On Theory X and What Matters Most

S. J. Beard and Patrick Kaczmarek

Centre for the Study of Existential Risk, University of Cambridge

Abstract. One of Derek Parfit's greatest legacies was the search for Theory X, a theory of population ethics that avoided all the implausible conclusions and paradoxes that have dogged the field since its inception: the Absurd Conclusion, the Repugnant Conclusion, the Non-Identity Problem and the Mere Addition Paradox. In recent years, it has been argued that this search is doomed to failure and no satisfactory population axiology is possible. This chapter reviews Parfit's life's work in the field and argues that he provided all the necessary components necessary for a Theory X. It then shows how these components can be combined together and applied to the global challenges Parfit argued matter most: preventing human extinction, managing catastrophic risks and eradicating global poverty and suffering. Finally, it identifies a number of challenges facing his theory and suggests how these may be overcome.

> *I regret that, in a book called On What Matters, I have said so little about what matters. I have been trying to defend the belief that some things really do matter. I hope to say more about what matters in what would be my Volume Four.*
>
> *- Derek Parfit, On What Matters, Volume 3*

At the time of his death in 2017 it is clear that Derek Parfit was still working on the problems of population ethics that had concerned him for over 40 years. He described the book mentioned in the above quote (in a book that would come out three weeks after he died) as a rewrite of Part 4 of his groundbreaking 1984 book *Reasons and Persons*, and he also indicated that it would contain a "longer and revised version" of a talk he gave in Stockholm in 2014. It is possible that this would have fulfilled his plan, described in a letter to his sister Theadora Ooms in 1996 and read at his memorial symposium, to write a book "if I survive ... that will address truth, evil, time, and the sublime". Sadly, we will never get to see this book, and as far as the authors have been able to determine, no version of it survives in his lengthy correspondence and drafts, which are currently in the possession of Larry Temkin and Jeff McMahan. However, we believe that by reading across the breadth of Parfit's work, both published and unpublished, it is possible to determine a trajectory for some of his ideas that we can use to glean what this book might have defended. In order to stimulate further discussion and scholarly attention to the late work and manuscripts of Parfit, and to help those who are less interested in such minutia to actually understand and apply the theories contained in it, we here offer a summary of our conclusions about this trajectory and its likely destination.

The portion of Parfit's work that we will focus on deals with the field of population ethics, which concerns the evaluation of outcomes that differ in terms of the size or composition of populations. According to Parfit, a satisfactory theory of population, which he labelled 'Theory X', "must solve the Non-Identity Problem, avoid the Repugnant and Absurd Conclusions, and solve the Mere Addition Paradox".[1] Since he first suggested these criteria, many philosophers have attempted to prove the impossibility of meeting these requirements, and thus the non-existence of Theory X. Notable amongst these have been Gustaf Arrhenius, Stewart

[1] (Parfit 1984, 443)

Rachels and Tyler Cowen.[2] Others have argued that one or more of these criteria should not be regarded as so inviolable as Parfit suggests, so that theories that fail to meet all of them might still be satisfactory. Notable amongst these have been John Broome, Torbjörn Tännsjö and Michael Huemer regarding the Repugnant Conclusion, Larry Temkin regarding the Mere Addition Paradox and David Boonin regarding the Non-Identity Problem.[3,4]

Despite this, Parfit continued searching for Theory X and devoted much time in his final years to writing a lengthy paper, "Towards Theory X" (Parts 1 and 2), in which he developed many new lines of argument concerning it. While this remains unpublished, Parfit did publish two shorter works that explicitly considered how to satisfy the conditions for Theory X.[5] In doing so he sketched three principles: the Wide Dual Person Affecting Principle, the Simple View and the Imprecise Lexical View. These combine a unique approach to evaluating the wellbeing of variable populations with a broader axiological pluralism that is also concerned about people's Quality of Life.[6]

In this chapter, we consider these principles in light of the whole range of Parfit's writings and argue that they can, collectively, satisfy all the criteria that Parfit set himself, so that, although he never lived to say so, his search for Theory X may have been nearing an end. We also begin the next important step of developing this theory by applying these principles to the global challenges that Parfit maintained "mattered most".[7]

1. Background

1.1. The Absurd Conclusion

Let us begin with the criteria for Theory X that has received the least attention, and which he was able to satisfy first, and most directly. The Absurd Conclusion is implied by the view that the number of people who exist at any time is less important than the length of time over which people continue to exist because the value of adding extra people diminishes as instantaneous population size increases, but the value of adding future generations does not.[8] At one point Parfit appears to have found this view attractive. However, he ultimately rejected it. He argued that it would be wrong for us to believe that the value of adding a person with a bad life, such as one plagued by unbearable suffering and containing little or no good things, should ever be diminished. However, if the value of adding people with good lives at one point in time diminished, while the value of adding people with bad lives did not, then this would imply the following:

> *Absurd Conclusion.* In one possible outcome, the future would consist of a single century with an enormous number of people, "[n]early all of [whom] would have a quality of life far above that enjoyed by most of the Earth's actual population" while one in ten billion would "have a life that was not worth

[2] (Arrhenius 2000; 2003; 2011; Rachels 2004; Cowen 1996)

[3] (Broome 2004; Tännsjö 2002; Temkin 1987; Boonin 2014)

[4] The Absurd Conclusion has received comparatively little discussion in the literature.

[5] See, respectively, (Parfit 2016; 2017b)

[6] Although Parfit himself did not capitalize the phrase 'Quality of Life', throughout this chapter we do so (except when quoting or paraphrasing Parfit) because, we believe, it helps highlight that we are using this phrase in a very specific way.

[7] (Parfit 2017a, 436-7)

[8] Parfit attributes this view primarily to Peter Singer. See (Parfit 1976).

> living". In a second possible outcome "there would be the same enormous number of extra future people, with the same high quality of life for all except the unfortunate one in each ten billion" but these would exist across ten billion future centuries, rather than occupying only one. "[T]he first outcome would be very bad [and] the second very good even though, in both outcomes, there would be the very same number of extra future people, with the very same high quality of life for all except the unfortunate one in each ten billion."[9]

There is certainly something absurd about this conclusion, especially if one reflects on the fact that space and time, while being experienced in distinctively different ways by human beings, are very similar according to the current best laws of Physics. However, Parfit never explicitly defines the nature of this absurdity.

One possibility, which Parfit does not explore, is that this conclusion is absurd because the implied asymmetry between space and time is absurd. If this were so, then we could avoid it simply by accepting the principle that the value of good, but not bad, lives diminishes as population size increases, regardless of how this population is distributed across space or time. This is not a position that Parfit himself adopts, and it seems to us that this is likely because this move would have other implications that are hardly less absurd than that specifically mentioned by Parfit. Consider:

> *Another Absurd Conclusion.* For a population with a certain number and distribution of lives that is very good, the mere addition of groups with the same number and distribution of lives can make it worse, and could eventually make it very bad.

For instance, a population in which almost everybody enjoys a very high quality of life and only a tiny minority of people live bad lives might be very good when the size of the population was relatively small but could become very bad as it grew, even where the proportion of good and bad lives remain unchanged, if the value of these good lives diminished with population size while that of the bad lives did not. This conclusion would strike us as being equally as absurd as that described by Parfit in his *Reasons and Persons*, even though it relies on no asymmetry between space and time.

To put this another way, if we simply try to avoid the Absurd Conclusion by prohibiting different aggregation methods for the value of lives based on their location in space and time, we might still be left with theories that would imply that it may be good that there was life on Earth, even though there is also some suffering on Earth, but that it would be bad if there were also life on many other planets, even if none of these planets, though also containing some suffering, had a population with lives that were any worse than those on Earth.

