 The Life-Comparative and TRI accounts

In this section, I explicate more fully the TRI account of prudential value, and give an initial discussion of why it is difficult to use it as the basis for a theory of ethics. Consider:

> **Emergency Room** (Table 1)[4]
>
> A doctor can treat only one of two patients; the other dies. In alternative *A*, the doctor treats a newborn, Teddy; in alternative *B*, the doctor treats a 20-year-old, Robert. In either case, whoever is treated goes on to live to 60, gaining welfare at a constant rate of ten units per year.

[4] See McMahan (2002, p. 185) for a similar case, 'Choice between Lives'.

One way to think about this case is that, if she saves the newborn, the doctor will grant someone 60 years of good life. If she saves the 20-year-old, she will grant only 40 years of good life. Now, 60 is more than 40. Therefore:

(ER-LC) Death would be worse for the newborn than it would be for the 20-year-old, because it would deprive him of more lifetime welfare.

(ER-LC) is the judgment of the *life-comparative* (LC) account of the badness of death. It identifies the badness of each death as the difference between the lifetime welfare of the person who dies and what their lifetime welfare would have been in the salient alternative. As I understand them, both the LC account and the TRI account are (at least in the first instance) accounts of the 'personal' or 'prudential' badness of death – how much death in the way specified makes things worse *for* the person who dies, when compared to the salient alternative. More precisely, we should understand them as accounts of how heavily the prospect of death should weigh in prudential deliberations by or on behalf of a person at a time: here, Teddy at birth or Robert at 20. From this perspective, the LC account of the badness of death is an application of what I called the 'standard' account of prudential value in section 1. It assumes that personal identity is the relation that underpins prudential interest. It also assumes that a human remains the same person more or less from conception until death.

There is a closely related judgment that the outcome A is *morally* preferable to the outcome B: the doctor ought to save the newborn. This judgment might be reached in various ways, but I wish to focus on the following two simple ones, which I will call the *lifetime-utilitarian* and the *person-affecting* explanations.

(LU) The doctor ought to save the newborn, because what matters is which outcome would, on balance, give the newborn and the 20-year-old more lifetime wellbeing.

(PA-LC) The doctor ought to save the newborn, because what matters is which outcome would be worse on balance for the newborn and the 20-year-old, and (ER-LC) is true.

These two ways of explaining the judgment are mathematically equivalent – they both boil down to the fact that 60 is more than 40 – but only the person-affecting explanation (PA-LC) depends on (ER-LC). In principle, one could accept (LU) as the right explanation of which outcome is better overall, or which outcome the doctor ought to choose, while denying (ER-LC). I will produce an example very soon.

In contradiction to the life-comparative account, many people have the intuition that death is not as bad for the newborn as it is for the 20-year-old. There is much one could say about this intuition, but there is at least one *prima facie* plausible justification for it, which is the basis of the TRI account. The idea is that the typical newborn is only weakly psychologically connected to his future self. Although he has many years of good life ahead of him, he does not have a real stake in them. If we want to know how much the newborn is harmed, we must only count the 60 years to the degree to which he has a stake in them. In contrast, the 20-year old has a full stake in his adulthood. He really would be deprived of 40 years of good life, without any discount, or perhaps with only a small discount. The question is not whether 60 is more than 40, but whether it is still more once we have applied the appropriate discounts.

It may help to see how this could work with a simplistic but concrete calculation. Suppose we divide the potential lives of these patients into three acts: infancy, from ages 0 to 3; childhood, from 3 to 15; and maturity, from 15 to 60. The welfare that accrues over these times is given in the following table:

		Age		
		0–3	3–15	15–60
Outcome A	Teddy	30	120	450
	Robert	30	120	50
Outcome B	Teddy	0		
	Robert	30	120	450

Perhaps a newborn has a full stake in the welfare of his infancy, a 1/2 stake in the welfare of his childhood, and no stake in the welfare of his maturity. Since the infant would have three good years of infancy (at 10 units of welfare a year), 12 good years of childhood, and 45 good years of maturity, his death effectively deprives him of

$$30 \times 1 + 120 \times 1/2 + 450 \times 0 = 90$$

units of wellbeing. But the 20-year old may have a full stake in the welfare accrued during the rest of his maturity. Thus death would deprive him of 400 units of wellbeing. The newborn is less harmed by death.

Thus, if we accept this calculation as representative, the TRI account yields:

(ER-TRI) Death would be worse for the 20-year-old than it would be for the newborn, because it would deprive him of more welfare (weighted by his stake in it).

One might conclude from this that doctor ought to save the 20-year-old. In particular, one might reason as follows.

(PA-TRI) The doctor ought to save the 20-year-old, because what matters is which outcome would be worse on balance for the newborn and the 20-year-old, and (ER-TRI) is true.

On the other hand, one might still accept (LU). In particular, one might reason like this.

> Yes, *A* is worse than *B* on balance for the newborn and the twenty-year-old, and yes, in a sense this accounts for all of the relevant people. But it doesn't account for all the relevant *person-stages*, nor for all of the relevant *welfare components*. Consider, in *A*, the twenty-year-old who the newborn eventually becomes. This twenty-year-old (a person-stage) has a whole year of good life (a positive welfare component), and surely that contributes to how good *A* is overall, even if we rightly discount it when asking how much 'the newborn' is harmed by the choice of *B* over *A*.

Proponents of the TRI account tend to accept (PA-TRI). That is, they not only see TRI as giving an account of the badness of death for a person at a time, they also take a specific stance on how the account feeds into judgments about what one ought to do, at least in simple cases like this. The conclusion that one ought to save the 20-year-old is one of the orthodox verdicts of the view. However, it has proved quite difficult to see how the TRI account can fit into a coherent moral theory, generalising (PA-TRI).

To get an initial diagnosis, let us consider more formally what the TRI account actually involves. The account depends on the idea of a person having a *stake* or *interest* in various welfare components. Here is how I interpret this idea. Suppose that I might have a three-hour headache tomorrow, or I might have a headache today. Suppose that, either way, the headache will make no difference to anything beyond my hedonic experience at the time it occurs. It will just make life less pleasant, to the tune of 1 unit of welfare per hour. If, as a matter of prudential rationality, I ought to be indifferent today between having a three-hour headache tomorrow and having a three-hour headache today, this indicates that I today have a *full stake* in my welfare tomorrow. If, rather, I ought to be indifferent today between having a three-hour headache tomorrow and having a *two*-hour headache today, this indicates that I today have a *two-thirds* stake in my welfare tomorrow. I will also say that the *strength* of my interest today in avoiding the headache tomorrow is the amount of welfare I could rationally give up today to forestall it tomorrow. This is essentially the eventual welfare-cost of the headache (three units of welfare for a three-hour headache) weighted by my stake in it. So if I have a full stake, then the strength of my interest amounts to three units of welfare; if I have a two-thirds stake, then the strength of my interest amounts to two units of welfare. In general, we might have something like this:

Time-Relative Interests. Person-stage S has an interest of strength i in outcome A over outcome B iff, as a matter of prudential rationality, S ought to be indifferent between A and the combination of B with an additional immediate welfare benefit to himself of size i, all else equal.

While the objects of time-relative interest are outcomes, states of affairs, or whatever else in general the objects of prudential preference might be, the things in which people have *stakes* are welfare components. In particular, the basic issue is the extent to which one person-stage has an interest in the welfare of another.[5]

Time-Relative Stakes. Person-stage S has a stake of size s in the welfare w of person-stage T in outcome A if and only if, as a matter of prudential rationality, that welfare counts in favour of A for S just as much as would an immediate welfare benefit to S himself of size sw, all else equal.

It follows that, if the only relevant difference between A and B is T's welfare, then the strength of S's interest in A over B is S's stake in T's welfare times

[5] Remember that when I speak of the welfare of a person-stage, I always mean the welfare that accrues to the person-stage as such, rather than anything like the lifetime welfare of the person of which the stage is a part. Note that the basic TRI account does not explain how to deal with welfare components that do *not* simply affect the welfare of person-stages. (Examples might include goods that arise from patterns of welfare within a life, and perhaps also goods that are not temporally localised, like the success of long-term projects.) However, this may just be because the basic account is incomplete. Indeed, the views I describe in section 6 are able to handle these holistic goods; I will say more about it then.

the size of the benefit to T of A compared to B.

This picture may be a little rough, but it is enough to give us a handle on what is going on. There are three important points about the formulation.

First, the unexplained notion of 'prudential rationality' is doing a lot of work. It means, first of all, that the strength of S's interests at t is a normative matter, not simply a descriptive matter of what S happens to be interested in. It also means that I am focusing on what we might call *self*-interest. I might rationally give up one unit of welfare in order to ensure that my friend gains two. In a general sense, this shows that I have a stake in my friend's welfare. But the reasons I have for accepting such trade-offs are not primarily a matter of self-interest. Indeed, prudential rationality requires me *not* to benefit my friend in this way (absent any knock-on effects on my wellbeing).[6]

The second point is that the strength of the interest I have today in my welfare tomorrow is not a matter of how much anything affects my welfare today. There is, perhaps, a sense in which now having a 'frustrated interest' can reduce my current wellbeing, but that is not what is going on here. My headache tomorrow is not painful to me *now*, but even a hedonist can recognise the sense in which I have a strong interest now in avoiding that headache. The headache should weigh heavily in my current prudential deliberations.

[6] I claim one can also have self-regarding preferences that are not egoistical or self-interested. If so, then I would also want to exclude from the domain of prudential rationality whatever standards govern such preferences. I suspect this distinction is important to understanding 'adaptation' cases, like that of the blind child who 'may rationally prefer to have been born blind' (McMahan, 2002, p. 295). That preference may not be based on the sense that one is *better off* being blind; though rational and self-regarding, it need not be *prudent*. Another ingredient in this case is the improper focus on the child's *actual* interests; I discuss the problems with actualism further below.

Third, the strength of my interests may indeed be *time-relative*. Tomorrow I will have a full stake in avoiding the three-hour headache, even if I have only a two-thirds stake today. Of course, we *might* hold that the strength of an interest is necessarily time-independent. Perhaps, in the relevant normative sense, I have now a full stake in the welfare of my organism at every other time. In that case, we could still talk about interests and stakes this way, but there would be no discounting. (For some ways of filling in the details, we would recover the LC account.) However, the teleological motivation for the TRI account is to justify intuitions in cases like Emergency Room in which the LC account seems to go wrong, and in particular this justification presumes that the strength of an interest can change with time. Specifically, the main thesis of the TRI account is

Psychological Reduction. S's stake in T's welfare in A measures the degree to which S is psychologically connected to T in A.[7]

Here is a point on which my formulation takes sides. Suppose that S exists in both outcomes A and B. Suppose also that S and T are strongly psychologically connected in outcome A, but not in outcome B. One might think that, contrary to my formulation of Psychological Reduction, if B actually occurs, then S does *not* have a big stake in T's welfare in A, because he is not actually strongly connected to T. This thought seems to reflect what McMahan says in *some* cases, like that of Prenatal Retardation, which I

[7] Why 'measures' instead of 'equals'? I am not sure there is an independent way of quantifying degrees of psychological connectedness with respect to which the claim of equality would make sense. However, we can certainly make qualitative comparisons, and the thought is that qualitative increases are suitably correlated with increases in stakes. This thought is made a little more precise in footnote 34 below.

discuss more in fn. 13 below. However, consider the basic Emergency Room case. There we clearly want to say that, even if Robert will actually die, he has a strong present interest in continuing to live. This interest can only be grounded in the psychological connections Robert *would* have had, had he lived.

We can now anticipate the basic difficulty for fitting the TRI account into a general normative theory. In Emergency Room, we identified the badness of death for the 20-year-old with the strength of his interest in avoiding death. But this was only the strength of his interest at a particular time. When he was a newborn, he did not have such a strong interest in the welfare of his maturity. If we want to know what the doctor ought to do, at what time should we evaluate the patient's interests? The time of the doctor's decision? The time of death? In the example, these times more or less coincide, but they need not in general, and, anyway, it is unclear why either of them is uniquely relevant. It may seem rather that we should somehow combine or aggregate the interests of different people at different times. That may be right, but note the following difficulty. In simple cases it is intuitively appropriate to aggregate the *welfare* of different people at different times. But the strength of an interest at a time is not the same as the impact on welfare at that time. So it is unclear whether we can reconcile interest aggregation with welfare aggregation in cases where the latter gives intuitive results. Returning to the example, suppose I today have a full stake in my welfare tomorrow. The headache will eventually cost me three units of welfare, so I have today and will have tomorrow a three-unit interest in avoiding the headache. But if I simply *add up* these interests, I double count: I get a disproportionate six-unit interest. The longer the strong interest persists, the larger the aggregate interest, even though the impact on welfare remains the same. It then seems absurd to focus on the aggregated interests rather than on the welfare effect. The same considerations rule out *averaging* instead of *adding*. Of course, there may be some other sort of aggregation rule that does a better job. My point is only that it isn't obvious what the right rule really is.

3 Problem Cases

Let us see how these general consideration play out in key cases. Return to Teddy the newborn and Robert the adult. Remember that each one will live to 60 at latest, gaining 10 units of welfare per year. But Teddy has only a small stake in his welfare after age 15, while Robert has a full stake in his.

> **Choice Between Deaths** (Table 2)[8]
>
> A doctor can either (A) save Teddy the newborn, or (B) let him die today. If the doctor saves him now, Teddy will foreseeably die when he reaches age 20. Sadly, there is no possibility that (C) Teddy lives a full life.

In Emergency Room, we said that Robert's death at 20 was worse than Teddy's death at birth. In Choice Between Deaths, we have to choose between *Teddy's* death at 20 and his death at birth. Blindly applying the earlier judgment that death at 20 is worse than death at birth, it seems we should

[8] A similar example is considered in McMahan (2002, p. 185)

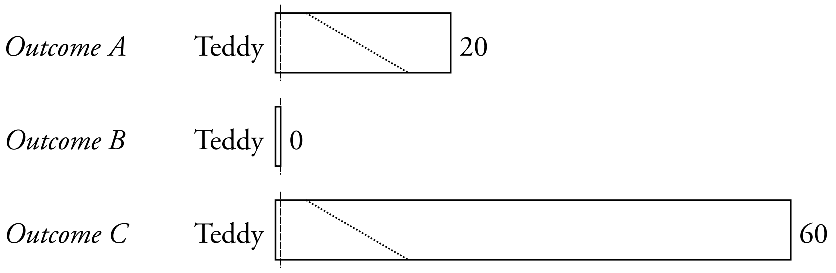

Table 2: Choice Between Deaths. The outcome C is unavailable.

let Teddy die at birth. But this must be wrong, given that Teddy's twenty years would be filled with happiness.

The problem is caused by insisting that death at 20 will be bad for Teddy. It will, of course, be bad compared to the unavailable option C in which Teddy survives and has a long and happy life. And of course, were C available, then the doctor ought to choose C over either A or B. But that is not the situation, and the comparison to C is irrelevant. We are asking whether A is better than B, or anyway whether A may be chosen over B.[9]

McMahan's view is that newborn Teddy has an interest in continuing to live, and what matters in the choice between A and B are Teddy's present

[9] McMahan's own diagnosis might appear to be different:

> Preventing the frustration of that later possible time-relative interest [i.e the death at 20] is important, though only certain means of prevention are acceptable. Others are self-defeating. If we could prevent the death at [20]...by *saving* the individual at that time..., our reason to do so would indeed be stronger than the reason the doctor has to save the infant now. But it is not a reasonable way of preventing the later time-relative interest from being frustrated to ensure that the individual will not then exist.McMahan (2002, pp. 187–8)

But I understand this to be compatible with what I have said about the case. McMahan is explaining *why* we might get a different answer when we compare A and B from the one we get when we compare them each separately to C.

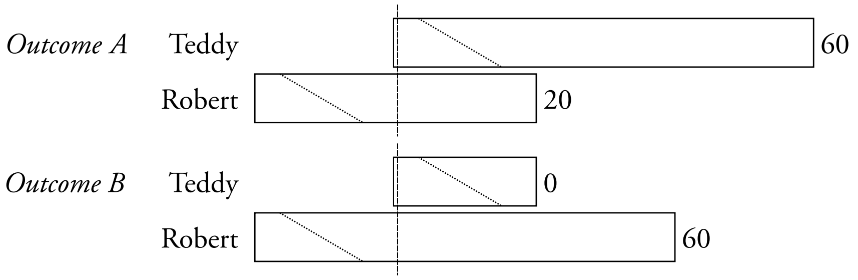

Table 3: Delayed Choice

interests. Now, I agree that this must be the upshot of the TRI account in this case. C must be morally preferable to A, and A must be morally preferable to B, and these verdicts coincide with Teddy's present interests. But the claim must not be that *in general* people's present interests are decisive. That is what the next case shows.

