The weight of suffering

Andreas Mogensen (Global Priorities Institute, University of Oxford)

Global Priorities Institute | May 2022

GPI Working Paper No . 4-2022

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

The weight of suffering

Abstract: How should we weigh suffering against happiness? This paper highlights the existence of an argument from intuitively plausible axiological principles to the striking conclusion that in comparing different populations, there exists some depth of suffering that cannot be compensated for by any measure of well-being. In addition to a number of structural principles, the argument relies on two key premises. The first is the contrary of the so-called *Reverse Repugnant Conclusion*. The second is a principle according to which the addition of any population of lives with positive welfare levels makes the outcome worse if accompanied by sufficiently many lives that are not worth living. I consider whether we should accept the conclusion of the argument and what we may end up committed to if we do not, illustrating the implications of the conclusions for the question of whether suffering in aggregate outweighs happiness among human and non-human animals, now and in future.

1. Introduction

There is both great happiness and great suffering in this world. Which has the upper hand? Does the good experienced by human and non-human animals in aggregate counterbalance all the harms they suffer, so that the world is morally good on balance? Or is the moral weight of suffering greater?

To answer this question, we need to know how to weight happiness against suffering from the moral point of view. In this paper, I present an argument from intuitively plausible axiological principles to the conclusion that in comparing different populations, there exists some depth of lifetime suffering that cannot be counterbalanced by any amount of well-being experienced by others. Following Ord (2013), I call this view *lexical threshold negative utilitarianism* (LTNU). I don't claim that we should accept LTNU. My aim is to explore different ways of responding to the argument. As we'll see, the positions at which we may arrive in rejecting its premises can be nearly as interesting and as striking as the conclusion.

In section 2, I define LTNU more rigorously and set out the argument. It relies on a number of structural principles governing the betterness relation on populations, together with two key premises. The first is the contrary of what Carlson (1998) and Mulgan (2002) call the *Reverse Repugnant Conclusion* (RRC). The second says, roughly, that the addition of lives with positive welfare levels makes the outcome worse if accompanied by sufficiently many lives that are not worth living. In section 3, I consider whether we should be willing to accept the argument's conclusion, especially given that LTNU has been thought to entail the desirability of human extinction or the extinction of all sentient life (Crisp 2021). In section 4, I discuss our options for rejecting the argument's structural principles. I argue that our options for avoiding the disturbing implications of LTNU discussed in section 3 are limited if we are restricted to rejecting one or more of these principles. In section 5, I consider the possibility of rejecting the first of the key non-structural premises. I focus on the possibility of rejecting the contrary of RRC without accepting RRC. This, I claim, is also not promising, considered as a way of avoiding the disturbing implications of LTNU discussed in section 3. I will have nothing original to say about RRC *per se*, except that the overarching argument of this paper may be taken as a reason to accept it. In section 6, I consider the possibility of rejecting the last remaining premise. Specifically, I consider the possibility that there are some lives so good that their addition to the population can justify the addition of any number of lives that are only barely not worth living, and the independent interest of this hypothesis to the project of reckoning the overall goodness of the world.

2. An Argument for LTNU

Before I present the argument for LTNU, I have to define the position more rigorously.[1]

I assume the existence of a totally ordered set of lifetime welfare levels, $\mathbb{L}$. I also assume the existence of a privileged zero level in $\mathbb{L}$. This is the level above which a life is worth living and below which it is not worth living.

Define a *population* to be a function from $\mathbb{L}$ to the natural numbers, telling us for each lifetime welfare level how many people are at that level. I'll use uppercase letters 'X', 'Y', 'Z' to denote populations, but when I need to convey specific information, I'll use the following notation. For any lifetime welfare level L in $\mathbb{L}$, $[L]$ is a population of exactly one person at L. For any natural number, m, $m[L]$ is a population of m people at L. I use the '+' symbol to denote combinations of populations. Thus, for any L and L' in $\mathbb{L}$, $m[L] + n[L']$ is a population of m people at L and n distinct people at L'.

I use the symbol '$\succcurlyeq$' to order populations in terms of moral value, taking '$X \succcurlyeq Y$' to mean 'population X is morally at least as good as population Y.' I define 'morally better than' ('$\succ$') in the standard way: $X \succ Y$ is equivalent to the conjunction $X \succcurlyeq Y$ and $Y \not\succcurlyeq X$. If there exist populations X and Y such that neither $X \succcurlyeq Y$, nor $Y \succcurlyeq X$, I shall write '$X \bowtie Y$'.

Using this notation, I define LTNU as follows:

LTNU: There is some extremely negative lifetime welfare level, $L_{tortured}$, such that for any population, X, it is the case that for any population, Y, consisting exclusively of lives with positive welfare levels, $X \succ X + Y + [L_{tortured}]$.

[1] In order to do so, I borrow notational conventions from Thomas (ms).

In other words, there is a certain kind of tortured life, such that it is better not to add even a single life like that to the population, no matter how many however good lives will be added alongside it.

I show how we can derive LTNU from a pair of individually plausible premises, given a small number of structural assumptions about the betterness relation on populations. The structural assumptions are as follows:

C: For any populations, X, Y: $X \succcurlyeq Y$ or $Y \succcurlyeq X$.

T: For any populations, X, Y, Z: if $X \succcurlyeq Y$ and $Y \succcurlyeq Z$, then $X \succcurlyeq Z$.

S: For any populations X, Y, Z: $X \succcurlyeq Y$ iff $X + Z \succcurlyeq Y + Z$.