However, Parfit's response to the Absurd Conclusion is equally able to handle both these sorts of absurdity. He sought to avoid the Absurd Conclusion by rejecting the view that "while there is no limit, in any period, to the disvalue of quantity, there is a limit to its value".[10] He would later formalize his position as:

[9] (Parfit 1984, 410-1)
[10] (Parfit 1984, 411-2)

> *The Simple View.* Anyone's existence is in itself good if this person's life is worth living. Such goodness has non-diminishing value, so if there were more such people, the combined goodness of their existence would have no upper limit.[11]

1.2. The Repugnant Conclusion

While avoiding the Absurd Conclusion, the Simple View makes it harder to avoid another problem; namely, the:

> *Repugnant Conclusion.* Compared with the existence of many people who would all have some very high quality of life, there is some much larger number of people whose existence would be better, even though these people would all have lives that were barely worth living.[12]

Indeed, Parfit noted that if one accepts the Simple View, then the only way to avoid this conclusion is to deny the principle that "a sufficiently large number of lives that all made the world better must, together, contribute more to the value of an outcome than a set number of other lives even if these were all, individually, much better".[13] Parfit was not the first to observe this conclusion;[14] however, he was the first to name it, draw attention to it and to explore its numerous ramifications.[15]

It is important to note that Parfit's description of the Repugnant Conclusion differed in some important ways from other philosophers who shared his concerns, both before and since.[16] The most notable of these is that, for Parfit, the Repugnant Conclusion was specifically, and always, a conclusion about people's Quality of Life, or the extent to which someone's life is *worth living*. 'Quality of Life' is a term that Parfit coined to describe the broadest possible conception of the value of a life, and one he intended to be broader than traditional conceptions of a person's welfare, wellbeing or utility and to capture more than simply that person's self-interest or what made their life go best.[17]

This leads to, and helps explain, another difference between Parfit's Repugnant Conclusion and that of many other philosophers. For Parfit:

> When we are most concerned about [the Repugnant Conclusion], our concern is only partly about the value that each life will have for the person whose life it is. We are also concerned about the disappearance from the world of the kinds of experience and activity which do most to make life worth living.[18]

Parfit, therefore, did not merely seek to incorporate one of the numerous, brilliant axiologies that have sought alternative means of aggregating value across lives,[19] but rather saw the

[11] (Parfit 2016, 112)
[12] (Parfit 2016, 110)
[13] (Parfit 2016, 112
[14] Perhaps the closest antecedent can be found in (McTaggart & McTaggart 1927). Similar observations can be found in (Sidgwick 1907), who notably did not find this conclusion implausible.
[15] (Parfit 1982, 142)
[16] (cf. Masny 2019)
[17] See (Parfit 1982, 117-8; 2016, 117).
[18] (Parfit 1986, 161)
[19] These are too numerous to mention. However, some notable examples include (Blackorby et al 2005; Hurka 1983; Thomas 2017).

Repugnant Conclusion as demanding a substantial account of what he called the 'qualitative differences' in the value of different lives. If it is to avoid the Repugnant Conclusion, therefore, Theory X must describe a broader axiological pluralism, evaluating outcomes according to both the wellbeing they contain and other facts that contribute to people's Quality of Life, and not simply propose an alternative method of aggregating value.[20]

Parfit's initial efforts to produce such a theory focused on a concern for the value of the *Best Things in Life*.[21] According to this view, which he labelled 'Perfectionism', "even if some change brings a great net benefit to those who are affected, it is a change for the worse if it involves the loss of one of the Best Things in Life".[22,23] The correct way to evaluate these goods, according to Perfectionism, is to view them pluralistically, considering both how they contribute to people's welfare as well as what Parfit sometimes referred to as their 'perfectionist value'. Thus, whether lives contain both kinds of value or whether they are valuable only because of the welfare they contain is more than a quantitative difference, it is a difference in quality. The potential loss of perfectionist value that only these Best Things in Life can bring justifies our belief that enough high-quality lives might be more valuable than any number of other lower-quality lives, allowing us to avoid the Repugnant Conclusion.

Yet, Parfit believed that Perfectionism in this form faced many objections and even called it "crazy".[24] Two of these objections are that it had elitist implications because it gave insufficient consideration to eliminating suffering (a point we will return to in §3) and that it could not clearly differentiate between the different kinds of good, or between their values.[25] Thus, while his earlier writings provide us with some of the ingredients necessary for his Theory X, these still require substantial elaboration.

2. Parfit's Final Papers

2.1. The Non-Identity Problem

A commonly-held intuition is that one of two outcomes cannot be worse if this outcome would be worse for no one.[26] However, Parfit famously demonstrated that this leads to a head-on collision with another well-worn intuition in variable population cases.

> *Non-Identity Problem.* That where different people existed in different outcomes, one of these outcomes could be clearly worse, for instance because

[20] We do not mean to imply that Parfit was the only person who shared this view, although he seems to have been in the minority. For other sources, see (Crisp 1992; Cowen 1996).

[21] The Best Things in Life "are the best kinds of creative activity and aesthetic experience, the best relationships between different people, and the other things which do most to make life worth living" (Parfit 2016, 161).

[22] (Parfit 2016, 163)

[23] Parfit may have held a similar view about rationality, arguing that a creator's desire for "his creation to be as good as possible" or a scientist's desire "to make some fundamental discovery" would be rational "even if this person knows that his act is against his own self-interest" (Parfit 1984, 192).

[24] (Parfit 1986, 164)

[25] (Parfit 2016, 163-164)

[26] Parfit dubbed this the 'Narrow Telic Principle' (Parfit 2017b, 118).

> the people in that outcome lived worse lives, even though it was worse for nobody, because these people did not exist in the other outcome.[27]

For much of his life, Parfit thought that the only way to avoid this problem was to endorse the impersonal view that it should make *no difference* whether an outcome was better or worse for any particular people.[28]

However, in a paper published posthumously, Parfit suggested that he considered this conclusion to have been mistaken. He maintained that, while the Non-Identity problem constitutes a strong objection to narrow person affecting views, which appeal to comparative harms or benefits, there were other *wide* person affecting views that could avoid this difficulty by appealing to whether an outcome was merely good, or bad, for people instead. According to these principles, I am benefited if you do what is good for me, even if I do not 'fare better' relative to some other outcome, where the degree to which an outcome is good for me depends only on how well I fare in that outcome. Choosing an outcome in which somebody does not exist can therefore be seen as denying that person a benefit that was ours to provide. Even if we can't quite describe this as a harm, because such an outcome is in no way bad for this person, the total lack of value for them is still an evaluative feature of this outcome, indicating that choosing it, rather than an outcome in which they exist with a good life, means choosing not to do what would have been good for this person.

In this way, we can avoid the Non-Identity problem while still accepting principles like the:

> *Wide Principle.* One of two outcomes would be in one way worse if this outcome would be less good for people, by benefiting people less than the other outcome would have benefited people.[29]

Such principles permit more sophisticated approaches to evaluating the lives of future people, which can avoid the worst excesses both of impersonal views (e.g., that we must give these people equal weight in our moral consideration) and of other person-affecting views (e.g., that, in non-identity cases, these people's wellbeing may not matter at all). This is because they do not necessarily imply that *existential benefits*, benefits of coming into existence with a life that is good for one, are always of equal worth to *comparative benefits*, benefits of being better off given that one already exists.[30]

One difference between existential benefits and comparative benefits is that, other things being equal, existential benefits change the number of persons who exist. Parfit saw this as an important difference, and argued that we should accept wide person-affecting principles that account for it. In particular, we ought to accept the:

[27] (Parfit 1984, 359)

[28] (Parfit 1984, 363-380)

[29] (Parfit 2017b, 129)

[30] Parfit credits Jeff McMahan with these terms. In earlier work, he referred to these two kinds of benefits as 'comparative' and 'non-comparative' and argued that acknowledging the existence of non-comparative benefits helps to bridge the gap between consequentialist and contractualist accounts of why we should care about future generations (Parfit 2011b). See also (Beard & Kaczmarek 2020).