> **Delayed Choice** (Table 3)[10]
>
> Teddy the newborn and Robert the 20-year-old are each such that, unless they are treated now, they will die in 20 years' time. The doctor can only treat one, and whoever is treated will live until age 60.

If, as in Choice Between Deaths, we are guided by present interests, Robert has a stronger interest in remaining alive than Teddy does. We should therefore save Robert. One difficulty is that it seems at least permissible to save Teddy. A related difficulty is inconsistency over time. Twenty years in the future, Teddy may have a stronger interest in staying alive than Robert

[10] A similar case is considered in Broome (2004, pp. 249–251); he presents it as a decisive counterexample to the TRI account.

will (400 units versus 200, on my simple model). So it seems that the doctor will regret a decision to save Robert, and ought to reverse his decision if he can, despite having no new information.

If we are to avoid inconsistency over time, Delayed Choice must be just like an 'Emergency Room' choice between saving the life of a 20-year-old (future Teddy) and a 40-year-old (future Robert). When I discussed Emergency Room, the decisive factor was the strength of the patients' present interests, which coincided with their interests at the time of premature death. Thus, in Delayed Choice, the decisive factor must be the patients' interests twenty years in the future. But we have to be a little careful. Suppose we modify the case so that, on option *A*, in which Teddy is saved, the first year of Teddy's life will be full of suffering. This must count against option *A*, but how? The suffering goes against his present interests, but not against his interests twenty years from now. We saw that it could not just be Teddy's present interests that count, and now we see that it cannot just be his interests twenty years in the future that count. It is obvious in any case that we cannot just look at the point of death, since the TRI account must be able to handle cases that are not cases of premature death. What counts must be some sort of amalgamation of interests at different times.[11]

McMahan's response to cases like Delayed Choice was to emphasise the importance of all *actual* interests, including but not limited to *present* ones. But even if it were clear what it meant to take all such interests into account, focusing on actual interests leads to a kind of modal inconsistency even worse

[11] Contrast Millum (2015, p. 292); he says of Delayed Choice that we must simply look at the time of death. While it may be that the interests at the time of death end up being decisive in Delayed Choice, that can't be the fundamental story.

than the temporal inconsistency mentioned above. In Emergency Room, if Robert is actually saved, then Teddy has only a weak actual interest in continuing to live, and we ought to save Robert; but if Teddy is actually saved, it turns out we are allowed to save *him* in order to satisfy the strong interest he eventually gains in the welfare of his childhood and maturity.[12] There are really two problems here. First, what we ought to do depends on what we actually do – that is the modal inconsistency I mentioned. Second, if we actually save Teddy, then the orthodox verdict in Emergency Room is mistaken. Similar points are explained clearly by Holtug (2011). I will make some further comments as we go, and especially in section 8, but I will not press this line of objection, since McMahan's commitment to actualism was never consistent, and he has more recently disavowed it.[13] Suffice to say that, in combination, Emergency Room, Choice Between Deaths, and Delayed Choice render it mysterious how the TRI account is to be understood. I think that the problems raised by the next examples are even more telling.

These next few examples involve patterns of psychological connections that are more extreme than those of Teddy and Robert. It is plausible that

[12] As I said, it is not clear in general what being guided by actual interests may entail. Here I rely on a piece of dominance reasoning, which ought to be relatively uncontroversial. Since Teddy's life is actually longer than Robert's, the interest that Robert has in *B* over *A* at any given age can be matched against the interest that Teddy has in *A* over *B* at the same age. So *A* satisfies Teddy's actual interests at least as well as *B* would satisfy Robert's actual interests. (Remember that, in this scenario, Robert's *actual* interests are the ones he has before he dies at 20.)

[13] Here is an example not noted by Holtug in which McMahan may be led astray by actualism. He is greatly troubled by the case of 'Prenatal Retardation' (McMahan, 2002, p. 323). His treatment of the case trades on the claim that if (as actually happens) a mother causes a foetus 'to have cerebral deficits sufficient for severe retardation, it will *never* have a significant time-relative interest in being a person' and therefore it appears that harming the foetus in this way is not very wrong. But if the mother *didn't* harm the foetus then it would eventually have very strong time-relative interests in being a person; actualism improperly prevents us from considering this fact.

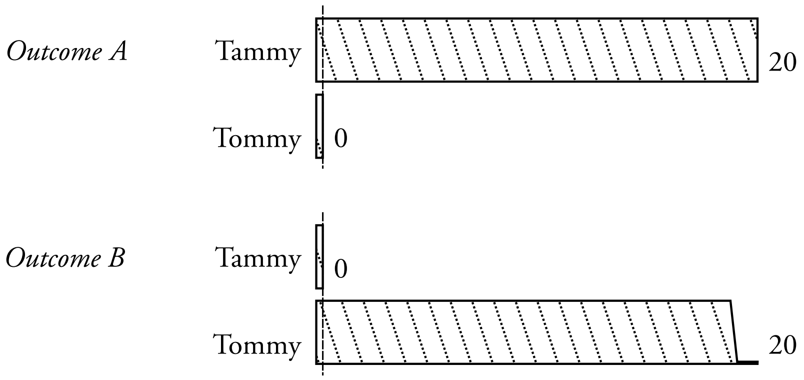

Table 4: Choice Between Horses. The dense diagonal lines schematically represent the assumption that a horse's prudential interests stretch no more than a year into the future.

many animals have limited psychological connections from one year to the next. Suppose Tommy the Horse and his sister Tammy have a large stake at any given time in their welfare over the next year, but no stake in anything beyond that. In these examples, Tammy and Tommy have just been born.

> **Choice Between Horses** (Table 4)[14]
>
> Tammy and Tommy are seriously ill, and the vet can save only one. Whichever one she saves will live until 20. If (*A*) Tammy is saved, she will accrue ten units of welfare per year. If (*B*) Tommy is saved, he will accrue ten units of welfare per year until age 19. But in his last, twentieth year, Tommy will accrue zero units of welfare.

[14] The character Tommy the Horse is drawn from Harman (2011) and discussed in McMahan (2015); but as far as I know, the problem raised by Choice Between Horses has not been previously considered in the literature.

This case looks similar to Emergency Room. However, Tammy's present interest in A over B has the same strength as Tommy's present interest in B over A. Going by present interests, it looks like the vet may as well save either horse. But this is quite odd. Surely she ought to save Tammy. Of course, we have already seen that we should not go by present interests alone, in general. On the other hand, it is not clear why this case is different from Emergency Room, in which present interests sufficed. Specifically: we claimed that the welfare that Teddy might have in middle age does not count in favour of saving him as a newborn, since he has no present stake in it. If that is correct, then the welfare that Tammy might have in her twentieth year should not count in favour of saving her either. But then it is hard to see what *does* count in favour of saving her over Tommy: their welfare in the twentieth year is the only potentially relevant fact.

A different kind of problem is raised by the following case.

> **Suffering Later** (Table 5)[15]
>
> In outcome A, Tommy the horse will die imminently. In outcome B, Tommy will have two years of slightly good life (one unit of welfare per year) followed by 18 years of miserable life (-10 units per year).

Following the logic of Choice Between Deaths, option A is worse for Tommy than option B. Although there is a great deal of misery in his future, he has no present interest in avoiding it, and, on the model of Choice Between Deaths, it is this present interest that is decisive. But it seems wrong

[15] A similar case is considered in McMahan (2015, p. 71)

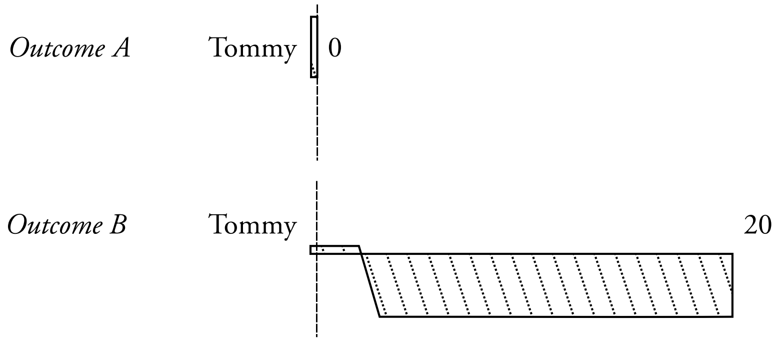

Table 5: Suffering Later

to prolong his life and so ensure immense suffering later on. We really ought to choose *A*.

Notice that this is a case in which the premature deaths occur more or less at the present moment. So it again shows that we cannot simply evaluate interests at the point of death, any more than we can simply evaluate interests at the present moment.

> **Suffering Now** (Table 6)[16]
>
> One would ordinarily expect Tommy to live for 20 more years (outcome *A*). The first 18 years would be very pleasant (10 units of welfare per year), but the last two would be full of suffering (−10 units per year). However, as it happens, Tommy has acquired a rare disease, which (*B*) will kill him imminently. The only treatment (*C*) will cause suffering for one year right away, but he will then have 19 years of very pleasant life.

[16] A similar case is considered in McMahan (2015, p. 72)

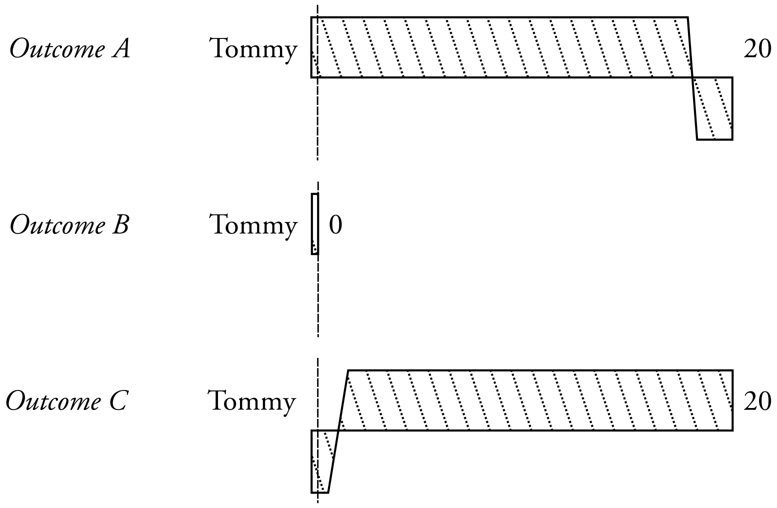

Table 6: Suffering Now

Going by Tommy's present interests, death now would be bad, compared to the ordinary expectations. It foils his strong present interest in the first year of very pleasant life, while he has no present interest in avoiding the suffering later on. On the model of Choice Between Deaths, *A* is morally preferable to *B*. What about treating Tommy? Stepping back, the consequences of treatment are clearly preferable to the ordinary expectations; they involve only two years of suffering, rather than the expected three. So *C* should come out preferable to *A*. Harman (2011) claimed that the TRI account could not deliver this verdict. After all, treatment is against Tommy's present interests, which emphasise his wellbeing over the first year. But the choice between *A* and *C* does not have the same structure as Emergency Room or Choice Between Deaths, in which present interests supposedly sufficed. So it is open to us to take into account Tommy's future interests. We lack a general story of how to do so, but in *this* case it is obvious what the upshot must be: Tommy's interest during the first year of life in A over C must be dominated by his interest during the last two years of life in C over A. So, as McMahan notes, the right judgment seems to be at least heuristically compatible with the TRI account.

Granted this result, C is morally preferable to A, and A is morally preferable to B. Surely, then, C is morally preferable to B: the cure is preferable to Tommy's death. However, the choice between B and C again has the structure of Choice Between Deaths. As in that case, we should compare B and C on the basis of Tommy's present interests. But Tommy's present interests favour B over C: death now would allow Tommy to avoid a year of suffering in which he has a large prudential stake.

If this is all correct, we have obtained a certain kind of intransitivity. If C is to be chosen over A, and A is to be chosen over B, it does not follow that C is to be chosen over B. The ordinary expectations are preferable to death now, and death now is preferable to treatment, but ordinary expectations are not preferable to treatment. As I mentioned in the introduction, this intransitivity is one reason I tend to talk about which options are morally preferable, rather than about which options are morally best. But, terminology aside, intransitivity causes problems when all three outcomes are available. We cannot simple-mindedly sort the options by the relation of preferability. Doing so would only drive us around in circles. Instead, we need an additional story about what to do in multi-option cases. It is less clear than ever what the TRI account amounts to.[17]

[17] McMahan (2015) considers a version of Suffering Now, but he does not explicitly

Recap

The main point of this discussion is just how hard it is to make sense of the TRI account, once we have moved beyond prudential value to questions of ethics. Doing so would, at a minimum, involve producing a reasonably natural moral theory that systematizes the recommended verdicts in the examples considered so far. Most of these recommendations were obtained by adapting the interest-based reasoning that underpins the orthodox verdict in Emergency Room. In that sense, they are also a part of the ethical orthodoxy associated with the TRI account. In other cases – in particular, in Choice Between Horses and Suffering Later – I appealed to strong theory-neutral intuitions. The overall pattern of verdicts is summarised in the following table, in which '$\succ$' denotes the relation of moral preferability.

Case	*Verdicts*
Emergency Room	$B \succ A$
Choice Between Deaths	$C \succ A, A \succ B, C \succ B.$
Delayed Choice	$A \succ B$
Choice Between Horses	$A \succ B$
Suffering Later	$B \succ A$
Suffering Now	$C \succ A, A \succ B, B \succ C.$

One natural objection is that, absent an actual theory, some of the interest-based reasoning used in the examples is not well motivated. For example, in Suffering Now, we appealed crucially to *present* time-relative interests. We know that we must not, in general, go by present interests alone. So

note the intransitivity, and this leads to some problems in his argument which I take up in section 8.

why may we do so in this case? Besides, the theory-neutral intuitions in Choice Between Horses and Suffering Later seem to directly contradict our understanding of what interest-based reasoning requires! Perhaps, then, the best way of systematising the interest-based reasoning would yield different verdicts from the ones I have, following McMahan, suggested here. However, until section 8, I will take these orthodox verdicts for granted. The question until then is how to incorporate them into a coherent moral theory.

4 The Different-Person Heuristic

In this section and the next, I explain how the problems faced by the TRI account are related to controversies in population ethics. The key issue in population ethics is whether creating a new person can make things better or worse overall. A connection to the TRI account may be drawn in several ways.

One way, which I point out only to set aside, involves a sequence of scenarios in which a person dies earlier and earlier. Naively, at least, the limiting case is one in which the person never exists at all. On the life-comparative account, dying as a newborn or a foetus is much worse than having a long and happy life. The earlier the death, the worse. This at least suggests that the limiting scenario is the worst one of all. It is very bad to prevent happy lives from coming into existence. On the time-relative interests account, dying as an infant, or as a foetus, is not much worse than having a long and happy life. (This is true in terms of the prudential interests of the foetus, and also true at the ethical level, given the orthodox verdict in

Emergency Room.) At a certain point, the sequence of scenarios gets better and better, in a way that suggests that the limiting scenario may not be bad at all. In other words, it is neither good nor bad to prevent happy lives from coming into existence.

However, this way of drawing the connection to population ethics is misleading. The application of the TRI account here depends on the person having less and less of a stake in its future welfare as the sequence proceeds. This is plausible for normal humans. A human foetus might not have much of a stake in its future welfare. But, in principle, a being could come into existence with a strong stake in its future welfare, right from the beginning. The TRI account would agree with the LC account about such creatures. Dying early would be much worse than having a long and happy life. So even if we countenance non-existence as the limit of shorter and shorter periods of existence, this application of the TRI account does not suggest any general lesson about the value of creating happy lives.

A second possible connection is the analogy between people and person-stages. Extending a life (thus adding person-stages) is in some ways akin to enlarging a population (thus adding people). However, this analogy does not seem to help with the TRI account. In ordinary cases, anyway, we have no prudential interest in the welfare of other people, but we do have at least some prudential interest in the welfare of future person-stages.

This section is devoted to developing a third point of view. As I explained in the introduction, it is a question of importing into ethics the idea that psychological connectedness is what matters. Since this question is of very broad interest, I will consider a wide range of examples. Then, in the next section, I will focus in on the intuitions in population ethics that corresponds to the orthodox verdicts of the TRI account.