The first principle, C, asserts that the relation expressed by 'morally, at least as good as' is *complete*, in the sense that for any two populations, we can always compare them by saying that one is at least as good as the other. The next, T, tells us that the relation 'morally, at least as good as' is *transitive*. Finally, S asserts that the value of a population is *separable*, in the sense that the ranking of one population relative to another doesn't vary as a result of how things stand with respect to some unaffected background population.

The two additional premises on which I rely are as follows. Firstly, we have the contrary of what Carlson (1998) and Mulgan (2002) call the *Reverse Repugnant Conclusion* (RRC). In Mulgan's formulation, RRC states that given a population "where ten billion people live long lives of unalloyed excruciating agony", there is a morally worse population in which "a vast number of people have lives which are almost but not quite worth living" (362).

According to Mulgan, this conclusion is at least as repugnant as the original *Repugnant Conclusion (*RC) formulated by Parfit (1984: 388): "For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better, even though its members have lives that are barely worth living." Assume, therefore, that we accept

CON-RRC: There is some extremely negative lifetime welfare level, $L_{tortured}$, and some positive integer m, such that for any merely mildly negative lifetime welfare level, L_{lousy}, and any positive integer n, $n[L_{lousy}] \succ m[L_{tortured}]$.

Here I use 'merely mildly negative' to denote the kind of lifetime welfare level at which a life is supposed to be, in Mulgan's phrase, 'almost but not quite worth living.' Thus, CON-RRC tells us that there is some population of terrible, tortured lives whose existence would be worse than any population of people whose lives are almost but not quite worth living.

Secondly, we have the claim that the value of additional good lives can always be outweighed by sufficiently many lives that are not worth living. More formally,

AO:[2] For any population X, it is the case that for any population, Y, consisting exclusively of lives with positive welfare levels, and any negative lifetime welfare level, L_{bad}, there is some n such that $X \succ X + Y + n[L_{bad}]$.

[2] Here 'AO' stands for '*a*lways *o*utweighable'.

The idea here is that, given any background population, adding some number of people who all have excellent lives and some number of people with negative lifetime welfare levels is for the worse if the latter are sufficiently numerous, and this holds true for any negative lifetime welfare level.

I think each of these premises is plausible on its face. Many find RRC very hard to believe. I assume they find CON-RRC plausible as a result. There is obviously the option to assert only the weaker claim that the existence of ten billion people living long lives of unalloyed excruciating agony is not better than – but not necessarily worse than - a vast number of people with lives that are almost but not quite worth living. I expect most who find RRC unacceptable find this too weak. I also think very many people will find it intuitively plausible to think that there are no lives so valuable that for their sake we should be willing to accept that arbitrarily many individuals will have to have lives that are not worth living. They might not be so confident if asked to imagine that these lives are *only just* not worth living. Still, I expect that many won't budge.

In due course, we'll examine our options for rejecting the two premises, along the lines discussed in the previous paragraph. What I want to show now is how LTNU follows from CON-RRC and AO given C, T, and S.

Informally, the argument works as follows. First, borrowing techniques from Jensen (2008) and Nebel (2022), I'll explain how, given C, T, and S, we can strengthen CON-RRC to get

CON-RRC*: There is some extremely negative lifetime welfare level, $L_{tortured}$, such that for any merely mildly negative lifetime welfare level, L_{lousy}, and any positive integer n, $n[L_{lousy}] \succ [L_{tortured}]$.

This tells us that a population of just a single life lived at a tortured level is worse than a population of any number of people whose lives are only barely not worth living.

The remainder of the argument then relies on the following idea. AO tells us that no number of additional happy lives can compensate for the addition of arbitrarily many merely mildly negative lives. CON-RRC* tells us that a single tortured life is even worse than arbitrarily many merely mildly negative lives. We infer that no number of additional happy lives can compensate for even a single additional life at the tortured welfare level.

Here is the proof stated more formally. I'll start by showing how to prove CON-RRC* from C, T, S, and CON-RRC. The proof is by *reductio*. Assume that CON-RRC* is false. Given C, it follows that there is some k such that

$$[L_{tortured}] \succcurlyeq k[L_{lousy}] \tag{2.1}$$

From S and (2.1), it follows that

$$2[L_{tortured}] \succcurlyeq k[L_{lousy}] + [L_{tortured}] \tag{2.2}$$

Similarly, from S and (2.1), it follows that

$$k[L_{lousy}] + [L_{tortured}] \succcurlyeq 2k[L_{lousy}] \tag{2.3}$$

Given T and (2.2) and (2.3), we infer that

$$2[L_{tortured}] \succcurlyeq 2k[L_{lousy}] \tag{2.4}$$

We can re-iterate this line of argument, relying on S to add either k lives at L_{lousy} or 1 life at $L_{tortured}$ to either side of (2.4), and then appealing to T to conclude that 3 lives at $L_{tortured}$ are at least as good as $3k$ lives at L_{lousy}. By iterating through $m-3$ additional steps, we arrive at the conclusion that $m[L_{tortured}] \succcurlyeq n[L_{lousy}]$, where $n = m \times k$. This contradicts CON-RRC. So CON-RRC entails CON-RRC* given C, T, and S.

Here is the second half of the proof, which shows that CON-RRC*, AO, T, and S together entail LTNU. From AO, for any population, X, any population Y, consisting exclusively of lives with positive welfare levels, and any merely mildly negative lifetime welfare level, L_{lousy}, we know there is some n such that

$$X \succ X + Y + n[L_{lousy}] \tag{2.5}$$

Given CON-RRC*, we have

$$n[L_{lousy}] \succ [L_{tortured}] \tag{2.6}$$

By S and (2.6), we infer that

$$X + Y + n[L_{lousy}] \succ X + Y + [L_{tortured}] \tag{2.7}$$

From (2.5), (2,7), and T, we infer the desired conclusion

$$X \succ X + Y + [L_{tortured}] \tag{2.8}$$

3. Following the Argument Where it Leads

LTNU follows from CON-RRC and AO given C, T, and S. Each of these principles, I've claimed, is to some extent intuitively plausible. Should we be willing to follow the argument where it leads?