> *Wide Dual Person-Affecting Principle.* One of two outcomes would be in one way better if this outcome would together benefit people more, and in another way better if this outcome would benefit each person more.[31]

According to this dual principle, the addition of lives that are worth living would always be in one way good, and this value is non-diminishing as the Simple View implies. However, comparatively benefiting people would also be good in another way. Thus, for any fixed quantity of wellbeing, this principle prefers an outcome where this is concentrated in fewer people to one where it is spread across these people and also other people who do not exist in the first outcome. As Parfit puts it "[t]he first outcome would be better because the benefits to each would be greater and the benefits to all would be the same".[32]

However, it is less clear what this principle should say about two outcomes that involve a tradeoff between the quantity of wellbeing and the number of people. For instance, suppose we must choose between:

> A = n people at level 100; and
>
> B = 2n people at level 75

In this case, choosing A would still benefit each person more, because all the people in A are better off than they would be in B, but choosing B can be said to 'together benefit people more' because it involves a larger quantity of wellbeing. Parfit offers no arguments for how this tradeoff should be made, and indeed in his draft paper, "Towards Theory X: Part Two", he provides some lines of argument suggesting that no precise tradeoff between these two ways in which outcomes could be better may be possible due to 'different-number-based imprecision'.[33]

[31] (Parfit 2017b, 154)

[32] (Parfit 2017b, 154)

[33] These arguments can be found in (Parfit ms2, 8-14). Different-number-based imprecision involves the claim that "[when] two possible worlds would contain different numbers of people, this fact makes these worlds less precisely comparable". He develops arguments for the view based on two sets of competing principles. Firstly, the 'Total-and-Average View', that:

> It would always be in one way better if (1) the people who exist would together have a greater total sum of well-being, and in another way better if (2) because this total sum of well-being would come to fewer people, there would be a greater average sum of well-being per person.

and secondly, the 'Total View' and the 'Part-Whole Principle', that

> If world T would contain more people than World S, and it is true both that everyone who would exist in S would either also exist in T and be better off, or would have some counterpart in T who would be better off, and that everyone else who would exist in T would have a life that is clearly worth living, these facts strongly support the view that T would be better than S.

However, it's not clear that Parfit actually believed any of these principles and views. Both the lines of argument he developed and the implication of different-number-based imprecision, came in for significant criticism; see e.g., (Arrhenius 2016). Parfit acknowledged that his claims "are hard to prove, since it makes sense to claim that all such truths are precise" and these lines of argument were dropped from later work, such as (Parfit 2016). However, there is at least anecdotal evidence that he continued believing in different-number-based imprecision and hoped to revise his arguments and respond to his critics in later work.

On the other hand, Parfit was also keen to stress that there are some cases in which such tradeoffs between the value of these two kinds of benefit should be made. For instance, if the difference between the wellbeing levels of people across two outcomes were very great, the existence of many more people at the lower level would be worse than the existence of fewer people at the higher level even if it meant more wellbeing, and hence a greater collective benefit.[34]

Parfit argued that his Wide Dual Person-Affecting Principle would agree since it "would often imply that one of two outcomes would be better though the people who exist would have a much smaller total sum of benefits".[35] However, he conceded that the principle might still form part of a view that implied the Repugnant Conclusion. Yet, he did not believe that this would be the case for his Theory X because:

> We might also justifiably believe that great losses in the quality of people's lives could not be outweighed by any increase in the sum of benefits, if these benefits came in the lives of people whose quality of life would be much lower.[36]

This brings us to the final component of Theory X.[37]

2.2. The Mere Addition Paradox

In the above quote, Parfit was almost certainly alluding to his 2016 paper "Can We Avoid the Repugnant Conclusion?", in which he sets out to defend the Simple View with the following:

> *Imprecise Lexical View.* If many people exist who all have some high quality of life, that would be better than if there existed any number of people whose lives, though worth living, would be, in certain ways, much less good.[38]

The proposed view contains two related principles, only one of which is explicitly stated here. The first of these, which is left implicit in the above statement, is the principle that many evaluative judgements cannot be made precisely, even if one is aware of all the underlying facts. Parfit's acceptance of this principle appears closely related to his value pluralism since one of the key sources he gives for this normative imprecision is that outcomes can differ in terms of the many features that determine people's Quality of Life. As previously alluded to, Parfit saw

[34] (Parfit 2017b, 155-6)

[35] (Parfit 2017b, 157)

[36] (Parfit 2017b, 157)

[37] We have focused on only one kind of difference between the value of existential and comparative benefits that Parfit's Wide Principle might allow. Yet, he acknowledged that there may be others. In particular, this principle might allow us to say that we have strongest reason to provide a comparative benefit though providing an existential benefit would produce a better outcome. For instance, he writes that it is both plausible and defensible to believe that "we ought to give some benefits to presently existing people rather than giving some greater benefits to future people" (Parfit 2017b, 148-9). This difference is also discussed at greater length in (Parfit ms3, 18-19). However, note that despite these differences between the moral significance of existential and comparative benefits, Parfit continued to defend a version of his 'No Difference' view, that: when our choice and acts would greatly lower the quality of life of future people "[t]hese choices and acts would make things go much worse, and would be wrong. It would make no moral difference… that these choices and acts would be worse for no one" (paraphrased from Parfit 2017a, 122-3).

[38] (Parfit 2016, 115)

these features as being distinct from, and broader than, those that determined people's wellbeing. Because of this distinctiveness, we cannot evaluate all these features on a common scale; this leads to the conclusion that we may not be able to produce a precise ordering of outcomes that contain these lives. As Parfit puts it:

> There can be fairly precise truths about the relative value of some things. One of two painful ordeals, for example, might be twice as bad as the other, by involving pain of the same intensity for twice as long. However, in most important cases relative value does not depend only on any such single, measurable property. When two painful ordeals differ greatly in both their length and their intensity, there are no precise truths about whether, and by how much, one of these pains would be worse. There is no scale on which we could weigh the relative importance of intensity and length. Nor could five minutes of ecstasy be precisely 7.6 times better than ten hours of amusement.[39]

In many cases, such as these, "when two things are qualitatively very different", outcomes containing them may be *roughly comparable*; meaning that it is "impossible either that one of these things is better than the other by some precise amount, or that both things are precisely equally good".[40] However, sometimes the qualitative differences between two outcomes, or the lives they contain, make them *utterly incommensurable* because there is simply no way in which the value of these outcomes can be compared on a single common scale, even roughly.

This, we contend, is where the second explicit feature of Parfit's view comes in, which is the analogue of his earlier Perfectionism. Where the qualitative difference between two goods, P and Q, reaches a point where their values are incommensurable, and yet P is still better than Q, the evaluative relation of lexical superiority and inferiority.[41] As Parfit defines this, "though the existence of more Qs would always be non-diminishingly better, the existence of some sufficient

[39] (Parfit 2016, 113)

[40] (Parfit 2016, 113)

[41] Parfit says little about his conception of lexicality. His fullest account appears to be the following passage from "Towards Theory X: Part One":

> The most important sense of 'good' and 'bad' are, I believe, reason-implying. One of two things is relevantly better if this thing has features that does or might give us stronger reasons to choose this thing, or to respond to this thing in other positive ways. If some things are lexically better than others in this reason implying sense, the strength of these reasons could not be represented by using either a scale or numbers. Though there would be no limit to the combined strength of our possible reasons to choose certain things, we would have stronger reason to choose certain other things.
>
> Such claims may seem to make no sense. If there is no limit to the combined strength of certain reasons, it may seem impossible that certain other reasons could be stronger. But this objection assumes that the strength of all our reasons must correspond to different positions on some scale, or be able to represented by numbers. As before, that is not so. It may help to mention here what Raz calls exclusionary reasons. These are reasons to ignore, or give no weight to, certain other reasons. Suppose, for example, that I am judging who deserves to win some prize, and one of the contestants is my best friend. Though I have a reason to want my best friend to win, my role as a judge gives me an exclusionary reason to ignore this friendship-based reason. This exclusionary reason doesn't defeat my friendship-based reason merely by being stronger. Excluding is not the same thing as outweighing. Even if we doubt the claim that there are such exclusionary reasons, we should admit that this claim makes sense (Parfit ms1,15-6).

number of Ps would be better than the existence of any number of Qs".[42] In the case of wellbeing and the Best Things in Life, this means that:

> [If in one world] there would be no art, or science, no deep loves or friendships, no other achievements, such as that of bringing up our children well, and no morally good people. [That world] would be much worse than [some other worlds] in what we can call qualitative or perfectionist terms .... This great qualitative loss would, I believe, make [this world] in itself a worse world, even [if it] would give, to the same number of people, a greater and more equally distributed sum of well-being.[43]

As we saw in §1.2, the Imprecise Lexical View does not require both imprecision and lexicality simply to avoid the Repugnant Conclusion, perfectionist lexicality alone would suffice. However, the combination of both these features can overcome at least one of the objections Parfit raised against this earlier Perfectionism, that the difference between the Best Things in Life and other goods could not be precisely defined. This second view also performed better than his earlier Perfectionism in relation to certain versions of the Mere Addition Paradox.