Methuselah

Consider Methuselah (Lewis, 1976), who lives until the ripe old age of 969. His psychology gradually changes over time, so that none of his memories reach more than 137 years into the past, nor does he plan more than 137 into the future. In general, the psychological connections within his life are limited to that extent. Although Methuselah may be always the same person in an ordinary sense, there is at least an informal sense in which Methuselah at 100 is 'a different person' from Methuselah at 900.[18] The idea I want to consider here is that this sense is *normatively relevant*. When two person-stages are psychologically unconnected, we ought to treat them as if they belong to different persons. I will call this *the different-person heuristic*. It may be more than a heuristic. Perhaps it can be explained in terms of a normatively relevant notion of personhood, or at least of personal identity. That is the sort of picture I will consider in sections 6 and 7, but for now I want to leave the matter open.

Of course, the different-person heuristic is a version of the idea already introduced, that it is psychological connectedness rather than personal iden-

[18] Recall that I consider personal identity as a relation between person-stages, and so often ask whether two person-stages 'are' the same person, rather than whether they 'belong to' the same person. This is just a matter of terminology, but I find it helps me to bracket certain irrelevant distinctions. Note also that Methuselah is very much like Tommy the horse, drawn out over a longer timescale. Tommy at 1 and Tommy at 19 are different persons (or different horses?) in just the same sense that Methuselah at 100 and Methuselah at 900 are different persons. But here I focus on Methuselah, because a richer variety of moral considerations plausibly apply in the case of a long-lived human being.

tity that matters. It reframes that idea in a useful way. Suppose we have a standard normative theory, one that gives persons a fundamental role. That gives us a reference point for what it means to treat two person-stages as if they belong to different persons, or to the same person. This reference point can help us find a revised theory that satisfies the heuristic.

All this is best understood through examples. Some of these pertain directly to the TRI account, but others do not. They show that the project of reconciling normative theories with the different-person heuristic is of very broad interest.

Let me begin with prudential value. The different-person heuristic suggests that Methuselah at 100 should prudentially be indifferent about what happens to Methuselah at 900, in the same way that in general what happens to other people is not a matter of prudential concern. This is not to deny that Methuselah at 100 might have some special reasons to help Methuselah at 900. He may rightly privilege the wellbeing of Methuselah at 900 over that of other people at other times. But, the thought goes, we can legitimately separate these reasons from *prudential* reasons, reasons of narrow self-interest. They are more like the special reasons we may have to care for our own children.

In particular, killing Methuselah at 100 would frustrate his prudential interests to the extent of at most 137 years of wellbeing, counting to the outer limit of psychological connectedness, not anything like 900 years. A little more precisely, Methuselah at 100 should prudentially regard dying immediately as at least no worse than living on with an immediate one-time loss of 137 years' worth of welfare. That is the extent of his interest in continuing to live. The reader will recognise here a judgment compatible with the TRI account of prudential value.

What about ethics? Again, killing Methuselah at 100 prevents Methuselah at 900 from ever existing. More than that, insofar as Methuselah at 900 is 'a different person', killing Methuselah at 100 prevents *that person* from ever existing. Many people think that preventing someone from ever existing is morally neutral, a view I will consider in detail in the next section. On this sort of view, the fact that Methuselah at 900 will not exist, and hence will not have whatever welfare he could have had, cannot count morally against killing Methuselah at 100. The moral cost of killing Methuselah at 100 is again a matter of at most 137 years of wellbeing. Again, this is the kind of ethical claim that is orthodox to the TRI account.

I earlier mentioned an analogy between persons and person-stages. That is not quite what is going on here. The thought is that Methuselah at 900 is (at least heuristically) a different person from Methuselah at 100, i.e. that these person-stages do not stand in a relation of personal identity. But Methuselah at 101 presumably *is* the same person as Methuselah at 100, or perhaps *mostly* the same person, even if they are different person-stages. So I am not drawing out an analogy between persons and person-stages, but rather gesturing towards a normatively salient identity-like relation.

As a different sort of example, consider the intuition against *replacement*: roughly speaking, two forty-year lives that occur in sequence do not seem as good as a single eighty-year one, even if the welfare facts at each moment are the same.[19] Taking this idea seriously, it would have been better had Methuselah at 100 been the same person as Methuselah at 900. All else equal, it would have been better had Methuselah been strongly psychologically connected from birth until to death, even if it made no difference to his welfare from one day to the next. Then the 969 years of life would have been had by a single person, in the sense towards which I am gesturing, rather than being shared by the person who is Methuselah at 100 and the person who is Methuselah at 900, and whomever else.

Here is an example that is not simply concerned with welfare. Consider a case in which I impose a cost on you, in order to give you a greater benefit in five minutes. Contrast this with a case in which I again impose a cost on you, but this time give the greater benefit to someone else. The second case strikes me as potentially *unjust* in a way that the first case does not. Although from a general point of view the harm is compensated by a benefit in either case, it is a matter of justice whether *your* harm is compensated by *your* benefit. If that's right, then it may be unjust in this same way to harm Methuselah at 100 in order to give him a greater benefit him at 900. The person who is harmed is not the person who benefits.

Or suppose you are deciding whether to confer a certain benefit on me now, or to do it in five seconds. It does not seem to matter which one you choose, or whether you flip a coin to choose. But if you are choosing between conferring the benefit on me or on someone else, it may be *fairer* to flip a

[19] See e.g. Broome (2004, §7.2). He concludes that we 'intuitively value longevity', but there may be better ways to conceptualise the intuition in question.

coin. So too, it may be fairer to flip a coin when deciding whether to benefit Methuselah at 100 or at 900.

Some philosophers are prioritarians, and this might be understood axiologically (as is common) or deontically (as seems more plausible to me). Roughly speaking, they think that badly-off people have a stronger claim than well-off people do to a benefit of any given size (Parfit, 1997). According to one common version of the view, whether someone is 'well off' or 'badly off' is a matter of her lifetime welfare. Even if someone is suffering now, she may not have a strong claim to priority, since her suffering may be compensated by happiness at another time. But insofar as Methuselah will be a different person at 900, his happiness then cannot compensate his suffering at 100. Contrast Methuselah with Methuselah's Twin, who is psychologically like Methuselah, and has the same birth-to-death aggregate welfare, but suffers more from age 0 to 500 and prospers more from 500 to 969. Would it be better to give a benefit to Methuselah at 100, or a same-sized to Methuselah's Twin at 100? Going by birth-to-death welfare, Methuselah and Methuselah's Twin both have the same claim to priority. But the different-person heuristic suggests that Methuselah's Twin at 100 has priority over Methuselah at 100. He suffers more, and his suffering is not compensated by later happiness, so he has a stronger claim.[20]

These examples show that the different-person heuristic may have wide-

[20] This is essentially the kind of 'prudential' prioritarianism defended by Holtug (2010). I discuss it more in section 7. Yet another kind of prioritarianism claims that people receive priority based on their welfare at the time in question, so that a low-welfare person-stage has a stronger claim than a high-welfare one. One could tweak the example given in the text to distinguish prudential prioritarianism from this alternative view.

ranging consequences. But there is a difficulty that I was able to ignore in the examples given so far. It is plausible enough that we should treat Methuselah at 100 and Methuselah at 900 as distinct persons. But what about Methuselah at 100 and Methuselah at 230? They are psychologically connected, but only to a low degree. Their relationship is not *very* different from the relationship between Methuselah at 100 and Methuselah at 900, but there is still a difference. Presumably, this means we should treat them to some extent as if they are distinct persons, and to some extent as if they are the same. But what that amounts to is unclear.

In some cases, it is relatively clear what to do. In the case of prudence, we have the standard life-comparative theory on which the prudential value of an outcome A to a person-stage S is the lifetime welfare of the person of whom S is a part. There is a fairly natural way to adapt this theory in light of the different-person heuristic. To treat my future person-stages 'to some extent' as if they belong to different persons is to partially discount their welfare in my prudential deliberations. This is how the different-person heuristic suggests the TRI account of prudential value.

The argument, though, is only suggestive for now. We lack a proper theoretical understanding of the heuristic, and how to implement it, especially for intermediate degrees of psychological connectedness. Once we have that deeper understanding, we can apply it to ethics as well, and thus embed the TRI account into a broader normative theory. That is the issue I explore in sections 6 and 7.

5 The Principle of Neutrality

First, however, let me consider a different question. The plan is to revise a standard theory of ethics in light of the different-person heuristic. But what standard theory should we choose? What features should the theory have in order to reproduce the orthodox verdict of the TRI account in Emergency Room, and to give the recommended answers in the problem cases of section 3?

The principle of neutrality

Let us look at Emergency Room again. The heuristic I am working with is that Teddy as a newborn is a different person from the one he will be if and when he reaches maturity. Let us call these persons Young Teddy and Old Teddy. (I will keep it vague what happens in between Young Teddy and Old Teddy; that is where we will have to revisit the issue of degrees.) Similarly, I can talk of Young Robert and Old Robert. Both Old Robert and Young Robert exist, on either outcome. Thus Emergency Room is analogous to the following case.

> **Harmful Creation**
>
> Three people, Young Teddy, Young Robert, and Old Robert currently exist, and the question is whether to create a fourth person, Old Teddy. Creating Old Teddy will benefit Young Teddy, but harm Old Robert to a greater degree. More precisely, the following outcomes are available:

	Outcome	
	A	*B*
Young Robert	90	90
Old Robert	110	510
Young Teddy	90	0
Old Teddy	510	

Table 7: Harmful Creation. Each entry gives the lifetime welfare of the person in question. The blank indicates that Old Teddy does not exist in outcome B. The reason for these specific welfare values will become clear in section 6.

Total utilitarianism, for example, rules that outcome *A* is preferable to outcome B. But the orthodox verdict in Emergency Room corresponds to the verdict that *B* is preferable to *A*. The most plausible and robust explanation is that Old Teddy's good life on *A* does not count in favour of *A* over B.[21] This is presumably because Old Teddy's very existence is contingent on the choice between *A* and B.

The *principle of neutrality*, as I will call it, is the rough principle that there is no moral reason to create additional people, like Old Teddy, with good lives. (In Narveson's famous formulation, 'We are in favour of making people happy, but neutral about making happy people' (1973, p. 80).) This is a popular intuition in population ethics, but it faces notorious difficulties. At the end of this section I will sketch one theory that overcomes them, by and large. But the immediate task is to relate them to the problem cases described in section 3.

[21] It could just be that Old Teddy's welfare counts to a diminished extent. However, the basic logic of the TRI account seems to be that it counts not at all, because Young Teddy has no stake whatsoever in Old Teddy's welfare.

Simple Necessitarianism

When thinking about the principle of neutrality, it will help to have in mind the simplest theory of population ethics that satisfies it. I call this *simple necessitarianism*. Suppose that we are comparing options A and B. Let $\mathcal{N}$ be the group of people who exist in both A and B, the 'necessary' people. (The other people are 'contingent'.) Simple necessitarianism says that A is preferable to B just to the extent that the total welfare of the people in $\mathcal{N}$ is greater in A than in B.

To illustrate: in Harmful Creation, $\mathcal{N}$ consists of Young Robert, Old Robert, and Young Teddy. Their total welfare in A is 290 and their total welfare in B is 600, so B is preferable to A. A similar analysis works for analogues of Choice Between Deaths and Delayed Choice. For example, the analogue of Choice Between Deaths is a situation in which Young Teddy necessarily exists, and we are deciding whether to create Old Teddy. Creating Old Teddy slightly benefits Young Teddy. In such a case, simple necessitarianism says that Old Teddy's welfare is irrelevant, but since creating him benefits the only relevant necessary person, we ought to do so. This corresponds to the orthodox verdict that we ought to save Teddy in Choice Between Deaths. Simple necessitarianism also supports the orthodox verdicts in Choice Between Deaths and Delayed Choice. I will leave the details to the reader. Since we were initially puzzled by how to reconcile these two cases with Emergency Room, it is an important observation that they can all three be explained by the combination of simple necessitarianism with the different-person heuristic.

In choosing to begin with simple necessitarianism, I have interpreted the principle of neutrality as discounting the welfare of *contingent* people. There are some closely related interpretations, which, however, are ultimately less plausible.[22] The way they go wrong is instructive in light of the point just made: they lead to bad results in Emergency Room and Delayed Choice, of a type we have already seen. One interpretation holds that only the welfare of *presently existing* people matters. In Harmful Creation, the explanation for why Old Teddy's welfare does not count is that he does not yet exist (whether or not he ever will). But suppose we could set in process a chain of events that would eventually harm Old Teddy, without affecting anyone else. On this 'presentist' view, it would not be wrong to do so. Once Old Teddy came to exist, however, we would surely regret our past action, and want to undo it if we could. This is the same kind of temporal inconsistency that threatened the TRI account in Delayed Choice. Indeed, the focus on presently-existing people corresponds to a focus on present time-relative interests, given the different-person heuristic.[23] Alternatively, one might think that we should not count Old Teddy's welfare if he does not *actually* exist. Assuming that we do not actually create him, it would be wrong to do so. This leads to the sort of modal inconsistency we encountered in section 3: if we *do* actually create Old Teddy, then it turns out we were *obliged* to do so. One problem is the inconsistency, and another problem is that this latter verdict contradicts the principle of neutrality.

[22] See Arrhenius (2013, ch. 10) for a more comprehensive discussion.

[23] The correspondence between presently existing people and present time-relative interests (or between actually existing people and actual interests) is not quite on-the-nose; after all, presently existing people may have different interests in the past or future, and actual people may have merely possible interests.

	Outcome		
	A	*B*	*C*
Adam	0	0	0
Eve			0
Steve	10		

Table 8: Adam and Steve

In summary, presentist and actualist interpretations of the principle of neutrality lead to temporal and modal inconsistency. We should reject these interpretations of the principle, for the same reason that we must reject ethical elaborations of the TRI account that rely only on present or actual interests.

The Non-Identity Problem

Now I will raise three crucial problems for the principle of neutrality. Afterwards I will relate them to the problem cases described in section 3.

Consider first the following case.

> **Adam and Steve** (Table 8)
>
> God can create either Adam and Eve, or Adam and Steve. Steve would have a good life; Adam's and Eve's lives would be neither good nor bad.

The principle of neutrality puts us under pressure to say that neither *A* nor *C* ought to be preferred over the other. And this is the verdict of simple necessitarianism, which only cares about the welfare of the necessary person, Adam. But intuitively *A* is preferable to *C*, because Steve is better off in *A* than Eve is in C. The pressure here is both formal and substantive. Formally: if the principle of neutrality holds that A and B are equally good, and B and C are equally good, then, by transitivity of 'equally good', A and C are equally good.[24] Substantively: one typical *justification* for thinking that A is not preferable to B, in line with the principle of neutrality, is the 'person-affecting' consideration that no one would be benefited by the occurrence of A instead of B. (Since Steve doesn't even exist in B, the thought goes, he can't be *better off* in A than in B). But this type of argument also seems to show that A is not preferable to C, since no person would benefit from the occurrence of A instead of C.

Intransitivity

One way to avoid the formal version of the non-identity problem, above, is to deny that 'equally good' is transitive. The person-affecting considerations about harm also seem to suggest intransitivity of choice. Consider, for example, the following case.

> **Threesome** (Table 9)
>
> God has created Adam and is wondering whether to create Eve and Steve. He can either (A) create them in a way that benefits Adam, although they themselves would have neutral lives; (B) not create them at all; or (C) create them in a way that harms Adam, yet gives them good lives.

[24] Let me say that two outcomes are *equally preferable* to mean that, given a choice between only the two of them, either option is permissible. There are many well-known cases that show that 'equally preferable', rather than 'equally good', need not be transitive. But the point remains that the most obvious axiological treatment of the principle of neutrality will

	Outcome		
	A	*B*	*C*
Adam	10	5	0
Eve	0		10
Steve	0		10

Table 9: Threesome

The person-affecting considerations suggests that A is preferable to B (it benefits Adam), B is preferable to C (it again benefits Adam), but C is preferable to A (the benefits to Eve and Steve outweigh the harm to Adam). This is indeed the verdict of simple necessitarianism. Such intransitivity is not automatically fatal; it does mean that we cannot simply use pairwise comparisons to choose when all three options are available. It also shows that permissibility cannot be a straightforward matter of maximizing value.[25]

Asymmetry

Although many people have the intuition that we have no moral reason to create happy people, few people have the intuition that we have no moral reason not to create unhappy people. There is an asymmetry between creating good lives and creating bad ones.[26] For example:

Long-Suffering Eve

God has already created Adam, and is wondering whether to

not work here.