Let's begin by considering whether LTNU might be independently plausible. It may be thought that it is and/or can be seen to be so by asking the reader to consider thought experiments like that found in Ursula K. LeGuin's well-known short story, "The Ones Who Walk Away from Omelas" (LeGuin 1973/1991) or its antecedents in James (1891) and Dostoevsky (1880/1994), as is suggested by Mayerfeld (1996: 327-8, 1998: 148).[3] However, I'm not persuaded of this claim.

In LeGuin's story, Omelas is initially presented to the reader as an unblemished utopia. Its citizens are said to enjoy "boundless and generous contentment, a magnanimous triumph felt not against some outer enemy but in communion with the finest and fairest in the souls of all men everywhere and the splendor of the world's summer" (LeGuin 1973/1991: 2). There turns out to be a catch. The reader is told that "under one of the beautiful public buildings of Omelas, or perhaps in a cellar of one of its spacious private homes," (ibid.: 3) there is a room in which a lone child suffers. "The child used to scream for help at night, and cry a good deal, but now it only makes a kind of whining … It is so thin there are no calves to

[3] Note that Mayerfeld focuses exclusively on the thought experiment in James (1891). I focus on LeGuin's, as I find it more vivid and powerful.

its legs … It is naked. Its buttocks and thighs are a mass of festered sores, as it sits in its own excrement continually." (ibid.: 4) If the child were to be spared her suffering, "in that day and hour all the prosperity and beauty and delight of Omelas would wither and be destroyed. Those are the terms." (ibid.: 4) The citizens all come eventually to know that "their happiness, the beauty of their city … depend wholly on the child's abominable misery." (ibid.: 4) For the most part, they reconcile themselves to this fact. However, sometimes a citizen of Omelas will instead "go out into the street, and walk down the street alone. … They leave Omelas, they walk ahead into the darkness, and they do not come back." (ibid.: 5)

Many readers sympathize with this reaction. They too would not accept those terms - or hope they would not. It might seem they must therefore believe, in accordance with LTNU, that there is a depth of suffering so great that it cannot be compensated for by any measure of bliss.

I disagree that pre-theoretic credence in LTNU best explains our intuitive reaction to LeGuin's story. It seems to me most plausible that we sympathize with those who walk away from Omelas because we judge intuitively that what is done to the child is *wrong*. Since we need not believe consequentialism, that judgment need not entail any axiological hypothesis, such as LTNU. The choice made by those who walk away is most naturally rationalized by agent-relative reasons to avoid complicity in wrongdoing. It is hard to rationalize as promoting the good, since it is not apparent what good is achieved by leaving the city and never returning. The child still suffers. Rather than a lexical axiological hypothesis, our response to LeGuin's story may reflect intuitive sympathy for the postulate of an absolute

agent-centred restriction against certain ways of using an innocent person as a means to the benefit of oneself and others.[4]

For his part, Mayerfeld (1996: 327-8) claims that his intuitions about this sort of case survive the removal of the assumption that the one's suffering is used as a means to the benefit of others.[5] But I, at least, don't share that intuition. If the child is imagined as inhabiting a far-off country, and if the boundless and generous contentment of Omelas is imagined as independent of her suffering, except in that it would have to be destroyed in the process of working to spare her from her misery, then I don't find I have the same reaction as before. I don't judge intuitively that the beauty and delight of Omelas does not suffice to allow the child's suffering to go on.

Let's now consider objections to LTNU, or, more exactly, a proper subset thereof. I set aside objections that apply to lexical axiological theories in general. For example, LTNU is subject to the objection that lexical axiological theories cannot be extended to cases involving risk or uncertainty without absurdity (Huemer 2010). I set aside this kind of objection here because the argument presented in the previous section uses a lexical axiological hypothesis as a premise, in the form of CON-RRC. I discuss the option of rejecting CON-RRC in section 5. For now, I concentrate on objections to LTNU that are not also independent objections to the acceptability of the individual premises of the argument.

---

[4] This kind of interpretation is suggested by Hooker (2002: 129 n.7).

[5] In line with footnote 3, I should be clear that, strictly speaking, Mayerfeld is discussing the analogous thought experiment in James (1891) from which LeGuin drew inspiration, and not LeGuin's story specifically.

Regardless of this restriction, the greatest cost of accepting LTNU is surely that it appears to support the desirability of human extinction or the extinction of all sentient life (Crisp 2021). In this respect, it resembles its namesake. The term 'negative utilitarianism' is most closely associated with Popper's moral and political philosophy. Popper (1966/2011: 602) claims that "pain cannot be outweighed by pleasure" and that our goal should be "the least amount of avoidable suffering for all." Taken literally, this suggests a moral view on which one ought to choose the best outcome, defined as the outcome with the lowest total ill-being. Call this view *classical negative utilitarianism* (CNU). Given CNU, we ought to bring about the immediate extinction of all sentient life, if we can, since doing so will minimize total suffering (Smart 1958). Many treat this as a knock-down objection.[6]

It may be thought that LTNU supports the same kind of conclusion and thus opens itself up to the same kind of objection. It is a virtual certainty, we might think, that there will exist additional people whose lifetime welfare falls below the relevant threshold for a tortured life. Therefore, given LTNU, it would be better if there never again came to exist anyone at all.