One of these versions of the Mere Addition Paradox, which both appeared especially troubling to Parfit, and was the last version he considered, concerns the following case:

> World A = N people at level 100;
>
> World Raised A Plus = N people at level 101 and a million times as many other people at level 95;
>
> World B = the same number of people as in Raised A Plus but all at level 99.

These worlds, together with the Z population from the Repugnant Conclusion, are illustrated in Figure 1.

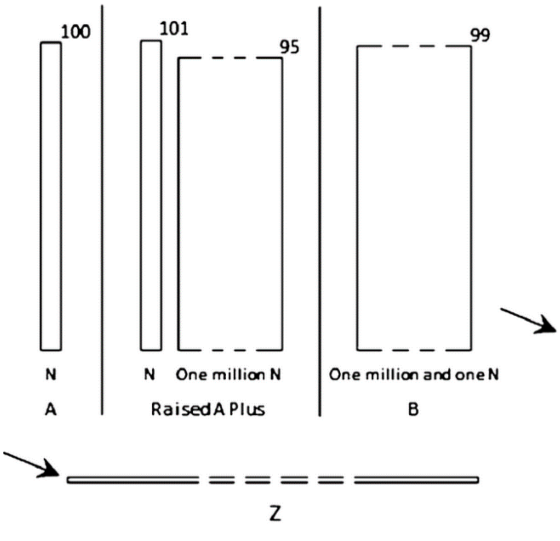

Figure 1. Source: (Parfit 2016, 117)

[42] (Parfit 2016, 112)

[43] (Parfit 2016, 123)

Parfit claims that somebody might argue:

> (H) Since N of the people in Raised A Plus would have a higher quality of life than all of the N people in A, and everyone else in Raised A Plus would have a quality of life that would not be much lower, Raised A Plus would be better than A.
>
> (I) Since World B would contain as many people as Raised A Plus, and these people would together have, in B, a greater and more equally distributed sum of well-being, B would be better than Raised A Plus.

Therefore:

> (J) B would be better than A. Compared with the existence of N people at level 100, it would be better if there existed instead a million times as many people all at level 99.[44]

However, if this was so, then it would seem that we could apply the same kind of argument to another three populations, taking B as our starting point and arguing that a Raised B Plus would be better than B and that C would also be better than Raised B Plus, and hence than A, despite everyone in C having a lower Quality of Life than those in B. Continuing down this chain of reasoning would produce the inevitable conclusion that World Z is better than World A. This is paradoxical because it implies the Repugnant Conclusion.[45]

According to the perfectionist part of this view, we can reject the conclusion of this chain of reasoning because the transition from A to Z involves the loss of the Best Things in Life. Yet, on its own, this argument gives us no reason to refute any of the individual steps in this chain, and at each of these steps, the judgement that will inevitably lead to the Repugnant Conclusion seems more plausible than its alternative, making the denial of any of these steps appear crazy. However, Parfit's new Imprecise Lexical View can counter this objection because it creates a new, more reasonable, way in which we can reject (I), namely we can claim that B and Raised A Plus are not precisely comparable. As Parfit puts it, "though World B would be better than Raised A Plus in utilitarian and egalitarian terms, B would be worse in qualitative terms, since the best things in people's lives would be worse in B".[46] This leads him to conclude that:

> (K) given the conflict between these values, Worlds B and Raised A Plus are only imprecisely comparable, and would be imprecisely equally good.

Parfit notes that, on its own, (K) may still appear less plausible than (I) as the qualitative difference between the people in Raised A Plus and B would be very slight. However, though B is in many ways better than Raised A Plus, there is at least this one way in which Raised A Plus is better, and since we cannot precisely weigh the relative strength of these different kinds of value, we should not accept (I), but can only say that B would be imprecisely better than Raised A Plus. Furthermore, Parfit claims, if we overlook the potential importance of this qualitative difference between Raised A Plus and B, by assuming that such a small difference must be trivial in comparison to the many ways in which B is better than A, then this will lead to the Repugnant Conclusion. As he puts it:

[44] (Parfit 2016, 124)

[45] But see also (Temkin 1987).

[46] (Parfit 2016, 126)

> (K) seems implausible because, in a change from Raised A Plus to B, there would be only a slight qualitative loss. The best lives would fall only from level 101 to 99. It may be hard to believe that this slight qualitative loss could make Raised A Plus not better than B, but only imprecisely equally good. But it would be much harder to believe that, compared with the existence of many people whose quality of life would be very high, it would be better if there existed instead some vast number of people whose lives were barely worth living. Since (K)'s way of rejecting premise (I) is less implausible than the Repugnant Conclusion, this argument fails. By appealing to this Imprecise Lexical View, we can justifiably reject this conclusion.[47]

This gives us reason to reject (I), and hence to avoid even this especially troubling version of the Mere Addition Paradox while still avoiding the Repugnant Conclusion.

Hence, we see how combining the Imprecise Lexical and Simple Views with the Wide Dual Person-Affecting Principle can produce a theory capable of meeting the four criteria for Theory X.

3. Applying Theory X

The previous sections outlined the basic principles of Theory X and how these met the theoretical conditions that Parfit wanted it to satisfy. In this section, we continue to explore this moral theory by applying it to the global challenges that Parfit argued mattered most.

3.1. Ensuring Humanity's Survival

Parfit addresses the question of 'what matters most?' in *Reasons and Persons*, *On What Matters, Volume I* and *On What Matters, Volume III*. Here is what he says in *On What Matters, Volume III*:

> What now matters most is how we respond to various risks to the survival of humanity. We are creating some of these risks, and we are discovering how we could respond to these and other risks. If we reduce these risks, and humanity survives the next few centuries, our descendants or successors could end these risks by spreading through this galaxy.[48]

This echoes a concern that goes back *Reasons and Persons*, and a thought experiment about the costs of nuclear war. He writes:

> I believe that if we destroy mankind, as we now can, this outcome will be much worse than most people think. Compare three outcomes:
>
> (1) Peace.
> (2) A nuclear war that kills 99% of the world's existing population.
> (3) A nuclear war that kills 100%.

[47] (Parfit 2016, 126-7)
[48] (Parfit 2017a, 436)

> (2) would be worse than (1), and (3) would be worse than (2). Which is the greater of these two differences? Most people believe that the greater difference is between (1) and (2). I believe that the difference between (2) and (3) is very much greater. … The Earth will remain habitable for at least another billion years. Civilization began only a few thousand years ago. If we do not destroy mankind, these few thousand years may be only a tiny fraction of the whole of civilized human history. The difference between (2) and (3) may thus be the difference between this tiny fraction and all the rest of this history.[49]

Theory X supports this conclusion. This is because, according to its principles, all potential future lives should be counted equally (the Simple View) and their wellbeing will contribute to the goodness of outcomes as much as our own, in one way at least (the Wide Dual Person-Affecting Principle).