[25] There are further potential problems arising from intransitivity beyond the reduced effectiveness of pairwise comparisons, but I will not consider them here. An additional point is that the verdict that B is strictly preferable to C may seem too strong. I will return to this consideration in section 8.

[26] See McMahan (1981) for an early discussion of this problem, and Roberts (2011b) for a more recent survey.

	Outcome	
	A	*B*
Adam	0	1
Eve		-10

Table 10: Long-Suffering Eve

> create Eve. Doing so will slightly benefit Adam, but Eve's life will be full of suffering.

Simple necessitarianism suggests that A is preferable to B, since it benefits Adam, and Eve's very existence is contingent on the choice. But it is hard to believe that the modest gain for Adam makes it obligatory, or even permissible, to create Eve with a miserable life.

Strictly speaking, the principle of neutrality is silent about this case, because the principle only concerns the creation of good lives. However, the example is still a problem for the principle, because the asymmetry is hard to maintain. First, the intuitive judgment in Long-Suffering Eve undermines the person-affecting motivation for the principle of neutrality. If Old Teddy cannot be benefited in Harmful Creation, then Eve cannot be harmed in Long-Suffering Eve. More generally, it is hard to justify the asymmetry theoretically, or to obtain it without putting it in by hand. Second, the obvious ways of putting in the asymmetry by hand make it too strong. For example, suppose you are considering creating a new person, and there is a 99.99% chance she will have a great life, and a 0.01% chance that she will have a slightly bad life. A naive application of the asymmetry suggests that one ought not to create the person, since the possibility of a great life provides no reason to create her, and the possibility of a bad life provides a reason not to do so. But this is a very high standard for permissibility which most acts of procreation cannot meet. As a slightly different example, a naive application of the asymmetry suggests that it would would be best for humanity to throw in the towel, since some people in the future are otherwise bound to have bad lives, and this cannot be outweighed by the fact that many people in the future are bound to have good ones.

The Problem Cases Revisited

I claim that the problems I just raised for the principle of neutrality correspond exactly to the problem-cases described in section 3, given the different-person heuristic. I have already mentioned the connection with Choice of Death and Delayed Choice; we should not formulate the principle of neutrality in terms of presently existing people, nor in terms of actually existing people.

According to the different-person heuristic, Choice Between Horses is an instance of the non-identity problem. It is closely analogous to Adam and Steve. Saving Tammy involves creating a future person (the one who is Tammy-aged-19) with high welfare; saving Tommy involves creating a different person (the one who is Tommy-aged-19) with lower welfare. Simple necessitarianism problematically suggests that either of these options is permissible. In contrast, if we have a population ethics that treats basic non-identity cases in the right way – favouring the creation of Steve over the creation of Eve – then the different-person heuristic will advise us to save

Tammy.

The problem in Suffering Later is the problem raised by the asymmetry. Just as in Long-Suffering Eve, saving Tommy involves a modest benefit to a person who already exists (the person who is Tommy at birth) but also the creation of a person, or indeed several people, with negative welfare. Any population ethics that reproduces the asymmetry – ruling against the creation of Eve in Long-Suffering Eve – will say that we ought to let Tommy die.

Suffering Now illustrates at least two points. One is that the asymmetry must be moderated. Given the different-person heuristic, the choice of A (the ordinary expectations) over B (death as a newborn) would involve a benefit to Tommy, but also the creation of a series of future persons, a few of which would have bad lives. I take it that A is pre-theoretically preferable to B, yet the number of bad lives seems sufficient to outweigh the small benefit to Tommy. Thus the additional *good* lives in A must somehow count in its favour, at least in the sense that they offset or defeat the reasons we have not to create the bad lives.[27]

Second, setting aside the asymmetry, Suffering Now illustrates the intransitivity that arises from the principle of neutrality. The case has the same structure as Threesome. (In Threesome I have made the welfare levels positive, in order to bracket the role of the asymmetry.) When we compare outcomes A and B in Threesome, Eve and Steve count as contingent people, and their welfare does not matter. So too when we compare B with C. But when we compare A with C, Eve and Steve no longer count as contingent, and their

[27] In McMahan's terminology, the creation of good lives must have 'offsetting' but not 'reason-giving' weight (McMahan, 2015, p. 76).

welfare matters. That is the source of intransitivity. In Suffering Now, the welfare of Tommy at age 19 does not matter when we compare *A* with *B* or *B* with *C*. By the different-person heuristic, he is a contingent person. But when we choose between *A* and *C*, the existence of Tommy at 19 is no longer contingent on the choice. His welfare must matter, and we obtain the intransitivity once more.

Suppose we can find a theory of population ethics that respects the principle of neutrality sufficiently to give the desired answer in Harmful Creation, but also handles the non-identity problem and the asymmetry, and gives plausible verdicts even in choices involving more than two options. Suppose further that we can combine this theory with the different-person heuristic, just as we adapted the standard theory of prudential value to get the TRI account of prudential value. In applying this adapted theory, we may not end up directly aggregating or otherwise appealing to time-relative interests. However, we will have a unified conceptual framework that includes the TRI account of prudential value, and, we can hope, makes clear how time-relative interests prove to be decisive in simple cases like Emergency Room.

A Theory of Population Ethics

It is possible to modify simple necessitarianism to handle the problem cases adequately. But it is not so easy to find a way that is really plausible overall, and that does not feel irredeemably *ad hoc*. I will now describe the best simple theory I know that does the trick, partly as a proof of principle, and partly to use as an example later on. For lack of a better name, I will call the theory *complex necessitarianism.*

Let me first explain the structure of the theory. Given a pair of options $\{A, B\}$, the theory associates values $V_{\{A,B\}}(A)$ and $V_{\{A,B\}}(B)$ to each of them. Given a choice between A and B, one ought to maximize this value. There are several different basically plausible ways to extend this rule to a choice among many options. I do not know which one is most plausible, and it does not matter very much for my current purposes. One can either try to define the multiple-choice rule directly, or one can construct it out of the pairwise comparisons. The most plausible rule of the latter kind is defended (indeed, axiomatically derived) by Schwartz (1972). Here is a simple reformulation of his rule. Let me say that a non-empty subset $\mathscr{S}$ of the available options is *deliberatively stable* if no available option outside $\mathscr{S}$ is preferable to an option inside $\mathscr{S}$. Schwartz claims that an option is permissible just in case it is contained in a *minimal* deliberatively stable subset of the available options. This amounts to optimization when preferability happens to be transitive. In the simplest intransitive case, when A is preferable to B and B is preferable to C but C is preferable to A, and no other options are available, any one of the three options is permissible. In this matter, I may as well go along with Schwartz.[28]

It remains to define the value $V_{\{A,B\}}(B)$. Here is the main idea; afterwards I will write down some formulae. The idea is to use a counterpart relation between the people in A and the people in B. The counterpart relation must satisfy two constraints. First, it should extend the relation of transworld

[28] See Ross (2015, §5) for a discussion of similar principles specifically in the context of person-affecting population ethics.

identity. That is, if a person is necessary, existing in both outcomes, he must be his own counterpart. Second, the counterpart relation must pair up as many people as possible from B with people in A. (In other words, if A has more people than B, then everyone in B should have a distinct counterpart in A; otherwise, everyone in A should have a distinct counterpart in B.) Relative to such a counterpart relation, the value of B is the total welfare of the people in B who have counterparts in A. The idea is that, relative to the counterpart relation, these people count as necessary people. In the spirit of simple necessitarianism, their welfare counts in favour of B. However, there are usually many counterpart relations meeting the two constraints. To get the final value $V_{\{A,B\}}(B)$, we just average the value of B relative to all candidate counterpart relations.[29] There is one exception to this rule, to handle the asymmetry. If the contingent people in B (those who do not exist in A) have negative total welfare, then this must count fully against B. In such a circumstance, $V_{\{A,B\}}(B)$ is just the total welfare in B.

The upshot of this construction can be expressed in terms of the following statistics:

$T_{\text{nec}}(B)$, the total welfare of the necessary people in B;
$N_{\text{con}}(B)$, the number of contingent people in B;
$T_{\text{con}}(B)$, their total welfare.

Simple necessitarianism evaluated B purely in terms of the welfare $T_{\text{nec}}(B)$ of the necessary people. Complex necessitarianism introduces an extra term

[29] This approach has a family resemblance to the 'saturating counterpart' theory of Meacham (2012). Instead of averaging over many admissible counterpart relations, he introduces an extra 'harm-minimizing' constraint that effectively picks out a unique one.

to overcome the non-identity problem and to account for the asymmetry:

$$V_{\{A,B\}}(B) = T_{\text{nec}}(B) + X.$$

To define X, there are two cases. First, suppose that the contingent people in B have negative or neutral total welfare. Then

$$X = T_{\text{con}}(B), \quad \text{if } T_{\text{con}}(B) \leq 0.$$

Otherwise,

$$X = \min(N_{\text{con}}(A), N_{\text{con}}(B)\frac{T_{\text{con}}(B)}{N_{\text{con}}(B)}, \quad \text{if } T_{\text{con}}(B) > 0.$$

Here $\min(N_{\text{con}}(A), N_{\text{con}}(B))$ is either the number of contingent people in A or the number of contingent people in B, whichever is smaller.

I can finally explain how complex necessitarianism respects the principle of neutrality, and handles the problems for that principle. In Harmful Creation, there are no contingent people in outcome B, and the sole contingent person in outcome A has positive welfare. In such a case, complex necessitarianism compares the options using the total welfare of the necessary people. It agrees with simple necessitarianism: only B is permissible.

In Adam and Steve, the non-identity case, Adam is the sole necessary person, for any pair of options, and any contingent people have non-negative welfare. Given a choice between A and B only or between B and C only, the decision is determined simply by the welfare of the necessary person, so either option is permissible. So far, this agrees with simple necessitarianism. But given a choice between A and C only, the view balances Eve's welfare against Steve's. They are treated as counterparts of each other. Specifically, we have $V_{\{A,C\}}(A) = 10$ and $V_{\{A,C\}}(C) = 0$, so only A is permissible. That is the verdict we desired. What if all three options are available? The deliberatively stable sets of options are $\{A, B, C\}$, $\{A, B\}$, $\{A\}$, and $\{B\}$. Thus the permissible options are A and B. I find that intuitions in multi-option cases are harder to sort out, but this seems a satisfactory verdict.

In Threesome, the example of intransitivity, A is preferable to B and B is preferable to C. In each of these cases, the decisive factor is Adam's welfare. But C is preferable to A, because, with respect to these two options, all three people are necessary, and they have higher total welfare in C. So far, these verdicts coincide with those of simple necessitarianism. The only new ingredient here is Schwartz's rule, when all three outcomes are available. It says that any one of them is permissible. I do not know whether that is the right verdict, but it is at least not clearly wrong.

Finally, in Long-Suffering Eve, B has a slight advantage with regards to Adam's welfare, but Eve's suffering counts strongly against it. We have $V_{\{A,B\}}(B) = -9$ and $V_{\{A,B\}}(A) = 0$. Thus only A is permissible. Moreover, the asymmetry here is a moderate one. Suppose in general that A differs from B only by the addition of some new people. Complex necessitarianism says that *not* adding them is always permissible, and adding them is permissible if and only if their welfare will be positive on average. Humanity need not throw in the towel.

Complex necessitarianism is a bit too *ad hoc* to be satisfactory – especially in its treatment of the asymmetry – but it is fairly simple and matches the desired verdicts in these problem-cases as well as any theory I know.[30]

[30] As far as the standard intuitions of population ethics go, there is one place where

I will present a theory which I slightly prefer in section 8, but for now let us turn to the question of how to adapt a standard theory of population ethics, and complex necessitarianism in particular, to deal with degrees of psychological connectedness.

6 Recasting the ‘Person’ Role

What I have called ‘standard’ normative theories presuppose that each outcome contains a set of people, and that the people are composed of non-overlapping sets of person-stages. Thus two person-stages either belong to the same person or they do not. Sticking to prudential rationality, and assuming that prudential concern is *self*-interest, this means that a person-stage S has a full stake in the welfare of person-stage T if T is the same person as S, and no stake otherwise. There seems to be no question of S having only a partial stake in T's welfare.

The default way of interpreting these theories takes a ‘person’ to be something more or less coextensive with a human organism. Lifetime welfare is accumulated more or less from conception until death, even in the case of Methuselah. But suppose we could identify some *other* entities to play the role

complex necessitarianism falls down. It leads to one form of the ‘sadistic conclusion’. The basic form of the sadistic conclusion, which sounds particularly bad, is that it is sometimes obligatory to create bad lives instead of good ones. But that is not the form that follows from complex necessitarianism. Instead, suppose you can create two groups of people, the Ps and the Qs, or two other groups, the Xs and the Ys. No one in one option is identical to anyone in the other. Suppose that the Ps are equinumerous with the Xs, with the same welfare levels instantiated. Suppose the Qs all have good lives, and the Ys all have bad ones. Then complex necessitarianism says that it is sometimes obligatory to create the Xs and the Ys instead of the Ps and the Qs. This is because the Xs and Ys may have higher welfare on average than the Ps and Qs. This judgment is defended by Kath (2016, pp. 176–8), and I am sympathetic to what she says. The theory she advocates, ‘shortfall utilitarianism’, is one inspiration for complex necessitarianism.

of persons, in a way compatible with the different-person heuristic. Having chosen a standard normative theory – for example, complex necessitarianism – we could then apply it to these entities. We would not really have to *adapt* the theory at all, except to reinterpret it in this way.[31] In this section, I consider three ways this might work.

Indeterminate Persons

The simplest way to take the different-person heuristic seriously is to suppose that, at some point, the person who is Methuselah at 100 ceases to exist, and a new person's life begins. The various person-stages who make up Methuselah are grouped into persons. *How* they are grouped may be rife with indeterminacy. Perhaps Methuselah at 100 is a different person from Methuselah at 300, but we cannot expect there to be a determinate point at which the first person is replaced by the second.

How might this picture depend upon *degrees* of psychological connectedness? Determinacy itself comes in degrees, often called degrees of truth. There will be some determinate cases of personal identity – those of full psychological connectedness – and some determinate cases of non-identity, and a spectrum of borderline cases. We may posit that the degree to which it is determinate that S and T are the same person equals the degree to which S and T are psychologically connected.

How can a standard normative theory cope with this kind of indeterminacy? To begin with prudential value, if it is indeterminate whether T is

[31] In other contexts, other entities may count as persons. The thought is only that there may be a normatively relevant way of talking about persons that validates the different-person heuristic.

personally identical to S, how should considerations of T's welfare figure in S's prudential reasoning? There are many possible answers. But if the project is to interpret the TRI account in these terms, then we know what the answer must be. It must work out that the weight that ought to be given to T's welfare in S's prudential deliberations tracks the degree to which S and T are psychologically connected, or the degree to which it is determinate that they are the same person. And it is quite natural to think of this weight as an objective *probability*.

How so? A standard idea from decision theory is to identify one's credence in a scenario with the weight that one gives to considerations about one's welfare in that scenario. If there is an objectively correct weight, or objectively correct credence, then that weight has the status of an objective probability.[32] So it is natural to try to understand the stake that S has in T's welfare as an objective probability governing S's credence in the scenario that T's welfare is a matter of self-interest for S, or that T is the same person as S. If this is right, we can set indeterminacy aside, and focus on uncertainty.[33]

To start afresh: suppose that it is objectively uncertain whether T is the same person as S. There *might* be a time somewhere between S and T where one person ends and another begins, but it is unknown, and unknowable, whether and where this happens. Then considerations of T's welfare ought to have only intermediate weight in S's prudential deliberations. Assume in particular that S's prudential deliberation is governed by expected utility

[32] I follow Lewis (1980) in thinking that chances are functionally characterised by the way in which they norm credences, through something like his 'Principal Principle'. Here the issue is the correct credences to have about personal identity, conditional on a complete specification of the psychological connections between person-stages.

[33] A similar line of thought is pursued by Williams (2014b).

theory, and suppose that there is a 2/3 chance that T is the same person as S. Then S ought to be prudentially indifferent between a three-unit benefit to T and an immediate two-unit benefit to himself. That is, relative to the objective probabilities, S has a 2/3 stake in T's welfare. In general, a version of the TRI account of prudential value will arise *automatically* on the hypothesis that the objective probability that T is the same person as S coincides with the degree of psychological connection between T and S. Call this the *probability hypothesis*.[34]

On the probability hypothesis, the TRI account *just is* the standard account of prudential value, interpreted in terms of the normatively relevant notion of personal identity. We do not have to adapt the standard account. We just use a resource already implicit in it: decision theory under uncertainty.