Now, there are at least two reasons why this objection to LTNU is not as strong as the corresponding objection to CNU. Firstly, the conclusion that there will exist additional people whose lifetime welfare falls below the relevant threshold for a harrowed life is not forced on us. While they might have the status of mere theoretical curiosities, we can imagine interpretations of LTNU that set the threshold for a life so wretched that its addition cannot be compensated for by any increase in well-being at so utterly abysmal a level of suffering

[6] But see Knutsson (2021).

that we may be sufficiently uncertain whether so horrible a life will ever be lived.[7] Secondly, CNU is a consequentialist normative ethical theory, whereas LTNU does not presuppose consequentialism, nor any other theory of normative ethics. It is a fragment of a population axiology. It does not of itself suggest any conclusions about what we ought to do. We may have powerful reasons to conserve humanity that are independent of population axiology, as highlighted in recent work by Scheffler (2013, 2018) and Frick (2017).

I expect many will nonetheless continue to believe that LTNU has unacceptable implications concerning the evaluation of extinction. Let's therefore consider our options for declaring the argument in section 2 unsound.

4. Rejecting the Argument's Structural Principles

We'll start by considering which among the structural principles of the argument we may be willing to reject.

For my own part, I cannot take seriously the idea of giving up T (*pace* Rachels 1998; Temkin 1996, 2012). I also think that S is extremely compelling, if less so than T. For example, I do not know of anyone who has claimed, as Broome (2004) does of T, that S is an analytic truth. Nonetheless, rejecting S carries serious costs. To do so is to open ourselves up to the possibility that the moral desirability of population changes within our power may depend on how things stand in regions of spacetime that we cannot affect. Granted, there are

[7] By 'sufficiently uncertain,' I mean unsure enough that the desirability of extinction does not follow given the right way of adapting LTNU to contexts involving risk, whatever exactly that is. On accommodating lexical axiological theories to contexts involving risk, see Lazar and Lee-Stronach (2019), Kosonen (2021), Nebel (2022), and Lee-Stronach (2022).

axiological theories that yield this result, such as *average utilitarianism* (AU) (Pressman 2015) and *variable value theories* (VVT) (Ng 1989; Hurka 1983). But this fact is typically wielded against them. As Parfit (1984: 420) memorably observes, "research in Egyptology cannot be relevant to our decision whether to have children."[8]

What about the option of rejecting C? Unlike rejecting T or S, this strikes me as plausible. In other words, it strikes me as plausible that some population outcomes are neither better than, worse than, nor exactly equally as good as one another (Chang 2016, 2022; Parfit 1984: 430-2, 2016). However, giving up C is not without cost. Plausibly, if C is false, a beneficent agent is warranted in having incomplete preferences over population outcomes. There are well-known problems and paradoxes associated with incomplete preference relations (Chang 1997; Broome 2000; Hare 2010; Schoenfield 2014; Gustafsson 2022). Furthermore, giving up C does relatively little to blunt the force of the argument presented in section 2.

C is only needed in the first half of the proof, where it is used in strengthening CON-RRC to obtain CON-RRC*. Given suitable modifications, the second half of the proof survives. More exactly, it can be shown that given S and T, CON-RRC and AO entail a weak form of lexical threshold negative utilitarianism, defined as follows:

WLTNU: There is some extremely negative lifetime welfare level, $L_{tortured}$ and some positive integer m, such that for any population, X, and any population, Y, consisting exclusively of lives with positive welfare levels, $X \succ X + Y + m[L_{tortured}]$.

[8] Compare McMahan (1981: 115). Thomas (2022: 278-80) provides a formal statement of the 'Argument from Egyptology.'

The proof is obtained through trivial modifications of (2.5) - (2.8) and is omitted.

If we believe that LTNU has unacceptable implications concerning the survival of humanity, I expect we will react similarly to WLTNU. WLTNU expands our options for avoiding the unacceptable implications that seem to follow from LTNU only in the following way. I noted in section 3 that the conclusion that there will exist additional people whose lifetime welfare falls below the relevant threshold for a harrowed life is not forced on us. We can imagine interpretations of LTNU that set the threshold for such a life at so utterly abysmal a level of suffering that we may be sufficiently uncertain whether so horrible a life will ever be lived. If we accept the weaker claim, we have more room to sow doubt about the desirability of extinction by specifying that the positive integer m quantified over in WLTNU is some very high number. But we may not have much additional room for manoeuvre. Recall, for example, that Mulgan relies on the figure of ten billion in his formulation of RRC. If WLTNU is inferred from CON-RRC given S, T, and AO, this would then be a reasonable value for the positive integer m quantified over in the statement of WLTNU. However, given the potential size of the future, ten billion people is a relatively small number. It has been claimed that, even with conservative estimates in place, the expectation of the total number of future people is 10^{28} – that is, 10 billion billion billion (Newberry 2021).

I conclude that our options for avoiding the disturbing implications of LTNU discussed in section 3 are limited if we are restricted to rejecting one or more of the structural assumptions of the argument set out in section 2. We must instead reject CON-RRC, AO, or both.

5. Rejecting CON-RRC

If we reject CON-RRC, a natural option would be to accept RRC. While not as widely debated as RC, RRC is a well-known principle within population axiology. It is typically wielded as an objection against *total utilitarianism* (TU) and/or *critical-level utilitarianism* (CLU) (Carlson 1998; Mulgan 2002; Gustafsson 2020; Nebel 2022). I expect most philosophers working in this area have a pre-set view about its acceptability. For example, Blackborby, Bossert, and Donaldson (2005: 135 n.7) claim not to be able to see why RRC is repugnant. I haven't anything original to say as regards the acceptability of RRC. Or, rather, I haven't anything original to say over and above noting that the argument of this paper may be taken as a reason to accept RRC in spite of any intuitive repugnance, and to go a little easier on TU and CLU. Doing otherwise may be thought to yield unacceptable results concerning the value of human survival, given other reasonable assumptions.