However, as we have seen, these are not the only principles that Theory X contains and, while its other principles do not undermine this conclusion, they do imply that this is not the only feature of human extinction that would be bad. This is best illustrated by another thought experiment from the end of Part 4 of his *Reasons and Persons*, concerning another version of the Mere Addition Paradox. Parfit asks us to consider the following possible futures for our species:[50]

> When the A+ Future begins, everyone enjoys an extremely high quality of life. After a thousand years, the surface of the earth becomes inhospitable, lowering everyone's quality of life. After another thousand years, the surface of the earth becomes uninhabitable, thereby ending human history.
>
> In the New A Future, the first two thousand years go even better. Everyone has a quality of life that is higher than that of the best-off people in the A+ Future. Near the end of this period, scientists predict that the earth's surface will shortly become uninhabitable, so people decide to dig deep caves, enabling humanity to survive. However, while life in these caves is worth living, it is far less good than it had been on the Earth's surface. Throughout the years lived in the caves, people's quality of life is very low, and their lives are only barely worth living. Yet, since the New A Future would be in no way worse than the A+ Future, and in at least one way better, it seems that the New A Future would be better.
>
> In the New B Future, people in the first two thousand years have decided, at a slight cost to their quality of life, to prepare the caves by stocking them with long-lasting resources that provide a great increase in the quality of life of everyone who lives during the first two thousand years in these caves, before these resources are exhausted, and their lives become barely worth living. People's quality of life would be the same for the first four thousand years. Although those living in the first two thousand years, would have a slightly lower quality of life, they would lose very much less than would be gained by as many people in the next two thousand years. This would seem to be an even better outcome.
>
> Similar remarks apply to the C-Future, in which people decide, at a slightly greater cost to themselves, to prepare the caves by stocking them with even

[49] (Parfit 1984, 453-4)

[50] (Closely paraphrased from Parfit 1984, 438ff).

> longer lasting resources that provide a slightly lower benefit to the people who will live in these caves, but will last for an additional four thousand years before they become exhausted. The people in the first four thousand years would be slightly worse off in terms of their quality of life, but there would be a very much greater gain for the people in the next four thousand years. Hence, this outcome seems even better
>
> By the same reasoning, so would the D-Future, and the E-Future, and so on. Therefore, the best of all these possible histories would seem to be the Z-Future. In this future, the people living in the first two thousand years decide, at very great cost to themselves, to prepare the caves by stocking them with extremely long-lasting resources that might only provide a tiny benefit to those who would live in these caves, but which would not be exhausted for millions of years, providing by far the greatest total benefit overall. In this outcome, throughout the rest of human history, the quality of life would not be much above what it would have been in the caves in the New A Future, and everyone would have lives that are barely worth living. This is the Repugnant Conclusion.

These futures are illustrated in Figure 2.

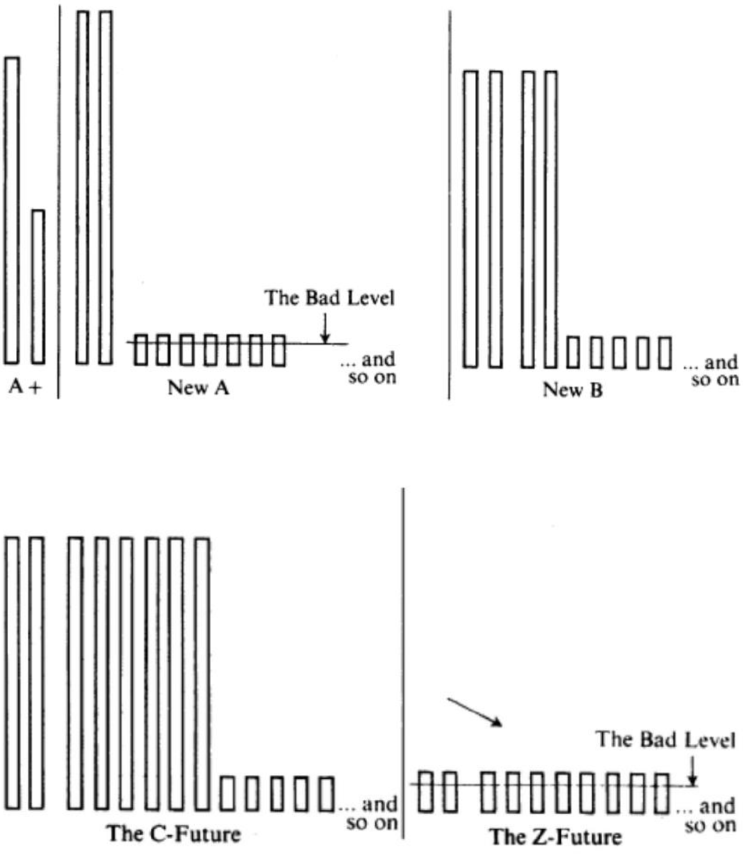

Figure 2. (Parfit 1984, 438-9)

As should be clear from the previous section, Theory X, because it contains the Imprecise Lexical View, would only be committed to the first step in this chain of reasoning, and would imply that even the transition from the New A Future to the New B Future would not be a precise change for the better but rather, at most, a change between two outcomes that are only roughly comparable. Though the New B Future would be better than New A regarding its total sum of benefits, New B would be worse in qualitative terms, given that people's Quality of Life would be lower in New B than in New A. Hence, New B would be no better than New A.

Thus, while Theory X gives us reason to seek to save humanity from extinction on the grounds of astronomical waste, it does not imply that we should place the maximization of future wellbeing ahead of all other concerns. For instance, if we find ourselves at the pinnacle of human history, facing a future in which things will only get worse (though they may never get bad), ensuring our species' future may not be worthwhile if it meant making certain sacrifices that would worsen this pinnacle, for instance by removing some of the Best Things in Life.

Of course, this is not the kind of future that Parfit imagines; indeed, his view is that in the future people will enjoy more of the Best Things in Life. As he puts it:

> Life can be wonderful as well as terrible, and we shall increasingly have the power to make life good. Since human history may be only just beginning, we can expect that future humans, or supra-humans, may achieve some great goods that we cannot now even imagine.[51]

Suppose, then, that some of the details of Parfit's nuclear war example change. Life is still good until nuclear war erupts, killing 99% of humanity and destroying our civilization. For several hundred years, the survivors manage to keep warm in tight-knit tribes during the longest winter on human record only to inherit what Tim Mulgan calls a 'broken world'.[52] Though humanity may continue to survive for millions of years, future people's lives will forevermore be only barely worth living. Thus, while human extinction would still involve the loss of a very large quantity of wellbeing, since no amount of wellbeing could ever be as valuable as the loss of the Best Things in Life it is now the shift from peace to war that constitutes the greater total loss of value and the difference between these two would be greater than that between human survival and extinction. While never explicitly stated by Parfit, the idea that there may be something even more important than the mere survival of humanity is reflected in his work. For instance, when discussing the badness of nuclear war, Parfit says:

> [Some] people believe that there is little value in the mere sum of happiness. For these people, what matters are what Sidgwick called the 'ideal goods'—the Sciences, the Arts, and moral progress, or the continued advance towards a wholly just world-wide community. The destruction of mankind would prevent further achievements of these three kinds. This would be extremely bad because what matters most would be the highest achievements of these kinds, and these highest achievements would come in future centuries.[53,54]

At this time Parfit made no pronouncement of whether he counted himself amongst these people or amongst those who cared most about the future quantity of wellbeing. However, as we have seen, his subsequent writings suggest that Theory X gives greater importance to something like these ideal goods.

[51] (Parfit 2017a, 436)

[52] (Mulgan 2015, 93)

[53] (Parfit 1984, 453-4)

[54] Sidgwick discusses the subject at (Sidgwick 1907, 114-5).

Hence, according to Theory X, while the avoidance of human extinction may be what matters most, it is avoiding the loss of the Best Things in Life that contributes most to this fact. Seen in this light, we may be said to currently care far less than we should for the preservation of what is best about our civilization, as well as the preservation of our species. Let us call such a potential loss of perfectionist value a 'P-catastrophe', and the risk of such catastrophes 'P-risks'. According to Theory X, P-risks present a morally significant and neglected area of concern.

3.2. Managing Global Catastrophic Risk

However, Theory X does not only have interesting implications about the value of human survival but also how to respond to risk and uncertainty. For instance, when we do not know if the survivors of a nuclear war would recover and rebuild a flourishing civilization. For ordinary people in the real world, who have no crystal ball and cannot know what the future holds, important decisions must invariably be made under such conditions.