The same move will work when we go beyond prudential value to ethics. Any well-developed theory of population ethics will tell us what to do in cases of uncertainty.[35]Given uncertain personal identity, it may be uncertain how many people there are and hence uncertain which lifetime welfare levels are instantiated. But as long as our favourite population ethics explains how to handle uncertainty, we can still apply it to these cases. This is a universal way of taking the different-person heuristic seriously, and extending it to cases

[34] My preferred way of understanding the probability hypothesis is that that the numerical representation of degrees of psychological connectedness is *defined* by its link to credences about personal identity, and thus to the size of prudential stakes; cf. footnote 7. The hypothesis is nonetheless a substantive constraint, as I will explain in my discussion of fission cases below.

[35] Unfortunately, few theories count as 'well-developed' by this criterion. There is often, at best, a vague gesture towards expected utility theory. But at some point the details matter. For example, I have already explained why it is difficult to reconcile the asymmetry with expected utility theory: it makes procreation impermissible. I have not yet explained how complex necessitarianism should incorporate uncertainty, but I will make a suggestion about that below.

of intermediate psychological connectedness. This universality is itself an important feature. It means we can understand the different-person heuristic while remaining neutral about the normative theory, and, in particular, whether or not we accept the orthodox verdicts of the TRI account.

Such flexibility even remains important when we focus on theories of prudential rationality. In my discussion of prudential value I have mainly focused on the welfare of person-stages. For example, I have explained what it means for one person-stage to have a stake in the welfare of another. But many people think there are considerations of prudential value beyond the welfare of person-stages. Suppose a life starts off badly and gets better. That may be better overall than a life in which the welfare of person-stages occurs in the opposite order, starting off well and getting worse. But this purported difference in lifetime welfare, or prudential value, cannot be attributed to the welfare of any person-stage. Or perhaps there are welfare components, like long successful projects, that make a life go well, but which do not contribute to the welfare of any person-stage in particular. If time-relative interests are just a matter of uncertainty about personal identity, then factors like these can be included automatically in the TRI account of prudential value. We can take the notion of lifetime welfare as a black box, even if it is uncertain which lives there are.

Examples

Before explaining the main difficulties for this approach, let me illustrate it with another visit to Emergency Room. Consider a simple model on which any human being starts off as one person and becomes another, and remains that other person until death. On this model, the transition happens somewhere between ages 3 and 15, but otherwise we are uniformly uncertain about when. To connect with my discussion of Harmful Creation, I will use 'Teddy' and 'Robert' simpliciter as names of the human organisms, individuated by physical continuity over time. In contrast, let 'Young Robert' be the person who Robert is as a newborn and 'Old Robert' the person who Robert is as an adult. Similarly for 'Young Teddy' and 'Old Teddy'. Young Robert's life has already ended no matter who is saved, but his lifetime welfare is uncertain. It consists of 30 units of welfare accrued in infancy plus some amount between 0 and 120 units accrued in childhood. (Despite my comments above, I will assume that lifetime welfare is just the total welfare of the relevant person-stages.) In expectation, Young Robert's total welfare is 90. If Robert is saved, then Old Robert has 510 units of expected welfare (450 from adulthood, and 60 in expectation from childhood); Young Teddy has zero, and Old Teddy does not exist. If Teddy is saved, then Old Robert has 110 units of expected welfare; Young Teddy has 90, and Old Teddy has 510. In other words, the *expected* welfare of the four persons is exactly as in the case of Harmful Creation.

What do these calculations mean for prudential rationality? The standard account says that the prudential value of each outcome for Teddy at birth should equal the expected lifetime welfare of Young Teddy. In other words, outcome *A* in which the doctor saves Teddy has value 90, and outcome *B* has value 0. Death would deprive Young Teddy of 90 units of welfare, in expectation. Similarly, death would deprive Old Robert of 400 units of welfare in expectation, and this represents the strength of interest that Robert at 20 has in B over A. So we have successfully reconstructed the values given by the TRI account in section 2.

So much for prudential value; what about ethics? One might guess from the calculations so far that we can just decide Emergency Room by looking at what our standard ethical theory says about Harmful Creation. But that is not quite automatic. The lifetime welfare values given in Harmful Creation coincide with the *expected* lifetime welfare values in Emergency Room. Some ethical theories can reach a verdict based on expected lifetime welfare, but others cannot.

As an example of the first kind, suppose that our axiology is the most common version of total utilitarianism.[36] The value of an outcome with no uncertainty is the total lifetime welfare contained therein; the value of an uncertain outcome is the expected total welfare, in line with expected utility theory. However, *expected total* welfare coincides with *total expected* welfare. (At least, this is true if there is no uncertainty about who exists, or if we may attribute zero welfare to non-existence, as total utilitarians are often happy to do.) On this view, therefore, the verdict in Emergency Room will be parallel to the verdict in Harmful Creation. Namely, we should save Teddy, since that results in 800 rather than 600 units of total expected welfare. This re-emphasises a point made in section 2: the lifetime utilitarian verdict in Emergency Room is formally compatible with the TRI account of prudential

[36] As an example of the second kind, consider the version of average utilitarianism that values an uncertain outcome by the expected average lifetime welfare. In general, expected average lifetime welfare cannot be deduced from the facts about expected lifetime welfare. As it happens, this form of average utilitarianism will still give parallel verdicts in Emergency Room and Harmful Creation. Giving an example that distinguishes them would take me too far afield.

value.

To get the more orthodox verdict, we can adopt complex necessitarianism. Now, I did not explain what complex necessitarianism says in cases of uncertainty. The simplest thing is to appeal to expected value. Let me say a little more, since the details are a bit finicky. Suppose we are choosing between two options A and B, each of uncertain outcome. Model this uncertainty in the following way. There is a set $\mathbb{S}$ of states of nature. Each $s \in S$ bears a probability Pr(s), and each option, like B, assigns to each s a distribution $B(s)$ of lifetime welfare levels over persons. We can then define the value of B relative to A to be the expected value

$$V_{\{A,B\}}(B) = \sum_{s \in \mathbb{S}} \Pr(s) V_{\{A(s),B(s)\}} B(s)$$

where $V_{\{A(s),B(s)\}}B(s)$ is calculated in the earlier, risk-free way. (Of course, we can replace the sum by an integral, as appropriate.)[37]

In my toy model of Emergency Room, each state of nature s corresponds to a length of time after the third birthday at which the transition from one person to another might occur. Effectively, $0 \leq s \leq 12$. For any given value of s, we have the following analogue of Harmful Creation:

[37] The general question of how to combine the principle of neutrality with uncertainty is a vexed one, which I cannot hope to settle here. And though I think that, for the particular *kind* of uncertainty I am presently concerned with, the standard expected value calculation may be right, there is another way of handling uncertainty which seems preferable for more ordinary cases of empirical uncertainty. Let me explain this briefly. I originally gave a formula for $V_{\{A,B\}}(B)$ in terms of the statistics $T_{\text{nec}}(B)$, $N_{\text{con}}(B)$, and $T_{\text{con}}(B)$. In cases of uncertainty, we could use the same formula, but interpret $T_{\text{nec}}(B)$ to be the *expected* total welfare of necessary people, $T_{\text{con}}(B)$ to be the *expected* total welfare of contingent people, and $N_{\text{con}}(B)$ to be the *expected* number of contingent people. The main reason for preferring this method is that it gives a better account of the asymmetry (recall my worry about the permissibility of procreation).

	Outcome	
	$A(s)$	$B(s)$
Young Robert	$30+10s$	$30+10s$
Old Robert	$50+10(12-s)$	$450+10(12-s)$
Young Teddy	$30+10s$	0
Old Teddy	$450+10(12-s)$	

We then have $V_{\{A(s),B(s)\}}(A(s)) = 230+10s$ and $V_{\{A(s),B(s)\}}(B(s)) = 600$ (in either case, the total welfare of the necessary people). Averaging over values of s, we find $V_{\{A,B\}}(A) = 290$ and $V_{\{A,B\}}(B) = 600$, so only B is permissible. We can run a similar analysis in other problem cases, to recover the verdicts recommended in section 3.

Supervenience and Fission

However, the indeterminacy-based account faces serious difficulties. The picture underlying the TRI account is that certain normative facts should be understood in terms of patterns of welfare and psychological connectedness. But the picture of uncertain or indeterminate person-boundaries invokes facts about personal identity, and these facts do not appear to supervene properly on facts about psychology. If we must appeal to irreducible facts about personal identity, then at best we are considering a rival to the TRI account, and a mysterious one at that. Here is one version of the supervenience claim that fails.

(S) If S is psychologically connected to T to the same degree that S' is psychologically connected to T', then S is the same person as T if and only if S' is the same person as T'.

The failure of (S) can already be seen in the case of Methuselah. He changes gradually over time, but the degree to which Methuselah-100 is psychologically connected to Methuselah-101 is the same as the degree to which Methuselah-101 is connected to Methuselah-102, and so on. At least, it seems we can stipulate this. According to (S), either Methuselah-100 is the same person as Methuselah-900, or, for every n, Methuselah-n is a different person from Methuselah-$(n + 1)$. We must deny both of these alternatives.

A natural response to such examples is to observe that (S) is stronger than we need in order to capture the idea that the normative facts depend on the psychological facts and not on additional facts about personal identity over time. For one thing, (S) requires that personal identity facts depend only on facts about *degrees* of psychological connectedness. But there are finer-grained ways to individuate psychological relations between person-stages. To give a crude example, it could be that S and T are connected by virtue of sharing certain memories, while S' and T' are connected by virtue of sharing certain intentions. Then even if the degrees of psychological connection are the same, the psychological facts vary in a way that could reasonably underpin variations in the facts about personal identity.

A more interesting way in which (S) may be unreasonably strong is that, according to the current picture, ethical judgments are guided not by the facts about personal identity *per se*, but rather by the distribution of objective probabilities over different ways the personal identity facts might be resolved. For normative facts to supervene on psychological facts, all we need is that this *distribution* is determined by the fine-grained psychological facts. The resulting picture of personal identity might be interpreted in terms of the theory of metaphysical vagueness developed by Wasserman (2015). On his view, indeterminacy is analogous to indetermin*ism*. He urges that grounding relations can be probabilistic in the same way as causal relations. In our case, the facts about personal identity will be probabilistically grounded in the psychological facts. On Wasserman's view, this is what *constitutes* their indeterminacy.

I find the picture just described interesting and in many ways attractive. But it requires a great deal of further development. We need a more specific theory of how the probability distribution is determined by the psychological facts. There may be many probability distributions satisfying what I called 'the probability hypothesis'. Which one is the one upon which the normative facts depend? At the same time, the probability hypothesis puts non-trivial conditions on the pattern of psychological connectedness, and it is not clear that these conditions must be met.

That second sort of difficulty becomes acute in fission cases. (Unsurprisingly so: it was this sort of case that drove Parfit (1986, ch. 12) to focus on psychological connectedness *instead* of personal identity.) Suppose Albert's brain is surgically divided and the two halves are transplanted into separate bodies, forming new person-stages Lefty and Righty. It is apparently possible for pre-fission Albert to be strongly psychologically connected to both Lefty and Righty – perhaps even *fully* psychologically connected to each of them. On the probability hypothesis, there is in such a case a high probability that Albert and Lefty are the same person, and a high probability that Albert and

Righty are the same person. If these two probabilities are each greater than 1/2, then there must be some probability that Lefty and Righty are the same person. But most people think that Lefty has no prudential stake in Righty’s welfare, and vice versa. They are certainly not the same person.

I find it hard to know what to make of these fission cases.[38] Even if Albert were fully psychologically connected to Lefty and Righty, it is not clear that he would have a full stake in the welfare of each. Many people intuit that he would have only half a stake. So perhaps the lesson is just that there are outré exceptions to the probability hypothesis, and to the claim that I called ‘Psychological Reduction’ in section 2. At the same time, I am not even sure that Lefty has no stake in Righty’s welfare. The situation is such a peculiar one that it is hard to know whether and how the notion of prudential interest applies. It is true that Lefty is not in the position of anticipating or remembering what happens to Righty, in a first-person sense, but they are still related to each other in another and extraordinary way. Alternatively, perhaps Lefty and Righty are the same person, but personal identity is not a sufficient condition for prudential interest. This is a difficult area to sort out, but the obstacles are not yet insurmountable, as far as I can see.

Where does this leave us? The framework of uncertain person-boundaries gives some hope that the different-person heuristic can be maintained with respect to a wide variety of normative theories. It gives at least a toy model

[38] See Ross (2014) for difficulties raised by fission cases for a wide range of theories. Williams (2014b, §3) combines a version of the probability hypothesis with a ‘subvaluational’ theory of indeterminacy to yield some intuitive results in fission cases. However, the non-classical logic associatied with subvaluation does not work well with standard ethical theories. (Williams only considers the prudential case.) Indeed, Williams is trying to interpret Lewis’s theory of overlapping persons, and faces the same kind of over-counting problems I will describe below.

for dealing with degrees of psychological connectedness, and this may be helpful in thinking about examples like Emergency Room. But to turn it into a genuine theory, or even to analyse examples systematically, we need a more complete story about how the person boundaries emerge, however indeterminately, from psychological connections.

Partial Persons

Now I turn, much more briefly, to two other ways in which one might try to ground the different-person heuristic casting other entities in the person-role.

The first depends on the following psychological fairytale. Think of a person, or a person at a time, as being composed of psychological constituents that persist from one time to another. McMahan considers a picture like this in relation to the beginning and end of life:

> It is quite natural to think of the person as going out of existence gradually as the mind is increasingly eroded, with its constituent psychological states and capacities disappearing as the tissues of the brain atrophy and die. (McMahan, 2002, p. 279)

During this phase of erosion, the person only 'partially exists'. If persons can be understood in this way, then perhaps we can conceptualise personal *identity* as a matter of psychological overlap. That is, two person-stages are personally identical to the extent that they have the same psychological constituents.[39] This relation will come in degrees, because the psychological

[39] What does it take for psychological constituents (e.g. states or capacities) of different person-stages to be 'the same'? Should we think of them as types or tokens? I don't know. But if we intuitively grasp what McMahan means when he talks about such constituents

constitutions of two person-stages can overlap to a greater or lesser extent. A mind, on this picture, is like the ship of Theseus. First it is built up plank by plank, and during this phase, there is only part of a ship. Then as time passes, planks slowly get replaced. In one sense, the same ship persists from one time to another. But in another sense – for my purposes the relevant one – the ship that is present at one time is only partially present at another, and after enough time, the ship that there is will be entirely distinct from the original. In the end it is left to rot on the beach.

Can this picture of shared psychological constituents be parlayed into a version of the TRI account? At the level of prudential value, the picture might look like this. A person-stage S will share some portion of T's psychological constituents; on that basis, it may have some proportionate stake in T's welfare. This stake could depend not only on the amount of psychological overlap, but also on what kinds of constituents are shared, and on the relationship between those constituents and T's sources of welfare. In the limiting case when S and T do not overlap at all, then S will have no stake in T's welfare.

What about ethics? My suggestion is that the psychological constituents could play something like the 'person' role in population ethics – or, a little more precisely, the role of *partial* persons. Roughly speaking, a psychological constituent c counts as a fraction of a person in proportion to the prominence it would have in an ordinary adult's psychological constitution. If the above account of prudential value works out, then we can effectively assign to c a 'lifetime welfare': for each person-stage T with constituent c, we can assign

'disappearing', we must also have some grasp on what it means for them to persist.

to c a portion of T's wellbeing.

In contrast to the preceding picture of indeterminate or uncertain personal identity, the current picture suggests a fairly concrete view about how personal identity is grounded in psychological facts. The main problem is its obviously speculative nature. I do not know whether this picture of psychology is really tenable. A narrower problem is again how it might work in fission cases. Albert divides into Lefty and Righty; what should we say about how their psychological constitutions overlap? One view is that all of Lefty's psychological constituents are shared with pre-fission Albert, as are all of Righty's. But if we want to deny that Lefty and Righty psychologically overlap, it must turn out that pre-fission Albert had a rich enough psychology to account for two persons all along.[40] The resulting picture – in fission cases only – bears a close resemblance to David Lewis's picture of overlapping persons, to which I now turn.