We can, however, consistently reject both CON-RRC and RRC. We can coherently claim that a population of sufficiently many tortured lives is neither always worse than - and nor ever better than - a population of lives that are only barely not worth living. However, if we reject both CON-RRC and RRC, then we are forced to give up C, given an additional, highly plausible assumption. Moreover, if we do not reject any of the remaining premises of the argument, we are forced to conclude, roughly, that it would not be worse for all of sentient life to be wiped out instantaneously if the future contains some absolute number of tortured lives.

Let me firstly explain why we are forced to give up C. Assume we reject RRC and so deny that for any m, there is some n such that $m[L_{tortured}] \succ n[L_{lousy}]$. Given C, this entails that for any m, n

$$n\big[L_{lousy}\big] \succcurlyeq m[L_{tortured}] \tag{5.1}$$

Furthermore, given that we reject CON-RRC, C also entails that for some n^*,

$$m[L_{tortured}] \succcurlyeq n^*\big[L_{lousy}\big] \tag{5.2}$$

Now, the following negative mere addition principle is extremely plausible:

NMA: For any population X, and any negative welfare level L_{bad}: $X \succ X + [L_{bad}]$

In other words, we always make the outcome worse by adding a life that is not worth living. It follows by NMA and (5.2) that for any n such that $n > n^*$

$$n^*\big[L_{lousy}\big] \succ n\big[L_{lousy}\big] \tag{5.3}$$

By T, from (5.2) and (5.3), we infer that that

$$m[L_{tortured}] \succ n\big[L_{lousy}\big] \tag{5.4}$$

for any $n > n^*$. But this contradicts (5.1). Thus, if we reject both CON-RRC and RRC, we must reject either C or NMA. I think it is preferable to reject C. We should posit that for some m, and for some n^*, for any $n > n^*$

$$m[L_{tortured}] \bowtie n[L_{lousy}] \tag{5.5}$$

As noted, we may not view giving up C as an especially significant cost. However, there are other, more costly conclusions that may be forced on us. In particular, by modifying (2.5) - (2.8), we can argue from (5.5) to the conclusion that it would not be worse for all of sentient life to be wiped out instantaneously if some sufficient absolute number of tortured lives will otherwise eventually be added to the population even, if all other future persons are extremely well-off and arbitrarily numerous.

Here is the argument. Given AO, for any population, X, any population Y, consisting exclusively of lives with positive welfare levels, and any merely mildly negative lifetime welfare level, L_{lousy}, there is some n such that

$$X \succ X + Y + n[L_{lousy}] \tag{5.6}$$

where n can be chosen so that $n > n^*$. By (5.5), it follows that

$$m[L_{tortured}] \not\succcurlyeq n[L_{lousy}] \tag{5.7}$$

Because S is a biconditonal, it follows from S and (5.7) that

$$X + Y + m[L_{tortured}] \not\succcurlyeq X + Y + n[L_{lousy}] \tag{5.8}$$

From (5.6), (5.8), and T, we can then infer that

$$X + Y + m[L_{tortured}] \not\succcurlyeq X \tag{5.9}$$

since if $X + Y + m[L_{tortured}] \succcurlyeq X$, it follows by (5.6) and T that (5.8) is false. Therefore, it would not be worse for all sentient life to be wiped out instantaneously, if the future contains some absolute number of tortured lives, no matter how many good lives will be lived alongside them, and no matter how blissful those good lives may be.[9]

While perhaps not as disturbing as the conclusions we have derived so far, I expect this is still a conclusion that many will be reluctant to endorse, insofar as they are disturbed by the conclusions considered previously. Most people, I expect, believe, roughly, that if the future will be sufficiently good for a suitably high *proportion* of future individuals, then it would be morally regrettable if humanity or sentient life in its entirely were to be wiped out within the present hour. The conclusion expressed by (5.9) tells us that this is not the case, provided that a large enough *absolute number* of future people have extremely bad lives. For reasons discussed previously, even if m is set at a number that might strike us intuitively as enormous, m individuals may still represent a miniscule fraction of the expected future

[9] It is possible to derive a strengthened version of (5.9) from CON-RRC, S, T, and AO, according to which for any population, X, and any population, Y, consisting exclusively of lives with positive welfare levels, $X + Y + [L_{tortured}] \not\succcurlyeq X$. By the argument obtained by iterating (2.1) - (2.4), we have already implicitly shown that CON-RRC, S, and T together entail that $[L_{tortured}] \not\succcurlyeq n[L_{lousy}]$ for any positive integer n. The strengthened form of (5.9) can then be derived by trivial modifications of (5.6) - (5.8). Since this argument does not rely on C, this illustrates yet another sense in which rejecting C is of itself of little help to us.

population, such that the per person probability of enduring a tortured life would have to be vanishingly small for the expected number of tortured lives to be less than m.

6. Rejecting AO

Suppose instead that we reject AO. The most natural way to do so is to posit that there is some possible population, Y, consisting exclusively of lives with positive welfare levels, and some merely mildly negative welfare level, L_{lousy}, such that for any population X, it is the case that $X + Y + n[L_{lousy}] \succcurlyeq X$ for any n. In other words, there are some lives so good that their addition to the population can justify the addition of any number of lives that are only barely not worth living.