Despite stressing the primary practical importance of risk and uncertainty, Parfit did little to address this pressing issue directly. For instance, one of his best-known thought experiments from *Reasons and Persons* asked readers to consider:

> *The Risky Policy*. As a community, we must choose between two energy policies. Both would be completely safe for at least three cen- turies, but one would have certain risks in the further future. This policy involves the burial of nuclear waste in areas where, in the next few centuries, there is no risk of an earthquake. But since this waste will remain radio-active for thousands of years, there will be risks in the distant future.[55]

Yet, at no point does he describe how we should respond to this risk; preferring, in cases like this, to assume that a negative outcome either obtains or does not.[56]

Parfit comes closest to setting out his reasons for this omission in the following passage from *On What Matters, Volume 1*:

> It is of great practical importance what we ought to do in cases that involve risk or uncertainty. These questions have been well discussed by many philosophers, decision theorists, and others. Certain other questions about reasons, though more fundamental, have been less well discussed. These are also questions about which people disagree more deeply.[57]

And later:

> Given the difference between these two sets of questions, they are best discussed separately. So I shall often suppose that, in my imagined cases, everyone would know all of the relevant facts.[58]

[55] (Parfit 1984, 371)

[56] At least one associate of Parfit's maintains that this was simply because he was too bad at mathematics to work directly on cases involving probability.

[57] (Parfit 2011a, 37)

[58] (Parfit 2011a, 162)

While acknowledging that Parfit could not possibly have dealt with every aspect of moral theory in his work, we believe that this omission was unfortunate and that his Theory X has more to offer to discussions about risk and uncertainty than he apparently realized. This is because it helps to address the difficult issue of how theories involving lexical superiority, or other forms of axiological absolutism, handle situations under risk.

One's knee jerk reaction may be that such theories are unduly insensitive to uncertainty and risk. That, on the one hand, it appears that any chance, no matter how small, of preserving a lexically superior good would be morally more important than even the certainty of losing a very large quantity of lesser goods. On the other hand, however, it would also seem that any chance, no matter how small, that this superior good might be lost would also be morally more important than even the certainty of gaining a very large quantity of lesser goods. A handful of philosophers argue that implications like this constitute decisive objections to such theories as being too demanding in the priority they give to higher goods and their insensitivity to risk, while efforts to make them less demanding, for instance by setting a threshold probability below which lexical superiority no longer applies, lead to paradoxical implications.[59]

However, other philosophers have found ways to counter these arguments. Seth Lazar and Chad Lee-Stronach, for instance, have recently shown how lexical, and other absolutist, axiologies can provide a coherent and undemanding response to cases involving risk and uncertainty, so long as they avoid assigning infinite value to superior goods, but rather posit that the value of lesser goods cannot exceed an upper bound. Orthodox Expected Utility Theory would then allow tradeoffs between superior and lesser goods under some conditions, but not others, without this having any paradoxical implications.[60]

They suggest that the most obvious way to achieve this is if "[any lesser good] has diminishing marginal value, which decreases asymptotically towards zero, so that the total value approaches a limit beneath the value of a single token of [a superior good]".[61] Reducing the probability of creating (or preserving) a superior good will then make it possible for a sufficient quantity of the lesser good to exceed its value, despite this diminishing tendency. So, for instance, we might believe that the value of adding to the total sum of benefits in an outcome decreases as the sum of benefits increases, so that there is a maximum value for the total quantity of welfare. We might represent this as a utility score of 10. Now, if a single instance of one of the Best Things in Life had a utility score of 20, this would explain why the loss of one of these things would be bad, no matter how much additional welfare was added. However, merely accepting a 25% risk that one of the Best Things in Life could be lost would only have an expected utility score of −5, meaning that there could still be some amount of welfare that would make accepting this risk worthwhile.

Since it includes the Simple View, that the value of additional good lives does not diminish, Theory X specifically excludes this approach. However, Lazar and Chad Lee-Stronach acknowledge this is not the only approach we can take to avoid these problems, and indeed there remains a significant problem with their proposed solution. This is that it can provide different assessments of groups of choices, depending on whether they are considered individually or together. For instance, if the above gamble were considered morally acceptable once, then it would seem no less acceptable to make it multiple times. However, if considered as a group accepting three separate 25% risks to one of the Best Things in Life would have an expected utility score of more than −13, meaning that there should be no quantity of wellbeing worth making this sacrifice for. The difficulty for lexical, and other absolutist, axiologies is how

[59] See for example (Huemer 2010).

[60] (Lazar & Lee-Stronach 2017)

[61] (Lazar & Lee-Stronach 2017, 100)

to decide whether to evaluate each of these gambles on their own, rendering them permissible, or collectively, rendering them impermissible.[62]

Parfit referred to this kind of problem as an 'each-we' dilemma and designed his theories specifically to address them.[63]

So, what approach might Theory X take to solving these challenges? We can start by considering some passages in which Parfit sets out a unique approach to the description of risk. On his view:

> To decide which of our possible acts would make things go expectably-best, we take into account both how good the effects of the different possible acts might be, and the probabilities, given our beliefs or the available evidence, that these acts would have these effects. When what matters is only the number of lives that are saved, some act's outcome would be expectably-best if this is the act that would save the greatest expectable number of lives. The expectable number that some act would save is the number of lives that this act might save, multiplied by the chance that this act would save these lives.[64]

He adds two clarificatory notes to this passage, pointing out that:

> Rather than talking of the expectable goodness of these outcomes, many people talk of their expected goodness. But that word is misleading, since such expectable goodness is often not goodness that either is, or should be, expected.[65]

And that:

> Expectabilists need not assume that the expectable goodness of outcomes depends only on the expectable sum of benefits. As Broome and Kamm suggest, for example, it may also matter how these benefits, or people's chances of getting these benefits, are distributed between different people .... And we might have reasons to be risk-averse, giving somewhat greater weight to avoiding the worst outcomes.[66]

Given that the language of expected value is standard across the literatures of ethics, economics and decision theory, this choice of terminology may appear pedantic. However, when seen in the light of his Imprecise Lexical View, we believe its importance becomes clearer.

According to Orthodox Expected Utility Theory, one should evaluate cases of risk merely by multiplying the value of each possible outcome by the probability of that outcome given one's choice. Such an approach can make good sense when all values are representable on a single scale. However, it makes less sense when they are not. For instance, Expected Utility Theory would say that we should treat a situation in which there is a 50% chance of losing all the Best Things in Life in the same way as we would a situation in which we would expect to lose 50% of the Best Things in Life. However, according to the Imprecise Lexical View, a situation in which 50% of the Best Things in Life would be retained would still be lexically superior to a situation

[62] (Lazar & Lee-Stronach 2017, 105-7)
[63] (Parfit 1984, 91)
[64] (Parfit 2011a, 160)
[65] (Parfit 2011a, 462)
[66] (Parfit 2011a, 462-3)

in which all were lost, while having a value comparable with preserving all the Best Things in Life, because the existence of these goods matters far more than their quantity.

In considering how we should respond to a situation in which the expectable outcome involves a risk to the Best Things in Life, we need to find a better way of representing this intermediate state in our axiology. The Imprecise Lexical View provides just such a way, because outcomes that are truly intermediate between lexically superior and inferior alternatives are not precisely comparable with either, but roughly comparable with both. This allows us to construct a framework that will be functionally like Lazar and Lee-Stronach's, but where it is not the value of a lesser good (here, wellbeing) that has an upper bound, but rather its comparability to the value of superior goods (the Best Things in Life). As we reduce the probability that superior good will exist, we effectively make the value of an outcome more comparable to one in which a lower good exists with certainty and less comparable to one in which a higher good exists with certainty. Depending on the degree of qualitative difference between these two goods, there will be a range of thresholds, representing the points at which this probabilistic good becomes increasingly incomparable with the certainty of a superior good and comparable with the certainty of a lesser good.[67]

On this framework, we might say that it would not be worse to sacrifice a small chance of preserving some of the Best Things in Life in order to increase the overall level of wellbeing, but it would still be, at most, a matter of rough comparability. Since, as we described in §2.2, rough comparability is not a transitive notion, there is no contradiction between believing that one person did no wrong by choosing between two roughly comparable options, but that many people together did great wrong by selecting a lexically inferior option, thus avoiding the each-

[67] Parfit sometimes used a 5-tier hierarchy of thresholds to describe such increasing incomparability (Parfit ms1, 23 and 26). This could be adapted for cases involving risk and uncertainty along the following lines:

> 1. Where both the qualitative differences between the two goods are very small and the probability of receiving the superior good is very low, then there's some quantity of the lesser good that would be better.
>
> 2. As the qualitative differences between the two goods either increases and/or the proba- bility of receiving the superior good increases, then:
>
> > (a) first it would become indeterminate whether there is some quantity of the lesser good that would be better or whether any quantity of the lesser good would at most be roughly comparable;
> >
> > (b) then any quantity of the lesser good would at most be roughly comparable;
> >
> > (c) and then it would become indeterminate whether any quantity of the lesser good would at most be roughly comparable or would be worse.
>
> 3. Finally, where either the qualitative differences between the two goods are very great or the probability of receiving the superior good is very high, any quantity of the lesser good would be worse.

we dilemma implied by Lazar and Lee-Stronach's approach. This seems to us like a promising proposal, and we believe that further work on this aspect of Theory X is needed.[68]

3.3. Alleviating Poverty & Suffering

While preventing human extinction is what Parfit thought mattered most, he also believed that it mattered a great deal that we do more to counteract the worst cases of human suffering and poverty in the world. As he puts it:

> One thing that greatly matters is the failure of we rich people to prevent, as we so easily could, much of the suffering and many of the early deaths of the poorest people in the world. The money that we spend on an evening's entertainment might instead save some poor person from death, blindness, or chronic and severe pain. If we believe that, in our treatment of these poorest people, we are not acting wrongly, we are like those who believed that they were justified in having slaves.[69,70]

Claims such as this are not hard to defend on purely utilitarian grounds and seem consistent with the principles in Parfit's Theory X. However, it is hard to miss the notion that there is something especially bad about suffering, which is not captured merely in its negative effect on individuals' wellbeing. This sense is brought into focus when one considers Parfit's specific reference to the wrongness of keeping slaves, something that he had previously discussed in his unpublished "Towards Theory X: Part 2". Here he writes:

> The ancient Greeks had amazing achievements in poetry, drama, architecture, sculpture, history, philosophy, mathematics, and some kinds of science. These achievements were in part made possible by slavery and unjust inequalities. When we consider these facts, we may … believe that these ancient Greeks acted wrongly, and ought to have abolished slavery and these inequalities, even if that would have prevented most of these great achievements. We may also believe that, though these ancient Greeks ought to have acted in these ways, these morally required acts would have made history go in ways that would not have been better, but only imprecisely equally as good. On this view, when we compare these ways in which history actually went and might have gone, there would be no precise truths about the relation between the perfectionist value of these great achievements, and the badness of Greek

[68] Some may object to the introduction of yet more imprecision into our evaluation of outcomes under risk and believe that, from being too demanding, this makes Theory X not demanding enough because we might plausibly do what we wish if we cannot precisely determine which of two outcomes would be best. However, Parfit was clear that he did not believe this was the correct response to such imprecision. For instance, in response to a different objection along similar lines, he writes that "we can reply that, if global warming would kill many people, and have other bad effects, making the world in some ways worse, and these bad effects would also predictably be greater than any good effects, that would be enough to make global warming a bad thing, which it might be our duty to try to prevent, or limit. Such claims would not be undermined if we can also predict that ... these two possible futures would be less precisely comparable, so that neither future would be worse all things considered" (Parfit ms2, 15-6). Thus, while it may weaken the absolute precision with which we can say that certain outcomes would be bad, this imprecision may not weaken our reasons for working to avoid them.

[69] (Parfit 2017a, 436)

[70] Parfit continues: "We ought to transfer to these people … at least ten percent of what we inherit or earn" and provides an endnote with practical advice about the most effective ways of doing this.

> slavery and the other injustices on which these achievements were partly based.[71]

For some people, this passage may seem troubling and highlight what is wrong with the perfectionist implications of Parfit's view—that it gives insufficient weight to suffering. How could the harm of slavery be considered imprecisely equally as important as the achievements of the ancient Greeks?

However, this reply misses an important implication of this passage. As we have seen, Parfit's Imprecise Lexical View implies that, all else being equal, the loss of any of the Best Things in Life, such as the achievements of the ancient Greeks, would be a change for the worse no matter how much additional wellbeing this created. Yet, in this case, he suggests that this is not so and that the loss of these things would only make our history imprecisely equally as good (or in his later terminology 'roughly comparable') to what it has been. There is thus no contradiction between Parfit's Imprecise Lexical View and his belief that this form of slavery was an inexcusable crime.[72] If this is so, then it follows that the badness of slavery is not only a function of its negative effect on people's wellbeing but must also involve qualitative changes to the value of outcomes that are in some way comparable to those of losing the Best Things in Life.

Parfit briefly addresses this issue at two other points in his work. Firstly, as we have seen, it is related to his earlier belief that Perfectionism appeared elitist and that we should reject "the Nietzschean view that the prevention of great suffering can be ranked wholly below the preservation of creation of the best things in life".[73] While at first he saw rejecting this concern as 'irrelevant' to his considerations, he appears to have returned to the issue again in 2016, noting that "if we care greatly about the quality of life, being in this sense Perfectionists, that would not make us elitists, who care most about the well-being of the best-off people".[74] We have already seen that Parfit saw the Quality of Life as being a broader notion than wellbeing. However, this assertion that his Imprecise Lexical View would not be elitist seems to suggest that its breadth may well incorporate more than simply wellbeing and the Best Things in Life.[75]

The other place in Parfit's work where he seems to connect the value of the Best Things in Life and the badness of suffering comes right at the end of *On What Matters, Volume Three*, where he chooses to close his argument for the moral importance of avoiding human extinction with the following observation:

> If we are the only rational beings in the Universe, as some recent evidence suggests, it matters even more whether we shall have descendants or successors during the billions of years in which that would be possible. Some of

[71] (Parfit ms2, 29)

[72] The nature of this crime is not considered further by Parfit but would seem to depend on the distinction he sometimes draws between the goodness of outcomes and what we ought to do. It would thus be similar to our potential obligations to benefit presently existing people, even when we could give a greater benefit to future people (discussed in footnote 37), and to do what would have predictably the best results, even if we cannot make precise judgements of the relative value of our choices (discussed in footnote 68). Since these do not relate to the goodness of outcomes, they do not form part of Theory X.

[73] (Parfit 1986, 20)

[74] (Parfit 2016, 117)

[75] Indeed, elitism is not the only charge that can be levelled against a purely perfectionist account of Quality of Life. Despite his statements to the contrary, Perfectionism on its own may also be unable to avoid the Repugnant Conclusion and may imply a version of Parfit's Ridiculous Conclusion, while more sophisticated accounts of the Quality of Life can also avoid these troubling implications as well—see esp. (Beard 2019).

> our successors might live lives and create worlds that, though failing to justify past suffering, would have given us all, including those who suffered most, reasons to be glad that the Universe exists.[76]

We, therefore, believe that, though he never stated this clearly, a special concern for the alleviation of suffering may be a feature of Parfit's Imprecise Lexical View, and hence of Theory X. Again, we believe that further work on this aspect of Theory X is needed.[77]

[76] (Parfit 2017a, 437)

[77] In this section, we have only considered how Theory X might account for the badness of suffering in terms of its impact on people's Quality of Life. Another aspect of Theory X on which further work is needed is how it fits in with Parfit's views on distributive ethics. Parfit famously argued against egalitarianism as a distributive principle and in favour of what he called 'The Priority View', according to which "we have stronger reasons to benefit people the worse off these people are" (Parfit 2012, 401).

However, it is not clear how this view relates to existential benefits. On the one hand, Parfit argued that "[it] is sometimes claimed, for example, that we have prioritarian reasons to have children, since we would thereby benefit some of the possible people who would otherwise be badly off, by never existing. …. But when we apply these distributive principles, we should not include, among the people who are badly off, possible people who never exist. Like the Principles of Personal Good, or Pareto Principles, the Prioritarian Principles that I have considered cannot be applied to cases in which, in the different possible outcomes, different people would exist" (Parfit 2012, 440).