Overlapping Persons

Lewis (1976) presented a theory on which personal identity tracks psychological connectedness. That may seem promising for our current purposes. On his theory, persons can overlap as four-dimensional entities, sharing person-stages. Because of this, there is no sense in general in asking whether person-stages S and T belong to *the* same person. Rather, the question is

[40] The analogy with the ship of Theseus may help: how can we take the planks from one ship and build two new ones? Only if the new ships are smaller than the original. What is puzzling in our case is that Lefty and Righty are not intuitively 'smaller' than pre-fission Albert. Note this is different from the fission-like scenario most commonly considered in the literature on the ship of Theseus, in which new planks are brought in (Hobbes, 1655, §11).

whether there is *a* person to whom S and T both belong. On his view, it is really this relation of co-belonging (Lewis called it 'the I-relation') which tracks psychological connectedness. More precisely, if S and T are pyschologically connected to an intermediate degree, then it is indeterminate whether S and T co-belong, but the degree to which it is determinate that they co-belong equals the degree of psychological connectedness.

Is there a way to use Lewis's overlapping persons to make sense of the different-person heuristic, and thus to develop the TRI account? I do not think there is a straightforward way to do so.

Let me start with prudential value. Again, we have the standard person-based theory of prudential value, the lifetime-comparative account. It says that the prudential value of an outcome B to a person-stage S equals the lifetime welfare in B of the person to whom S belongs. This formulation does not sit well with Lewis's picture. One reason is that there is, in general, no single person to whom S belongs. Perhaps we could get around this by some sort of averaging. There is a subtler underlying problem. Consider a case in which, in outcome A, Methuselah lives a full life to 969, while in outcome B, he dies at 500. Let me recall Lewis's characterisation of when a collection of person-stages composes a person. We can say that two person-stages are *strongly connected* if they are psychologically connected to a sufficient degree, surpassing some particular but indeterminate threshold. According to Lewis, a collection of person-stages composes a person if and only if all the person-stages are strongly connected to each other, and the collection is maximal with respect to this property. The problem is that there is then no person in A whose life is cut short in B.

Why not? Suppose, for illustration that the threshold for strong connectedness is such that, at any given time, Methuselah is strongly connected to person-stages up to 100 years into the past and 100 years into the future. So, for example, one person who exists in both outcomes *A* and *B* is the composition of Methuselah-100 through Methuselah-200. What about the composition of Methuselah-450 through Methuselah-550? This is a person in outcome *A*, and superficially his life is cut short in outcome *B*. But in fact he simply does not exist in outcome *B*. In *B* there is the composition of Methuselah-450 through Methuselah-500, but this collection of person-stages does not compose a person. It is not maximal with respect to the condition of strong connectedness. So it turns out that the composition of Methuselah-450 through Methuselah-550 is a *contingent* person.

Because of this, it does not seem to be in any person's interest that Methuselah lives on, unless we admit, against the grain of the principle of neutrality, that contingent people have an interest in coming to exist.

Similar problems arise when we turn to ethics. Again, Lewis's theory looks like it might help us. It delivers a set of persons in each outcome, so, to evaluate that outcome, we could try to aggregate the lifetime welfare of those persons. But it seems wrong to do this in any ordinary way, precisely because the persons overlap. The welfare of a person-stage will be over-counted if it belongs to many different persons, compared to others. For example, the first year of Methuselah's life belongs to only one person, whereas his hundredth year belongs to many. It is implausible that what happens in his hundredth year is enormously more important than what happens in his first year. A second problem is a version of the one I mentioned for prudential value. The outcome B in which Methuselah dies at 500 does not harm any Lewisian person who exists in A. So, given the principle of neutrality, there is no reason to prefer A over B.

Perhaps there is a way to finesse these problems. (The picture I present in the next section could be interpreted in this way.) But at any rate we cannot just plug Lewisian persons into a standard normative theory.

7 Prudential Counterparts

The main view I considered in the previous section claimed that personal identity between person-stages is often indeterminate, and that the degree of determinacy matches the degree of psychological connectedness. The basic problem was spelling out how the facts about personal identity could be determined, probabilistically or not, by the psychological facts. The picture I consider here is a variation on that one, but it appears more promising. I give a counterpart-theoretic treatment of personal identity over time. The key point is that I can adapt some of Lewis's ideas to explain how the counterpart relation supervenes on psychological connections. This gives a huge advantage over the earlier view.

The bad news is that the counterpart relation is not an equivalence relation. As a result, it does not partition person-stages into persons. So there is no general recipe that combines the counterpart relation with a wide variety of ethical views. We cannot simply claim to have identified the normatively relevant 'persons'. Still, in many cases there are some plausible ways to proceed. I will illustrate this with a few examples, including complex necessitarianism. We thereby get a fairly concrete ethical theory that reproduces the orthodox verdicts of the TRI account.

Prudential Value

Let me start with prudential value. Throughout this section I am going to assume that all welfare is a matter of the welfare of person-stages, so that, on the ordinary way of thinking, lifetime welfare is the total welfare of the relevant person-stages. (The need for this assumption is a significant disadvantage of the present approach, in comparison to the approaches in the previous section. But since the TRI account seems to rely on this assumption – see footnote 5 – it may be admissible here.) The standard theory of prudential value therefore says that the value of an outcome B for a person-stage S is the total welfare of the person-stages in B who stand in the relation of personal identity to S.

I claim that we can formulate the TRI account of prudential value in essentially the same way, replacing the relation of personal identity with what I will call *the prudential counterpart relation*. Like personal identity, and unlike psychological connectedness, this will be an on/off relation, not one that comes in degrees. Namely, T is a prudential counterpart of S in B just in case T is strongly connected to S in B. Recall from my discussion of Lewis that this means that the degree to which they are psychologically connected exceeds some particular but indeterminate threshold. In fact, in the terminology I used there, the prudential counterpart relation, strong connectedness, co-belonging, and Lewis's I-relation are all coextensive. The different names suggest different applications, and 'the prudential counterpart relation' suggests that it underpins prudential interest.

Now, at a first pass, the proposal is that the prudential value of B for S is the total welfare of the person-stages in B who are prudential counterparts of S. However, we have to remember that the prudential counterpart relation is indeterminate, because it depends on the threshold for strong connectedness. As in section 6, I suggest that the appropriate response to this kind of indeterminacy is to average over precisifications of the prudential counterpart relation. We could take this averaging as primitive, or, as I prefer, interpret it in terms of objective probability. Indeterminacy rationally demands uncertainty about the value of the threshold for strong connectedness, and thus uncertainty about the extension of the prudential counterpart relation. We will recover the TRI account of prudential value as long as the following condition holds: the objective probability that T is a prudential counterpart of S equals the degree of psychological connectedness between them.

Can we ensure that this condition holds? In Lewis's discussion of the I-relation, he effectively shows us how. Lewis assumes that we can measure degrees of psychological connectedness using numbers between 0 and 1. So a precisification of the threshold for strong connectedness is just given by a number in this interval. Lewis shows that, with respect to the uniform probability measure on the interval, the proportion of precisifications on which T is I-related to S equals the degree to which T is psychologically connected to S. The upshot: the TRI account follows from the hypothesis of *uniform* objective uncertainty about the location of the threshold for strong connectedness.[41]

To summarize the picture so far, let me say that the proportion of precisifications on which T is a prudential counterpart of S is the *degree* to which they are prudential counterparts. (More strictly, though, it is the degree to which it is *determinate* that they are prudential counterparts.) The TRI account says that, to calculate the value of B for S, add up the welfare of every person-stage in B, weighting the welfare of each one by the degree to which it is a prudential counterpart of S. The key difference from the earlier picture of indeterminate personal identity is that the prudential counterpart relation need not be transitive, so it need not divide up what person-stages there are into disjoint groups.[42] As I have said, Lewis interprets the prudential counterpart relation as a relation of co-belonging, which leads to overlapping persons for this very reason. But what I am suggesting does not rely on the idea that prudential value involves harms and benefits to persons. The theory of overlapping persons does not do any work here, and may be a distraction.

Ethical Theories

What if we go beyond prudential value? We can try the same basic strategy. Given a standard theory of population ethics, we can try to reformulate it in terms of person-stages and the personal identity relation, rather than in terms

[41] For Lewis's argument, see Lewis (1976, p. 70) and Williams (2014b, Appendix A). The use of the uniform measure may seem unjustified. But it is natural if we we think that the numerical representation of degrees of psychological connectedness is *defined* by its link to credences about the prudential counterpart relation; cf. footnotes 7 and 34.

[42] Here is an example of the intransitivity. Assuming that the threshold for strong connectedness is not very high, Methuselah at age $n+1$ is a prudential counterpart of Methuselah at age n, for any n between 100 and 900; but Methuselah at age 900 is not a prudential counterpart of Methuselah at age 100.

of persons *per se*. And then we can substitute the prudential counterpart relation for the personal identity relation. I will now illustrate this with a few examples, working my way up to complex necessitarianism.

Utilitarianism

As usual, there is no difficulty with total utilitarianism, on the standing assumption that lifetime welfare is the total welfare of person-stages. The value of an outcome is just the total welfare in that outcome. Personal identity plays no role in the theory, so the substitution is trivial. Average utilitarianism is a harder case. Let us take the most obvious way to apply average utilitarianism in cases of uncertainty: the value of an option is its expected average utility. (This matters insofar as we are going to treat indeterminacy in the prudential counterpart relation as giving rise to objective uncertainty.) The average utility is the total utility divided by the number of persons. Interpreting the total utility is no problem, but what of the number of persons? On the views I considered in the previous section, we had a set of persons (or else parts of persons), and therefore a number of persons. Here the explanation cannot be so simple. But there are still some things to try.

Here is one. To count persons in the standard framework, using personal identity, choose as many person-stages as you can such that no two of them stand in the relation of personal identity. The number you get is the number of persons. But now we can do the same thing, using the prudential counterpart relation. For each outcome, we get a number n, a version of 'the number of persons' based on psychological connectedness. The value of the outcome is the total welfare in it divided by this number n. That is the answer relative to any precisified threshold for strong connectedness. To resolve the indeterminacy in the threshold, use the expected value: average over precisificiations.

There are other ways to count persons. Divide up the person-stages into groups, such that the stages within each group stand pairwise in the personal identity relation. What is the least number of groups we need? That is the number of persons. Or: count how many person-stages there are, weighting each one, S, in inverse proportion to the length of the life to which S belongs. How long is that life? It's the total temporal extension of the person-stages that stand in the relation of personal identity to S. Each of these options suggests a subtly different adaptation of average utilitarianism.

Which way of 'counting persons' is the right way? It depends. Why did you want to average over persons in the first place? Answer that, and you may see what to do. There is no substitute, in the end, for building the theory from the ground up. For now I am content to say that there are *some* basically plausible ways to modify average utilitarianism, as well as other theories, in light of the different-person heuristic.

Prioritarianism

The upshot of prioritarianism is often explained in the following terms: the value of an outcome is given by weighted total lifetime welfare. What is the weight given to the welfare of a each person? That is itself a function g of the person's lifetime welfare. If the lifetime welfare is low, the person gets relatively high weight; if the lifetime welfare is high, the person gets relatively low weight.[43] The story so far assumes there is no uncertainty. When the outcome is uncertain, the most standard move is to use its expected value.

It is easy to reformulate prioritarianism in terms of person-stages and personal identity. The value of an outcome A is the weighted total welfare of the person-stages in it. The weight of a person-stage S is a decreasing function g of the prudential value of A for S. That is, we apply g to the total welfare of the of the person-stages that stand in the relation of personal identity to S.

Now we can formulate the analogous view using the prudential counterpart relation instead of personal identity. Explicitly, the value of A relative to a precisification is the weighted total welfare of the person-stages in it, with the weight of S a decreasing function g of the total welfare of the person-stages that are prudential counterparts of S. At the end, we average over precisifications.[44]

[43] Often the value function of prioritarianism is presented as first transforming each lifetime welfare by an increasing, concave function f, and then adding up the values. In other words, the contribution of a person with welfare w is $f(w)$. But we can usually write $f(w) = g(w)w$, where g is the weight function I mention in the text. (Here I am setting the zero of the welfare scale so that $f(0) = 0$. In other words, I am assuming there is no 'critical level'.)

[44] The result appears to be a version of the 'Prudential Prioritarianism' defended by Holtug (2010, §10.6) – although I have trouble understanding the details of his view. In any case, note that there is an obvious alternative. We could just declare that the weight given to a person-stage S is a decreasing function of the prudential value of A for S as given by the TRI account. This alternative view corresponds to a different ('ex ante') extension of prioritarianism to situations of uncertainty, in which one applies the prioritarian aggregation rule to expected lifetime welfare. I hope to examine these variations in more detail in other work.

Simple Necessitarianism

According to simple necessitarianism, the value of an outcome B in a choice between A and B is the total welfare in B of the necessary people. To put it another way, it is the total welfare of the person-stages in B who stand in the relation of personal identity to person-stages in A. On the new, modified version of simple necessitarianism, the value of B is the total welfare of the person-stages in B who are prudential counterparts of person-stages in A. That's the answer relative to any precisification of the threshold for strong connectedness. To get an answer overall, we can again appeal to expected value.

Complex Necessitarianism

It is not much harder to get a version of *complex* necessitarianism. As with average utilitarianism, the real problem is that there are several ways one could proceed, and it is not clear which of them is most reasonable. But for my current purposes it does not matter much, either. Here is one possibility.

Remember that, in complex necessitarianism, the value of B relative to A was expressed in terms of the three statistics $T_{\text{nec}}(B)$, $T_{\text{con}}(B)$, $N_{\text{con}}(B)$. (On page 221 I discussed how to extend this to cases of uncertainty.) So, first we can redescribe these statistics in terms of person-stages and personal identity. As in the preceding discussion of simple necessitarianism, we can redescribe $T_{\text{nec}}(B)$ as the total welfare of the person-stages in B who are personally identical to stages in A. Similarly, $T_{\text{con}}(B)$ is the total welfare of the person-stages in B who are *not* personally identical to stages in A. Finally, we can redescribe the number $N_{\text{con}}(B)$ of contingent people in any one of several ways, just as in the case of average utilitarianism. For example, we can identify it as the largest number of person-stages in B, no one of which is personally identical to a stage in A and no two of which are personally identical to each other.

Finally, we can modify complex necessitarianism in light of the different-person heuristic. Use the same formula to define the value function $V_{\{A,B\}}$, but define T_{nec}, T_{con}, and N_{con} in terms of the prudential counterpart relation instead of personal identity.

It may be objected that the theory I have just sketched is extremely ad hoc, and that is true. One big source of adhockery is complex necessitarianism itself. I have not properly justified the way it handles the non-identity problem and (especially) the asymmetry. Another source of adhockery is the way I defined a proxy for 'the number of contingent people'. The best way to think through that latter issue is presumably to revisit the original counterpart-theoretic explanation of complex necessitarianism, on page 210.[45] But getting the best possible version of the view is not my ambition here. For now, the conclusion is just that there is no fundamental barrier to making sense of the orthodox verdicts of the TRI account, modulo the admittedly formidable problems of population ethics. For these purposes, we can treat the modified version of complex necessitarianism as a useful model. It is internally consistent, it gets the main heuristics right, it is not too complicated,

[45] The idea would be to balance the welfare of the person-stages in A against the welfare of those in B, by averaging over counterpart relations between person-stages. These counterpart relations should satisfy certain constaints. As an initial proposal, they should (i) extend the relation of transworld identity; (ii) respect the prudential counterpart relation; (iii) pair as many person-stages as possible subject to (i) and (ii).

and perhaps it is not too far from the truth. It can serve as a guiding light in the quite general project of building a credible moral theory from the ground up.

8 An Alternative to the Orthodox Account

I've argued that complex necessitarianism can give a reasonable version of the principle of neutrality, with intuition-friendly consequences for the non-identity problem and the asymmetry. Moreover, it can formally be adapted in light of the different-person heuristic to give a coherent normative theory encapsulating the orthodox judgments of the TRI account.

But are those judgments right? The confusion that arose from the problem-cases in section 3 suggests that those initial verdicts were under-theorised. For example, in Emergency Room, Choice of Deaths, and Suffering Now, the suggested ethical verdicts were inspired by considerations of the present time-relative interests of the patients involved. But we *also* know that present interests are not the only ones that matter; why should they be decisive in these cases?

One of the advantages of the population-ethics-first strategy developed in this paper is that it invites us to balance these initial verdicts against considerations, both intuitive and theoretical, arising from population ethics itself. We see this in the case of the non-identity problem and, perhaps more especially, in the case of the asymmetry. The need to provide a theoretical justification for the asymmetry is so pressing that a plausible solution may lead us to revise some of our initial verdicts, especially in examples which, on a second look, do not appear so clear.[46]

I claim that Emergency Room may be such an example.