A population axiology that yields this striking result is *Lexical Total Utilitarianism* (LTU) (Carlson 2007; Thomas 2018; Nebel 2022). LTU is like TU, in that the value of a population is the sum of individual welfare. It differs from TU (or, at least, TU as it is typically understood) in that each person's welfare level is represented by a vector of real numbers, rather than a scalar. Specifically, LTU assumes that each person's welfare is represented by a vector of the form (a, b). To a first approximation, think of a as a measure of whether someone's life goes well or badly or is neutral in respect of what is truly deep and important, such as autonomy or meaningfulness. Think of b as a measure of how well someone's life goes in respect of more trivial goods and bads. We say that if a life is only barely not worth living, $a = 0$ and $b < 0$; and if a life is only barely worth living, $a = 0$ and $b > 0$.

Assume that vectors are added piecewise: $(a, b) + (c, d) = (a + c,\ b + d)$. Thus, it makes sense to speak of the sum of each individual's welfare. It is also possible to order vectors. Simplest is the lexicographic order: $(a, b) \geq (c, d)$ iff $a > c$ or $a = c$ and $b \geq d$. Nebel

(2022) argues for a more complex ordering principle for vectors representing (sums of) individual welfare, on which for some $\Delta, \delta > 0$ $(a, b) \geq (c, d)$ iff either (1) $a - c > \Delta$, or instead (2) $a \geq c$ and (i)$b \geq d$ or (ii) $\frac{a-c}{d-b} > \frac{\Delta}{\delta}$. The goal here is to capture the intuition that in order to be better overall, one population needs to be better by some non-trivial amount in respect of what is deep and important.[10] Either way, we are able to rank (at least some) outcomes in respect of the sum of the welfare of each individual, with welfare as a vector, rather than a scalar quantity.

A population axiology that ranks outcomes in this way entails the falsity of both RC and RRC. In rejecting the assumption that welfare is a scalar quantity, it undercuts many of the impossibility theorems that have been taken to show that RC cannot be avoided without sacrificing other intuitively compelling axiological principles (Thomas 2018). Notably, then, LTU entails the falsity of AO. It counts the addition of some number of lives that are sufficiently good in respect of what is truly deep and important as making the outcome better even if accompanied by arbitrarily many lives that are not worth living, provided that they are only barely not worth living.

This may well strike us as a drawback of LTU, since we may find AO intuitively plausible. However, the argument of this paper may make us more willing to reject AO and less reluctant to accept LTU, by suggesting that adherence to AO yields unacceptable results concerning the evaluation of the survival of humanity or of all sentient life, given other reasonable assumptions.

Unlike RRC, AO is not widely discussed explicitly as a principle of population axiology. Philosophers may therefore have less settled views on its truth or falsity. The

[10] Note that this ordering entails the falsity of C given LTU.

argument of this paper may provide some impetus to adopt a considered view. Moreover, there is good reason to believe that the truth or falsity of AO may be of significant independent importance to the project of reckoning the overall goodness of the world as it currently exists. Here is why.

Almost all welfare subjects are non-human animals, very many of which live for only a short time and plausibly have only a limited capacity for well-being and limited experience of positive welfare states. Some examples of especially abundant animals (and their approximate standing populations) include ants (10,000 trillion (Hölldolber and Wilson 1995)), bristlemouth fish (100s of trillions (Broad 2015)), and chickens (20 billion (Robinson et al. 2014)). Even if the lives of these animals are worth living, they are very unlikely to exceed a level above which a life is better than only just worth living, assuming that such a level is chosen to satisfy our intuitions about RC. Thus, as one possible realization of the population of drab lives quantified over in RC, Parfit (2016: 118) imagines a population of "animals who [have] lives that [are] just worth living, because these animals [have] enough slight pleasures like those of cows munching grass or lizards basking in the sun." Conversely, we might think that such lives will be only just not worth living, if in fact they are beneath the neutral level.

The view that most sentient animals have lives that are not worth living is disturbingly plausible. Many wild non-human animals adopt life-history strategies that emphasize high rates of reproduction and minimal parental investment, producing offspring in quantities several orders of magnitude greater than the replacement rate, almost none of which survive to reproduce. Hapgood (1979: 44-45) writes: "In her lifetime a lioness might leave 20 cubs; a pigeon 150 chicks; a mouse, 1,000 kits; a trout 20,000 fry, a tuna or a cod, a million fry or more". The many offspring who do not survive to reproductive maturity may have very brief lives whose affect balance is dominated by a painful death. These many, many brief lives might

therefore very well not be worth living. If so, their sheer abundance could be thought to make it the case that total suffering exceeds total enjoyment among wild non-human animals (Ng 1995; Tomasik 2015; Horta 2017).[11] Nonetheless, because these lives are so very short, and because the juveniles of the animals that reproduce in the highest numbers are sufficiently cognitively rudimentary at the point of death that they might be barely conscious (Browning and Veit 2021), these many, many lives that are not worth living might be only barely so.

Consider, next, domesticated animals. Setting aside fish and farmed invertebrates, the population of domesticated animals consists in large part of chickens raised for meat. The standing global population of chickens is about 20 billion individuals, compared to only 1.5 billion cattle, 2 billion sheep and goats, and 1 billion hogs (Robinson et al. 2014). In 2007, nearly 9 billion 'broiler' chickens were produced for slaughter in the US, compared to an annual inventory of 320 million egg laying hens (Norwood and Lusk 2011). Broilers are typically harvested at just 4-8 weeks old. At the time of slaughter, birds have on average 0.7 square feet of space. They suffer from leg problems caused by their oversized breasts, with one study of birds just prior to slaughter finding that around 28% experience some pain due to leg problems and 3.3% are almost unable to walk (Knowles et al. 2008). Herzog (2010: 156) describes the conditions in which these broilers are raised as "Dante-esque: the chicks will never see sun nor sky. Because they are so top-heavy, broiler chickens spend most of their day lying down, often in litter contaminated with excrement. As a result, many will develop breast blisters, hock burns, and sores on their feet." Nonetheless, they have ample food and water, as well as litter for scratching. Some authors therefore maintain that broiler chickens in the US on average have lives worth living. Norwood and Lusk (2001: 131) write that "after reviewing