On the other hand, in an unpublished draft of the paper that would become (Parfit 2017b), he begins considering a version of the priority view that would apply to these people. He writes:

> Suppose instead that the possible outcomes are these:
>
> > A: Tom's total will be negative 100, Dick's total will be 100
> >
> > C: Tom and Dick will never exist
>
> Though having zero well-being is different from never existing, this difference seems unimportant here. Most of us would believe that A would be worse than C. If someone's total level of well-being would be below zero, this person's life would be bad to live, and worse than lives that are not worth living. It would be better if such lives are not lived because such people never exist. In defending our belief that A would be worse than C, we could not appeal to the simplest version of the Priority View .... We could appeal, however, to another version of the Priority View. We could claim that
>
> > (J) the badness of someone's being existentially harmed is greater than the goodness of someone's receiving an equally great existential benefit.
>
> If someone is existentially harmed by being caused to have a life that is intrinsically bad, this person would be worse off than someone who is existentially benefited by being caused to have a life that is intrinsically good. Since benefiting people does more good the worse off these people are, preventing people from being existentially harmed does more good than giving people equally great existential benefits. That is why A would be worse than C (Parfit ms3, 5).

Parfit does not make it clear whether he endorses this latter view. However, he appears to view it more sympathetically than the alternative view he is considering at this point in the paper, that there is a fundamental asymmetry between existential benefits and existential harms. Our view is that Parfit may have wished to incorporate prioritarianism into his theory, so long as he could do so in a way that would not give us additional reasons to have children. One proposal towards this end would be to apply the

4. Conclusion

In this chapter, we have argued that, although he did not live long enough to complete it, Parfit's work contains all the components necessary to construct a Theory X that would fulfill the requirements of his lifelong search. The combination of a Wide Dual Person-Affecting Principle, concerning the value of wellbeing, and his Simple and Imprecise Lexical Views, which place this into a wider pluralism about the value of lives, can avoid the Absurd and Repugnant Conclusions, solve the Non-Identity Problem and escape the Mere Addition Paradox. This combination can also be applied to those cases that Parfit asserted mattered most, and support his conclusions about them.

Naturally, in combining these principles and applying them to these cases, we have come across issues in need of further work. Yet, these represent not only problems for this theory to overcome but also opportunities to break fresh ground in moral philosophy. Thus, while the search for Theory X may not quite be over, we believe that Parfit has laid the foundations for a truly innovative, insightful and compelling approach to assessing and responding to the global challenges we now face. If it is not presumptuous for us to say, we feel that, for Parfit, as for Sidgwick before him, the words of F. W. H. Myers are appropriate:

> He pointed to a definite spot; he vigorously drove in the spade; he upturned a shining handful; and he left us as his testament, *Dig Here*.[78,79]

priority view to Parfit's Wide Individual Principle but not his Wide Collective Principle, yielding the following:

> *Prioritarian Dual Wide Person Affecting Principle.* One of two outcomes would be in one way better if this outcome would together benefit people more, and in another way better if this outcome would benefit each person more, where benefiting each person matters more the worse off these people are.

Another proposal towards this end would be to apply the priority view only to people whose lives are bad but not those whose lives are good.

[78] (Schultz 2004, 719)

[79] We would like to thank the editors of this volume, Jeff McMahan, Tim Campbell, James Goodrich and Ketan Ramakrishnan. Jeff McMahan and Tim Campbell also provided feedback on earlier drafts of this work, as did Larry Temkin, Michael Otsuka, Shlomi Segall and participants at the workshop "Evaluating Extreme Technological Risk", hosted by the Centre for the Study of Existential Risk at the University of Cambridge, as well as colleagues at a work in progress seminar, all of whom we gratefully acknowledge. S. J. Beard worked on this paper while a Visiting Scholar at the Centre for Philosophy of Natural and Social Science at the London School of Economics. Patrick Kaczmarek received support during the writing of this paper from the Berkeley Existential Risk Initiative and the Effective Altruism Foundation. This publication was made possible through the support of a grant from Templeton World Charity Foundation, Inc. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of Templeton World Charity Foundation, Inc.

References

Arrhenius, G., An Impossibility Theorem for Welfarist Axiologies, in *Economics and Philosophy* 16/2 (2000): 247-266.

——————— The Very Repugnant Conclusion, in K. Segerberg and R. Sliwinski (eds.), *Logic, Law, Morality: thirteen essays in practical philosophy in honour of Lennart Åqvist* (Uppsala University, 2003), 167-180.

——————— The Impossibility of a Satisfactory Population Ethics, in E. Dzhafarov and L. Perry (eds.), *Descriptive and Normative Approaches to Human Behavior* (World Scientific, 2011), 51-66.

——————— Population Ethics and Different-Number-Based Imprecision, in *Theoria* 82/2 (2016): 166-181.

Beard, S. Perfectionism and the Repugnant Conclusion, in *The Journal of Value Inquiry* 54 (2019): 119-140.

Beard, S. and Kaczmarek, P., The Wrongness of Human Extinction, in *Argumenta* 5/1 (2020): 85-97.

Blackorby, C., Bossert, W. and Donaldson, D., *Population Issues in Social Choice Theory, Welfare Economics, and Ethics* (Cambridge University Press, 2005).

Boonin, D., *The Non-Identity Problem and the Ethics of Future People* (Oxford University Press, 2014).

Broome, J., *Weighing Lives* (Oxford University Press, 2004).

Cowen, T., What Do We Learn from the Repugnant Conclusion?, in *Ethics* 106/4 (1996): 754-775.

Crisp, R., Utilitarianism and the Life of Virtue, in *The Philosophical Quarterly* 42/167 (1992): 139-160.

Huemer, M., In Defence of Repugnance, in *Mind* 117/468 (2008): 899-933.

——————— Lexical Priority and the Problem of Risk, in *Pacific Philosophical Quarterly* 91/3 (2010): 332-351.

Hurka, T., Value and Population Size, in *Ethics* 93/3 (1983): 496-507.

Lazar, S. and Lee-Stronach, C., Axiological Absolutims and Risk, in *Noûs* 53/1 (2017): 97-113.

Masny, M., On Parfit's Wide Dual Person-Affecting Principle, in *The Philosophical Quarterly* 70/278 (2020): 114-139.

McTaggart, J. and McTaggart, E., *The Nature of Existence Volume II* (Cambridge University Press, 1927).

Mulgan, T., Utilitarianism for a Broken World, in *Utilitas* 27/1 (2015): 92-114.

Parfit, D., On Doing the Best for Our Children, in M. Bayles (ed.), *Ethics and Population* (Schenkman Pub. Co., 1976), 100-115.

——————— Future Generations: Further Problems, in *Philosophy & Public Affairs* 11/2 (1982): 113-172.

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

——————— Overpopulation and the Quality of Life, in P. Singer (ed.), *Applied Ethics* (Oxford University Press, 1986), 145-164.

——————— *On What Matters, Volume 1* (Oxford University Press, 2011).

——————— *On What Matters, Volume 2* (Oxford University Press, 2011).

——————— Another Defense of the Priority View, in *Utilitas* 24/3 (2012): 399-440.

——————— Can We Avoid the Repugnant Conclusion?, in *Theoria* 82/2 (2016): 110-127.

——————— *On What Matters, Volume 3* (Oxford University Press, 2017).

——————— Future People, The Non-Identity Problem, and Person-Affecting Principles, in *Philosophy & Public Affairs* 45/2 (2017): 118- 157.

———————*Towards Theory X: Part One* (unpublished ms1).

———————*Towards Theory X: Part Two* (unpublished ms2).

——————— *The Non-Identity Problem* (unpublished ms3).

Rachels, S., Repugnance or Intransitivity: A Repugnant But Forced Choice, in T. Tännsjö, & J. Ryberg (eds.), *The Repugnant Conclusion* (Springer, 2004), 163-186.

Schultz, B., Henry Sidgwick—Eye of the Universe: An Intellectual Biography (Cambridge University Press, 2004).

Sidgwick, H., *The Methods of Ethics, Seventh Edition* (Hackett Publishing Company, 1907).

Tännsjö, T., Why We Ought to Accept the Repugnant Conclusion, in *Utilitas* 14/3 (2002): 339-359.

Temkin, L., Intransitivity and the Mere Addition Paradox, in *Philosophy & Public Affairs* 16/2 (1987): 138-187.

Thomas, T., Some Possibilities in Population Axiology, in *Mind* 125/507 (2017): 807-832.