In this section I present a theory of population ethics that is both formally simpler than complex necessitarianism and offers a conceptually more natural treatment of the asymmetry. It also resonates with the actualist strain in McMahan's discussion of the TRI account, without the fatal modal variance that actualism entails. The theory, which I call *regret minimization*, uses the same ingredients as complex necessitarianism, and so can be adapted to the different-person heuristic in essentially the same way. In short, it speaks to many of the core preoccupations of the TRI account, and deserves serious consideration from those who share in them. The reason I started with complex necessitarianism rather than regret minimization is that the latter gives an unorthodox answer in Emergency Room, and in a few other cases. It says that it is permissible to save either the newborn or the adult.

Let me begin by explaining the view, and its approach to the asymmetry. Then I will go on to argue that its departures from orthodoxy should be welcomed by proponents of the TRI account. Whether or not regret minimization is on the right track, the reader should bear in mind the main point I am using it to illustrate: by situating the TRI account in the context of population ethics, we can draw on a wider range of theoretical and intuitive considerations in developing and evaluating the view.

[46] To say that justifying the asymmetry is a pressing need is not to say that the asymmetry is untouchable. Rather, given the indisputable prevalence of asymmetric intuitions, we should ask how they can best be justified, and evaluate them in that context.

Regret Minimization and the Asymmetry

I know of essentially two distinct ways of arriving at the asymmetry. One is to start, as I did in discussing necessitarianism, from the idea that bringing someone into existence neither harms nor benefits them. At this stage we have no reason to create good lives – this is half of the asymmetry – but we also have no reason not to create bad ones. The reason not to create bad lives must be reached in a different way, either by positing an additional impersonal disvalue to suffering, or by identifying some sort of deontic constraint. The asymmetry arises because this additional ingredient operates in terms of bad lives but not in terms of good ones.

The alternative, which I prefer, is to say that someone who has a good life benefits from it, and someone who has a bad life suffers from it – but if the relevant person does not exist, then there is no one who benefits or suffers. If a person X has a good life in A and not in B, and that is the only difference between these options, then in B there is no one who has a claim on the wellbeing that X has in A. In that sense, failing to create a person with a good life should carry with it no regret. Creating a person with a good life carries no regret, either. Quite the opposite: doing so creates a person who benefits from being alive. The act of creating a good life will only be regrettable if the harms to others are greater than the benefit that accrues to the new person. The one regrettable scenario, in a simple choice between creating and not creating, is the scenario in which one creates a person with a bad life. For then there actually is someone who suffers from the choice.

According to regret minimization, that is the one scenario to be avoided.[47]

Note how this account automatically produces a *moderate* asymmetry, in the sense that it can be permissible to create many lives even if some of them are bad ones. For such an act creates some people who benefit from existence, and others who suffer from it; there is only cause to regret the situation overall if the suffering outweighs the benefits. By the same token, creating good lives can be permissible even if it harms pre-existing people, as in Harmful Creation. To see if such an act is regrettable overall, we must balance those harms against the benefits conferred on the people who come to exist. I will say more about this below, when I consider in detail how regret minimization revises the orthodox verdicts of the TRI account.

To arrive at the proper formulation of regret minimization, we must try to understand the preceding comments in a way that overcomes the non-identity problem. In the case of Adam and Steve, we must recognise a sense in which creating Eve would be regrettable. There is a narrow sense in which it would not be regrettable, since Eve's life is not a bad one, and the alternative for her is non-existence. But, in a broader sense, we should recognize Steve in A as the counterpart of Eve in C. Creating Eve is regrettable because the extra person, *de dicto*, could have been better off. How to make sense of this in more complicated situations – in particular, when the relevant

[47] This basic way of justifying the asymmetry was suggested to me by Daniel Cohen. Note that the asymmetry enters in through the focus on regret-minimization as opposed to contentment-maximization (supposing that 'contentment' is the opposite of 'regret'). Other proposals in the same family are found in Roberts (2011a,b), and in the so-called 'harm-minimization' views going back to McDermott (1982); see also Meacham (2012); Ross (2015) for recent developments. In those views, the asymmetry arises through the focus on harm-minimization rather than benefit-maximization. In developing regret minimization I have reinterpreted a view that Cohen develops in unpublished work, and modified it to overcome the non-identity problem.

options contain different numbers of contingent people – is a delicate matter. The particular form of regret minimization I describe below is motivated by averaging over counterpart relations, in the way I explained in my discussion of complex necessitarianism. At any rate, in this respect regret minimization is certainly no worse off than complex necessitarianism and other similar theories.

To conclude this preliminary discussion, let me contrast regret minimization with 'actualist' views. On the simplest version of actualism, each option is evaluated on the basis of the welfare of the people who actually exist. So in Harmful Creation, if if outcome B actually occurs, then outcome A is to be evaluated only on the basis of the welfare of Young Robert, Old Robert, and Young Teddy; outcome A is worse than outcome B, and therefore impermissible. McMahan often appears to appeal to this sort of thinking, but it leads to the serious problems I have pointed out before, in sections 3 and 5.

Regret minimization in many ways respects the actualist thought that non-actual people are morally irrelevant. But it does so in a more plausible way, avoiding the problem of normative variance. In evaluating how regrettable each outcome would be, it will consider only the welfare of the people who *would* exist, *were* that outcome actual. This is another reason to consider regret minimization in this context. It allows one to make sense of the actualist strain in McMahan's development of the TRI account.

With all that preliminary, regret minimization is easy to state. The theory has the following structure. For each pair of options A and B, there is a value of B for the people in A, denoted $V_A(B)$. Given a set $\mathscr{S}$ of available options, the *regret* of A, $\mathrm{Reg}(A)$, is the extent to which A falls short of being the best for its own people:

$$\mathrm{Reg}(A) = \min_{B \in \mathscr{S}}(V_A(B) - V_A(A)).$$

Finally, an option is permissible if and only if it minimizes regret.[48]

It remains to define the value $V_A(B)$. The idea here is to choose counterparts in B for as many people in A as possible, in a way that extends transworld identity. (That is, if X exists in both A and B, then X must be its own counterpart.) Then $V_A(B)$ is just the total welfare of the people in B who are counterparts of people in A, averaged over all different ways of choosing the counterpart relation.

The upshot of this is that, if there are fewer contingent people in B than in A, $V_A(B)$ is the total welfare in B; if there are more contingent people in B than in A, $V_A(B)$ is the total welfare of the necessary people in B, plus the total welfare that the contingent people in A would have if they had the same average welfare as the contingent people in B. In symbols:

$$V_A(B) = \begin{cases} T_{\mathrm{nec}}(B) + T_{\mathrm{con}}(B) & \text{if } N_{\mathrm{con}}(B) \leq N_{\mathrm{con}}(A) \\ T_{\mathrm{nec}}(B) + T_{\mathrm{con}}(B)\frac{N_{\mathrm{con}}(A)}{N_{\mathrm{con}}(B)} & \text{if } N_{\mathrm{con}}(B) > N_{\mathrm{con}}(A). \end{cases}$$

This is the 'standard' version of regret minimization, which takes persons as basic. But it is simple enough to reformulate it in terms of person-stages and personal identity, exactly as I did for complex necessitarianism. We can then replace personal identity by the prudential counterpart relation, and finally

[48] Alternatively, we could use this characterisation of permissibility in two-option cases, and then extend the rule to multiple options using Schwartz's method mentioned above. I do not know which of these routes is more reasonable.

resolve indeterminacy by averaging over precisifications.[49] We thereby obtain a version of regret minimization founded on psychological connectedness, and vindicating the different-person heuristic.

Now that we have a precise statement of the view, let me conclude simply by revisiting the asymmetrical case of Long-Suffering Eve. In that example, the regret of A is 1 (Adam, the only person in A, could have been better off) but the regret of B is 9 (Adam is better off than he might have been, but Eve need not have suffered). This seems to me to capture particularly well the intuitions behind the asymmetry.[50]

Embracing Heterodoxy

Regret minimization entails certain amendments to the orthodox verdicts of the TRI account. I conclude by arguing that these amendments ought to be embraced.

Consider again a case like Harmful Creation, which is related to Emergency Room through the different-person heuristic. When is it permissible to harm a person X (or to withhold a benefit) while creating a second person Y? Complex necessitarianism says that it is never permissible to do so, for the harm to X cannot be compensated by any welfare that the new person might have. In contrast, regret minimization says that it is permissible to do so as

[49] An alternative, as in fn. 37, is to interpret the statistics T_{nec}, T_{con}, and N_{con} as expected values.

[50] With regard to the standard problems of population ethics, regret minimisation moderates the problem that complex necessitarianism had with the sadistic conclusion (see fn. 30): it says that either option (creating the Xs and Ys or creating the Ps and the Qs) is permissible. On the other hand, unlike complex necessitarianism, regret minimization also controversially rules that either option is permissible in a choice like that of the 'repugnant conclusion'. I think both of these verdicts are defensible, but I will not defend them here.

long as the lifetime welfare of the new person is greater than the harm to X. But (in line with the principle of neutrality) it is never *obligatory* to create the new person. We can explain this in McMahan's terms of 'offsetting weight'. Just as McMahan (2015) claims that the creation of good lives can offset the reason we have to avoid creating bad lives, so too, on regret minimization, the creation of good lives can offset the reason we have to avoid harming necessary people, even though we never have a positive reason to create those lives. I can only say that this seems a matter of consistency: it is hard to see why good contingent lives would have offsetting weight in one way but not in another.

But what about the intuition in Emergency Room? In McMahan's original version of the case, the choice is between saving a foetus at the cost of its mother's life, or saving the mother only. At a first look, McMahan may seem right to say

> Most of us believe that the doctor ought not to save the fetus. And the basis for this belief is our sense that the death of the fetus would be less bad – that it would be the less serious of the two possible misfortunes. (McMahan, 2002, p. 186)

Although I get the intuition mentioned in the first sentence of this passage, I am much less confident than McMahan is of the explanation offered in the second. There are all sorts of confounding factors. For example, in McMahan's version of the case, it is hard to avoid thinking about the special relationship between mother and child, both during gestation and afterwards. There could be many factors contributing to the intuition here, including the propensity to relate to the adult as a friend, at whose death we ourselves would feel more significant grief. As Greaves (MSa) argues, focusing on our sense of misfortune is suspect; there can be no general principle of minimizing 'tragedy', another word McMahan frequently employs. It may be the case that our sense of tragedy tracks the present time-relative interests of the patients in cases like these. But, if so, that fact has no great moral authority.

It also seems useful to think through the case retrospectively. Suppose the foetus is saved. Despite the loss of its mother it grows up to have a rich and meaningful human life. Once I focus on the details of that outcome, it becomes repugnant to me to think, 'It would be better if his mother had not died and he had never existed'. As unfortunate as her death may have been, in this retrospective evaluation, the fact of the child's flourishing does intuitively offset the misfortune of his mother's death, even though, as a foetus, he had no interest in that flourishing. Now, to focus *exclusively* on this retrospective evaluation risks falling prey to the modal variance of actualism. But there may still be *something* legitimate about it, which is nicely reflected by my technical notion of regret.

Consider also Suffering Now, and in particular the choice between B (Tommy's death) and C (one year of suffering followed by many years of good life). The orthodox judgment of the TRI account is that B is preferable to C. McMahan himself admits that this judgment may strike many as 'implausible' (p. 72). But from the point of view of the different-person heuristic, this is yet another case of harmful creation. The present-bound animal is harmed and this harm is not (on McMahan's interpretation) offset by the creation of future happy animal-stages. It is hard to see how to deny this counter-intuitive verdict in Suffering Now while accepting the orthodox verdict in Emergency Room. This should at least cast some doubt of the significance of the intuitions in these cases. Regret minimization goes for the middle ground here: it will say that these cases are parallel, and that either option in them is permissible.[51]

In short, I am skeptical of the orthodox intuition in Emergency Room, and of the related judgment in Harmful Creation. But, much like psychological connectedness, heterodoxy comes in degrees. Denying the orthodoxy in these specific cases need not involve a significant deviation from McMahan's vision of the TRI account. First of all, it does nothing to undercut the basic account of prudential value. Nor does it require us to give up on the different-person heuristic. Finally, it does not require us to give up on the principle of neutrality, which, I have suggested, is the basis for the broad ethical outlook associated with the TRI account.

Let me illustrate this with two examples in which McMahan applies the orthodox TRI account, and in which it might seem that the orthodox judgment in Harmful Creation plays a crucial role.

The first application is in support of a liberal attitude towards abortion. To this end, it is important to avoid the result in Emergency Room that we *must* save the foetus or newborn, and especially to avoid that result in similar cases where the potential harm to the adult is more modest. Regret minimization gets us this far, since it will typically say that either option is

[51] As this shows, regret minimization is a fairly permissive view. This goes against the claim in Broome (2004, p. 169) that the neutrality of adding lives should not not be 'greedy'. I disagree with him about whether having greediness at all is counterintuitive, but it is a fair question whether regret minimization ends up being too permissive.

permissible, as it does in Harmful Creation. It would be significantly stronger – I think too strong – to conclude on the basis of time-relative interests that saving the adult is *obligatory*. (It may be so for other reasons related to the adult's moral and legal status.) At any rate the stronger conclusion is not necessary to justify a liberal attitude.

The second example is more complicated. It concerns McMahan's project to explain how the TRI account can justify the intuition that, for low-level animals, suffering is worse than death. He introduces that intuition in terms of a

> version of humane omnivorism [which] implies that while it would be permissible to deprive an animal of many years of life as a means of providing each of twenty people a certain amount of pleasure, it would not be permissible to cause that animal more than a small amount of suffering as a means of providing the same pleasure to the same people.
>
> According to this version of humane omnivorism, therefore, it is worse to cause an animal to experience a small amount of suffering than it is to kill the animal, even when killing it would deprive it of many years of life without significant suffering, so that the good life it would lose would be good for it by much more than the suffering would be bad for it. (McMahan, 2015, pp. 65–66).

The switch here from permissibility in the first paragraph to axiology in the second is confusing. But later in the article, McMahan effectively takes the conclusion expressed in the second paragraph to mean that the animal should be killed in Suffering Now. Since he takes the TRI account to deliver that verdict in Suffering Now, he concludes that the TRI account supports the intuition that 'suffering is worse'. On the face of it, it seems that overturning the orthodox judgment in Suffering Now would throw a spanner in the works.

However, understood as an argument for a specific verdict in Suffering Now, the argument from humane omnivorism is invalid. There are really two problems with it. First, the argument has the following form: between A and B, A is permissible; between B and C, B is obligatory; therefore, between A and C, A is obligatory. But there are many plausible counterexamples to this form of argument, with the best-known ones arising in cases where some of the options are incommensurable, or on a par. Suppose that B is better than C, but B and C are both incommensurable with A. Then the most common view is that the pattern of permissibility is exactly the one just described.[52]

What if we modified the original premisses to claim that, on humane omnivorism, between killing the animal and failing to benefit the humans, killing the animal is obligatory (assuming, of course, that these are the only options)? Then the argument would be as follows: A is preferable to B, and B is preferable to C, so A is preferable to C. This form of argument is more widely considered valid, *but* we have also seen in section 3 that it is not valid according to the orthodox judgments of the TRI account. Even with the strengthened premiss, we cannot rely on the provided argument from

[52] See, for example, Rabinowicz (2012) for a recent systematic development of this view in terms of permissible preferences.

humane omnivorism to a specific answer in Suffering Now.

Because of this, overturning the implausible but orthodox verdict in Suffering Now does not necessarily undermine the TRI account's support for humane omnivorism.[53] Indeed, consider the following pure population-ethics case.

> **The Bobs** (Table 11)
>
> In the status quo (outcome *B*), Hannibal will have a good life, and there will be a long series of people called Bob, with lives that are about half as good as Hannibal's. We face one of two choices. In the first choice, we have the option (*A*) of bestowing a small benefit on Hannibal while preventing the existence of all but the first Bob. In the other choice, we has the option (*C*) of bestowing the same small benefit while harming the first Bob to a greater degree. The total welfare that the other Bobs would have, if they existed, is greater in magnitude that the possible harm to the first Bob in *C*.

Hannibal is the analogue of the omnivore who can benefit by either killing or hurting an animal. The Bobs are analogous to different life stages of the same animal, whose psychological connections extend only briefly into the future and past. Thus painlessly killing the animal is analogous to preventing the existence of some of the Bobs, and harming the animal at a given time

[53] Here I note that the practical import of the omnivore's position depends on the details of the case: killing the animal will frustrate its time-relative interests, even if these are fairly modest, and this must be weighed against the arguably even smaller benefit to humans of eating the meat. What we are discussing is the structure of the view, rather than the how the details would realistically weigh up.