[11] But see Groff and Ng (2019), Browning and Veit (2021).

all the obstacles to welfare and the nature of the birds, in our assessment, broiler farms do not cause large-scale suffering." I know many who find this claim very hard to believe, because the lives of broilers seem awful. However, Norwood and Lusk's conclusions are generally even-handed and clearly explained, and they do not shy away from condemning the lives of many farmed animals as worse than death.[12] If in fact they are wrong, I would expect that they are not too far wrong. Thus, while it may seem plausible to us that broiler chickens in their billions have lives that are not worth living, they might well be only barely so.

At the start of this paper, I asked whether the good experienced by human and non-human animals in aggregate suffices to counterbalance all the harms they suffer, such that the world is morally good on balance, or whether the moral weight of suffering is greater. In light of the evidence reviewed above, our answer could turn on AO. Given the sheer number of lives that might not be worth living that are added to the population year on year, if AO is true, the weight of suffering could very easily be greater. If AO is false, the optimist can more easily stand her ground. Assuming that we are willing to say that animals with lives like those reviewed above have lives that are at worst only just not worth living, we may then go on to assert that lives of that kind, no matter how abundant, cannot together outweigh the value inherent in at least the very best human lives.

[12] Specifically, they suggest that broiler breeders, caged egg-laying hens, veal calves, and all pigs except the relatively few raised in the shelter-pasture system have negative welfare scores. Broiler breeders are birds raised to produce broilers. Breeders live for one to two years, and are forced to live on a restricted diet, consuming only 25-35% of the feed they would prefer to consume. As of 2007, the US annually used around 62 million birds as broiler breeders, compared with a staggering 9 billion broilers.

7. Conclusion

In this paper, I have set out an argument for LTNU, which proceeds from a small handful of intuitive structural principles and two individually plausible additional premises to the conclusion that there exists some depth of lifetime suffering that cannot be counterbalanced by any measure of well-being experienced by others. As I've shown, there are significant costs to accepting this conclusion. We have little hope of avoiding these costs by questioning the argument's structural principles. We do not do much better in giving up CON-RRC, unless we are willing to accept RRC. By giving up AO, we are able to assert that there are some lives so good that their addition to the population can counterbalance the addition of any number of lives that are only barely not worth living. Independently, this conclusion could provide a way of avoiding the dreary conclusion that suffering in aggregate morally outweighs happiness among human and non-human animals, at least for now.

Bibliography

Blackorby, Charles, Walter Bossert, and David Donaldson (2005) *Population issues in social choice theory, welfare economics, and ethics*. Cambridge: Cambridge University Press.

Broad, William J. (2015) An ocean mystery in the trillions. *The New York Times* 29.06.2015 <https://www.nytimes.com/2015/06/30/science/bristlemouth-ocean-deep-sea-cyclothone.html> Accessed 28.02.2022.

Broome, John (2000) Incommensurable values. In Crisp and Hooker, eds. *Well-being and morality: essays in honour of James Griffin*, 21-38. Oxford: Oxford University Press.

Broome, John (2004) *Weighing lives.* Oxford: Oxford University Press.

Browning, Heather and Veit, Walter (2021) Positive wild animal welfare. PhilSci Archive: <http://philsci-archive.pitt.edu/19608/> Accessed 22.02.2022.

Carlson, Erik (1998) Mere addition and two trilemmas of population ethics. *Economics and Philosophy* 14, 283-306.

Carlson, Erik (2007) Higher values and non-Archimedean additivity. *Theoria* 73, 3-27.

Chang, Ruth (1997) Introduction. In Chang, ed. *Incommensurability, incomparability, and practical reason*, 1-34. Cambridge, MA: Harvard University Press.

Chang, Ruth (2016) Parity, imprecise comparability, and the Repugnant Conclusion. *Theoria* 82, 183-215.

Chang, Ruth (2022) How to avoid the Repugnant Conclusion. In McMahan, Campbell, Goodrich, and Ramakrishnan, eds. *Ethics and existence: the legacy of Derek Parfit*, 389-429. Oxford: Oxford University Press.

Crisp, Roger (2021) Would extinction be so bad? *The New Statesman* 10.08.2021 <https://www.newstatesman.com/ideas/agora/2021/08/would-extinction-be-so-bad> Accessed 22.11.2021.

Dostoevsky, Fyodor (1880/1994) *The Karamazov brothers*, transl. Avsey. Oxford: Oxford University Press.

Frick, Johann (2017) On the survival of humanity. *Canadian Journal of Philosophy* 47, 344-67.

Groff, Zach and Ng, Yew-Kwang (2019) Does suffering dominate enjoyment in the animal kingdom? An update to welfare biology. *Biology and Philosophy* 34, 40.

Gustafsson, Johan (2020) Population axiology and the possibility of a fourth category of absolute value. *Economics & Philosophy* 36, 81-110.

Gustafsson, Johan (2022) *Money pump arguments*. Cambridge: Cambridge University Press.

Hapgood, Fred (1974) *Why males exist: an inquiry into the evolution of sex*. New York, NY: William Morrow.

Hare, Caspar (2010) Take the sugar. *Analysis* 70, 237-47.

Herzog, Hal (2010) *Some we love, some we hate, some we eat: why it's so hard to think straight about animals.* New York, NY: HarperCollins.