	Outcome		
	A	B	C
Hannibal	15	10	15
Bob_1	5	5	–5
Bob_2		5	5
Bob_3		5	5
Bob_4		5	5
Bob_5		5	5

Table 11: The Bobs

is analogous to harming one of the Bobs. The *humane* omnivore's view corresponds to the verdicts that in the first choice between A and B, A is at least permissible, while in the second choice between B and C, C is obligatory. These are indeed the verdicts of regret minimization. So regret minimization still supports humane omnivorism. Note, by the way, that the choice between A and B is similar to the one in Harmful Creation. The orthodox verdict, then, is that between A and B, A is obligatory – which seems to support omnivorism too strongly!

To sum up, regret minimization broadly respects the ethical views associated with the TRI account, while offering plausible amendments in key cases, including the paradigmatic but dubious case of Emergency Room. It has the advantage of giving a better answer in the case of Suffering Now. Even if regret minimization is not quite right, there seem to be good reasons for revising the orthodox judgments of the TRI account in this way, especially if doing so allows for a theoretically plausible treatment of the asymmetry.

9 Conclusion

The TRI account is founded on the view that what matters is not personal identity but psychological connectedness, and this comes in degrees. For prudential value this has fairly clear implications (at least if we set aside fission and other outré circumstances). What it means for ethics is less clear. The orthodox ethical verdicts of the TRI account entail (in the special case of all-or-nothing psychological connectedness) specific popular but controversial commitments in population ethics. The strategy I have proposed in this paper is to work backwards from these commitments to reconstruct a concrete version of the view. I considered several ways of doing this; the most promising seems to be a theory of prudential counterparts. I illustrated this with complex necessitarianism, whose verdicts correspond very closely to the orthodox verdicts of the TRI account. But I have also presented a theory of population ethics – and thus indirectly a new way of extending the TRI account into the ethical domain – that does a better job of justifying some of the key population-ethical commitments, including the asymmetry. If this amendment is on the right track, it illustrates the utility of starting from population ethics and working backwards, rather than trying to construct a theory directly in terms of time-relative interests.

Bibliography

Arrhenius, G. (2000a). *Future Generations: A Challenge for Moral Theory*. Uppsala: University Printers.

Arrhenius, G. (2000b). An impossibility theorem for welfarist axiologies. *Economics and Philosophy*, 16(02):247–266.

Arrhenius, G. (2003). The very repugnant conclusion. In Segerberg, K. and Sliwinski, R., editors, *Logic, Law, Morality: Thirteen Essays in Practical Philosophy in Honour of Lennart Aqvist*, Uppsala Philosophical Studies. Uppsala University.

Arrhenius, G. (2009). One more axiological impossibility theorem. *Logic, Ethics, and All That Jazz. Uppsala Philosophical Studies*, 57.

Arrhenius, G. (2011). The impossibility of a satisfactory population ethics. In Dzhafarov, E. N. and Perry, L., editors, *Descriptive and Normative Approaches to Human Behavior*, pages 1–26. Singapore: World Scientific Publishing Co.

Arrhenius, G. (2013). *Population Ethics*. Manuscript.

Arrhenius, G. and Rabinowicz, W. (2005). Millian Superiorities. *Utilitas*, 17(2):127–146.

Arrhenius, G. and Rabinowicz, W. (2015). The value of existence. In Hirose, I. and Olson, J., editors, *The Oxford Handbook of Value Theory*. Oxford University Press.

Bacon, A. (2015). *Vagueness and Thought*. Draft of 15 January 2015; available from http://www-bcf.usc.edu/∼abacon/.

Bader, R. M. (2014). Neutrality and conditional goodness. Manuscript.

Bales, A., Cohen, D., and Handfield, T. (2014). Decision theory for agents with incomplete preferences. *Australasian Journal of Philosophy*, 92(3):453–470.

Barnes, E. and Williams, J. R. G. (2011). A theory of metaphysical indeterminacy. In Bennett, K. and Zimmerman, D. W., editors, *Oxford Studies in Metaphysics volume 6*, pages 103–148. Oxford University Press.

Barry, B. (1977). Rawls on average and total utility: a comment. *Philosophical Studies*, 31(5):317–325.

Barry, B. (1989). *Theories of Justice*, volume 1. Univ of California Press.

Blackorby, C., Bossert, W., and Donaldson, D. (1995). Intertemporal Population Ethics: Critical-Level Utilitarian Principles. *Econometrica*, 63(6):1303–1320.

Blackorby, C., Bossert, W., and Donaldson, D. (2005). *Population Issues in Social-Choice Theory, Welfare Economics and Ethics*. Cambridge University Press, New York.

Bostrom, N. (2011). Infinite ethics. *Analysis and Metaphysics*, 10:9–59.

Broome, J. (1991). *Weighing Goods: Equality, Uncertainty and Time*. Wiley-Blackwell.

Broome, J. (2004). *Weighing Lives*. Oxford University Press.

Buchak, L. (2013). *Risk and rationality*. Oxford University Press.

Bueno, O. and Colyvan, M. (2012). Just what is vagueness? *Ratio*, 25(1):19–33.

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(02):283–306.

Chang, R. (2002). The possibility of parity. *Ethics*, 112:659–688.

Constantinescu, C. (2014). Moral vagueness: A dilemma for non-naturalism. In Shafer-Landau, R., editor, *Oxford Studies in Metaethics, Vol. 9*, pages 152–185. Oxford University Press.

DeRose, K. (2008). Gradable adjectives: A defence of pluralism. *Australasian Journal of Philosophy*, 86(1):141–160.

Diamond, P. A. (1967). Cardinal welfare, individualistic ethics, and interpersonal comparison of utility: Comment. *The Journal of Political Economy*, 75(5):765–6.

Dougherty, T. (2014). Vague value. *Philosophy and Phenomenological Research*, 89(2):352–372.

Dunaway, B. (2016). Ethical vagueness and practical reasoning. *The Philosophical Quarterly*. Forthcoming.

Edgington, D. (1996). Vagueness by degrees. In Keefe, R. and Smith, P., editors, *Vagueness: A Reader*. MIT Press.

Field, H. (2000). Indeterminacy, degree of belief, and excluded middle. *Noûs*, 34(1):1–30.

Greaves, H. (MSa). Against 'the badness of death'. In Gamlund, E. and Solberg, C. T., editors, *Saving Lives from the Badness of Death*. Oxford University Press. Forthcoming.

Greaves, H. (MSb). A reconsideration of the Harsanyi-Sen-Weymark debate on utilitarianism. *Utilitas*. Forthcoming.

Greaves, H. and Lederman, H. (MS). Extended preferences and interpersonal comparisons of well-being. *Philosophy and Phenomenological Research*. Forthcoming.

Hare, C. (2010). Take the sugar. *Analysis*, 70(2):237–247.

Harman, E. (2011). The moral significance of animal pain and animal death. In L., B. T. and Frey, R. G., editors, *The Oxford Handbook of Animal Ethics*, pages 726–737. Oxford University Press.

Harsanyi, J. (1953). Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy*, 61.

Harsanyi, J. (1976). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. In *Essays on Ethics, Social Behavior, and Scientific Explanation*, volume 12 of *Theory and Decision Library*, pages 6–23. Springer Netherlands.

Harsanyi, J. (1977). Morality and the theory of rational behavior. *Social Research*, 44(4):623–656.

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63(4):309–321.

Harsanyi, J. C. and Rawls, J. (1975). Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory. *The American Political Science Review*, 69(2):594–606.

Hobbes, T. (1655). *Elementorum Philosophiae Sectio Prima De Corpore*.

Holtug, N. (2010). *Persons, Interests, and Justice*. Oxford University Press.

Holtug, N. (2011). Killing and the time-relative interest account. *The Journal of Ethics*, 15(3):169–189.

Huemer, M. (2008). In Defence of Repugnance. *Mind*, 117(468):899–933.

Hurka, T. (1983). Value and population size. *Ethics*, pages 496–507.

Hurka, T. M. (1982a). Average utilitarianisms. *Analysis*, 42(2):65–69.

Hurka, T. M. (1982b). More average utilitarianisms. *Analysis*, 42(3):115–119.

Kamp, J. A. W. (1975). Two theories about adjectives. In *Formal semantics of Natural Language*, pages 123–155. Cambridge University Press. Cambridge Books Online.

Kath, R. (2016). *Shortfall Utilitarianism: A Theory for Variable Population Decisions*. PhD thesis, University of Sydney.

Kavka, G. S. (1975). Rawls on average and total utility. *Philosophical Studies*, 27(4):237–253.

Keefe, R. (2000). *Theories of Vagueness*. Cambridge University Press.

Keefe, R. (2015). Prefaces, sorites, and guides to reasoning. In Walters, L. and Hawthorne, J., editors, *Conditionals, Probability, and Paradox: Themes from the Philosophy of Dorothy Edgington*. Oxford University Press. Forthcoming.

Kitcher, P. (2000). Parfit's puzzle. *Noûs*, 34(4):550–577.

Knapp, C. (2007). Trading quality for quantity. *Journal of Philosophical Research*, 32(1):211–233.

Lewis, C. I. (1946). *An Analysis of Knowledge and Valuation*. Open Court.

Lewis, D. (1976). Survival and identity. In Rorty, A. O., editor, *The Identities of Persons*, pages 17–40. University of California Press.

Lewis, D. (1980). A subjectivist's guide to objective chance. In Jeffrey, R. C., editor, *Studies in Inductive Logic and Probability*, pages 83–132. University of California Press.

McCarthy, D. (2015). Distributive equality. *Mind*, 124:1045–1109.

McCarthy, D. (MS). The priority view. *Philosophy and Economics*. Forthcoming.

McCarthy, D., Mikkola, K., and Thomas, T. (2016). Utilitarianism with and without expected utility. Preprint. https://mpra.ub.uni-muenchen.de/72931/.

McDermott, M. (1982). Utility and population. *Philosophical Studies*, 42(2):163–177.

McMahan, J. (1981). Problems of population theory. *Ethics*, 92(1):96–127.

McMahan, J. (2002). *The Ethics of Killing: Problems at the Margins of Life*. Oxford University Press.

McMahan, J. (2015). The comparative badness for animals of suffering and death. In Višak, T. and Garner, R., editors, *The Ethics of Killing Animals*. Oxford University Press.

Meacham, C. J. (2012). Person-affecting views and saturating counterpart relations. *Philosophical Studies*, 158(2):257–287.

Mill, J. S. (1863). *Utilitarianism*.

Millum, J. (2015). Age and death: A defence of gradualism. *Utilitas*, 27(03):279–297.

Mongin, P. (1994). Harsanyi's aggregation theorem: multi-profile version and unsettled questions. *Social Choice and Welfare*, 11(4):331–354.

Mongin, P. (2001). The impartial observer theorem of social ethics. *Economics and Philosophy*, 17(02):147–179.

Moss, S. (2016). Time-slice epistemology and action under indeterminacy. *Oxford Studies in Epistemology*, 5. Forthcoming.

Mulgan, T. (2002). The Reverse Repugnant Conclusion. *Utilitas*, 14(03):360–364.

Myerson, R. B. (1981). Utilitarianism, egalitarianism, and the timing effect in social choice problems. *Econometrica: Journal of the Econometric Society*, pages 883–897.

Narveson, J. (1973). Moral problems of population. *The Monist*, 57(1):62–86.

Ng, Y.-K. (1983). Some broader issues of social choice. In Pattanaik, P. and Salles, M., editors, *Social Choice and Welfare*. North-Holland Publishing Company.

Ng, Y.-K. (1989). What should we do about future generations? *Economics and Philosophy*, 5(02):235–253.

Parfit, D. (1986). *Reasons and Persons*. Oxford University Press.

Parfit, D. (1997). Equality and priority. *Ratio*, 10(3):202–221.

Parfit, D. (2004). Overpopulation and the quality of life. In Ryberg, J. and Tännsjö, T., editors, *The Repugnant Conclusion*, pages 7–22. Kluwer Academic Publishers.

Parfit, D. (2012). Another defence of the priority view. *Utilitas*, 24(03):399–440.

Parfit, D. (2016). Can we avoid the repugnant conclusion? *Theoria*, 82(2):110–127.

Pivato, M. (2014). Additive representation of separable preferences over infinite products. *Theory and Decision*, 77(1):31–83.

Qizilbash, M. (2005). Transitivity and vagueness. *Economics and Philosophy*, 21(1):109–131.

Rabinowicz, W. (2012). Value relations revisited. *Economics and Philosophy*, 28(2):133–164.

Rabinowicz, W. (MS). Incommensurability meets risk.

Rachels, S. (2004). Repugnance or intransitivity: a repugnant but forced choice. In *The Repugnant Conclusion*, pages 163–186. Springer.

Rawls, J. (1999/1971). *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, Mass.

Rinard, S. (2015). A decision theory for imprecise probabilities. *Philosophers' Imprint*, 15(7).

Roberts, M. A. (2011a). The Asymmetry: A Solution. *Theoria*, 77(4):333–367.

Roberts, M. A. (2011b). An Asymmetry in the Ethics of Procreation. *Philosophy Compass*, 6(11):765–776.

Ross, J. (2014). Divided we fall. *Philosophical Perspectives*, 28(1):222–262.

Ross, J. (2015). Rethinking the person-affecting principle. *Journal of Moral Philosophy*, 12(4):428–461.

Schiffer, S. (2000). Vagueness and partial belief. *Noûs*, 34(s1):220–257.

Schiffer, S. (2002). Moral realism and indeterminacy. *Noûs*, 36(s1):286–304.

Schoenfield, M. (2014). Decision making in the face of parity. *Philosophical Perspectives*, 28(1):263–277.

Schoenfield, M. (2015). Moral vagueness is ontic vagueness. *Ethics*, 126(2):257–282.

Schwartz, T. (1972). Rationality and the myth of the maximum. *Noûs*, 6(2):97–117.

Sen, A. (1977). Non-linear social welfare functions: A reply to Professor Harsanyi. In Butts, R. E. and Hintikka, J., editors, *Foundational Problems in the Special Sciences: Part Two of the Proceedings of the Fifth International*

Congress of Logic, Methodology and Philosophy of Science, London, Ontario, Canada, 1975, pages 297–302. Springer Netherlands, Dordrecht.

Shafer-Landau, R. (1995). Vagueness, borderline cases and moral realism. *American Philosophical Quarterly*, 32(1):83–96.

Sider, T. R. (1991). Might theory X be a theory of diminishing marginal value? *Analysis*, 51(4):265–271.

Smith, N. J. J. (2013). *Vagueness and Degrees of Truth*. Oxford University Press.

Tännsjö, T. (2002). Why We Ought to Accept the Repugnant Conclusion. *Utilitas*, 14(03):339–359.

Temkin, L. S. (2012). *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. Oxford University Press.

Varzi, A. C. (2007). Supervaluationism and its logics. *Mind*, 116(463):633–675.

Vickrey, W. (1945). Measuring marginal utility by reactions to risk. *Econometrica*, 13(4):319–333.

Vickrey, W. (1960). Utility, strategy, and social decision rules. *The Quarterly Journal of Economics*, 74(4):507–535.

Von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Science Editions. Princeton University Press.

Voorhoeve, A. and Fleurbaey, M. (2016). Priority or equality for possible people? *Ethics*, 126(July):929–954.

Wasserman, R. (2015). Vagueness and the laws of metaphysics. *Philosophy and Phenomenological Research*, 93(2).

Weymark, J. A. (1991). A reconsideration of the harsanyi–sen debate on utilitarianism. In Elster, J. and Roemer, J. E., editors, *Interpersonal comparisons of well-being*, pages 255–320. Cambridge University Press, Cambridge.

Williams, J. R. G. (2011). Degree supervaluational logic. *Review of Symbolic Logic*, 4(1):130–149.

Williams, J. R. G. (2012). Indeterminacy and normative silence. *Analysis*, 72(2):217–225.

Williams, J. R. G. (2014a). Decision-making under indeterminacy. *Philosophers' Imprint*, 14(4).

Williams, J. R. G. (2014b). Nonclassical minds and indeterminate survival. *Philosophical Review*, 123(4):379–428.

Wilson, J. (2016). Are there indeterminate states of affairs? Yes. In Barnes, E., editor, *Current Controversies in Metaphysics*. Taylor and Francis. Forthcoming.