Holldöbler, Bert and Wilson, Edward O. (1995) *Journey to the ants: a story of scientific exploration*. Cambridge, MA: Harvard University Press.

Hooker, Brad (2002) *Ideal code, real world: a rule-consequentialist theory of morality*. Oxford: Oxford University Press.

Horta, Oscar (2010) Debunking the idyllic view of natural processes: population dynamics and suffering in the wild. *Télos* 17, 73-88.

Huemer, Michael (2010) Lexical priority and the problem of risk. *Pacific Philosophical Quarterly* 91, 332-51.

Hurka, Thomas (1982) Value and population size. *Ethics* 93, 496-507.

James, William (1891) The moral philosopher and the moral life. *International Journal of Ethics* 1, 330-54.

Jensen, Karsten Klint (2008) Millian superiorities and the Repugnant Conclusion. *Utilitas* 20, 279-300.

Knowles, Toby, Steve Kestin, Susan Haslan, Steven Brown, Laura Green, Andrew Butterworth, Stuart Pope, Dirk Pfeiffer, and Christine Nicol (2008) Leg disorders in broiler chickens: prevalence, risk factors, and prevention. *PLoS ONE* 3: e1545.

Knutsson, Simon (2021) The world destruction argument. *Inquiry* 64, 1004-23.

Kosonen, Petra (2021) Discounting small probabilities solves the intrapersonal addition paradox. *Ethics* 132, 204-17.

Lazar, Seth and Lee-Stronach, Chad (2019) Axiological absolutism and risk. *Nôus* 53, 97-113.

Lee-Stronach, Chad (2021) Morality, uncertainty. *Philosophical Quarterly* 71, 334-58.

LeGuin, Ursula K. (1973/1991) The ones who walk away from Omelas. *Utopian Studies* 2, 1-5.

Mayerfeld, Jamie (1996) The moral asymmetry of happiness and suffering. *The Southern Journal of Philosophy* 34, 317-338.

Mayerfield, Jamie (1998) *Suffering and moral responsibility*. Oxford: Oxford University Press.

McMahan, Jefferson (1981) Problems of Population Theory. *Ethics* 92, 96-127.

Mulgan, Tim (2002) The Reverse Repugnant Conclusion. *Utilitas* 14, 360-4.

Nebel, Jacob (2022) Totalism without repugnance. In McMahan, Campbell, Goodrich, and Ramakrishnan, eds. *Ethics and existence: the legacy of Derek Parfit*, 200-231. Oxford: Oxford University Press.

Newberry, Toby (2021) How many lives does the future hold? *Global Priorities Institute Technical Report* No. T2-2021. <https://globalprioritiesinstitute.org/how-many-lives-does-the-future-hold-toby-newberry-future-of-humanity-institute-university-of-oxford/> Accessed 28.02.2022.

Ng, Yew-Kwang (1989) What should we do about future generations? Impossibility of Parfit's Theory X. *Economics & Philosophy* 5, 235-253.

Ng, Yew-Kwang (1995) Toward welfare biology: evolutionary economics of animal consciousness and suffering. *Philosophy and Biology* 10, 255-285.

Norwood, F. Bailey and Lusk, Jayson (2011) *Compassion by the pound: the economics of farm animal welfare*. Oxford: Oxford University Press.

Ord, Toby (2013) Why I'm not a negative utilitarian. <http://www.amirrorclear.net/academic/ideas/negative-utilitarianism/> Accessed 22.11.2021.

Parfit, Derek (1984) *Reasons and persons*. Oxford: Oxford University Press.

Parfit, Derek (2016) Can we avoid the Repugnant Conclusion? *Theoria* 82, 110-27.

Popper, Karl (1966/2011) *The open society and its enemies* Abingdon: Routledge.

Pressman, Michael (2015) A defence of average utilitarianism. *Utlitas* 27, 389-424.

Rachels, Stuart (2005) Counterexamples to the transitivity of better than. *Australasian Journal of Philosophy* 76, 71-83.

Robinson, Timothy, William Wint, Giulia Conchedda, Thomas Van Boeckel,,Valentina Ercoli,, Elisa Palamara, Giuseppina Cinardi, Laura D'Aietti, Simon Hay, and Marius Gilbert (2014) Mapping the global distribution of livestock. *PLoS One, 9*, e96084.

Scheffler, Samuel (2013) *Death and the afterlife*. Oxford: Oxford University Press.

Scheffler, Samuel (2018) *Why worry about future generations?* Oxford: Oxford University Press.

Schoenfield, Miriam (2014) Decision making in the face of parity. *Philosophical Perspectives* 28, 263-77.

Smart, R. N. (1958) Negative utilitarianism. *Mind* 67, 542-3.

Temkin, Larry S. (1996) A continuum argument for intransitivity. *Philosophy and Publica Affairs* 25, 175-210.

Temkin, Larry S. (2012) *Rethinking the good: moral ideals and the nature of practical reasoning*. Oxford: Oxford University Press.

Thomas, Teruji (2018) Some possibilities in population axiology. *Mind* 127, 807-32.

Thomas, Teruji (2022) Separability and population ethics. In Arrhenius, Bykvist, Campbell, and Finneron-Burns, eds. *The Oxford handbook of population ethics*, 271-95. Oxford: Oxford University Press.

Thomas, Teruji (ms) Reconstructing Arrhenius's Impossibility Theorems. <https://users.ox.ac.uk/~mert2060/webfiles/Reconstructing-Arrhenius-for-web.pdf> Last accessed 22.11.2021.

Tomasik, Brian (2015) The importance of wild animal suffering. *Relations: Beyond Anthropocentrism* 3, 133-52.