Ergo AN OPEN ACCESS JOURNAL OF PHILOSOPHY

Non-Additive Axiologies in Large Worlds

CHRISTIAN TARSNEY
Population Wellbeing Initiative, University of Texas at Austin

TERUJI THOMAS
Global Priorities Institute, Faculty of Philosophy, University of Oxford

Is the overall value of a world just the sum of values contributed by each value-bearing entity in that world? *Additively separable* axiologies (like total utilitarianism, prioritarianism, and critical level views) say 'yes', but non-additive axiologies (like average utilitarianism, rank-discounted utilitarianism, and variable value views) say 'no'. This distinction appears to be practically important: among other things, additive axiologies generally assign great importance to large changes in population size, and therefore tend to strongly prioritize the long-term survival of humanity over the interests of the present generation. Non-additive axiologies, on the other hand, need not assign great importance to large changes in population size. We show, however, that when there is a large enough 'background population' unaffected by our choices, a wide range of non-additive axiologies converge in their implications with additive axiologies—for instance, average utilitarianism converges with critical-level utilitarianism and various egalitarian theories converge with prioritarianism. We further argue that real-world background populations may be large enough to make these limit results practically significant. This means that arguments from the scale of potential future populations for the astronomical importance of avoiding existential catastrophe, and other arguments in practical ethics that seem to presuppose additive separability, may succeed in practice whether or not we accept additive separability as a basic axiological principle.

1. Introduction

Is the overall value of a possible world just the sum of values contributed by individual value-bearing entities in that world? This question represents a central dividing line in axiology, between axiologies that are *additively separable* (hereafter

Contact: Christian Tarsney <christian.tarsney@austin.utexas.edu>
Teruji Thomas <teruji.thomas@philosophy.ox.ac.uk>

https://doi.org/10.3998/ergo.5714

usually abbreviated 'additive') and those that are not. Additive axiologies allow the value of a world to be represented as a sum of values independently contributed by each value-bearing entity in that world, while non-additive axiologies do not. *Total utilitarianism*, for example, claims that the value of a world is simply the sum of the welfare of every welfare subject in that world, and is therefore additive. On the other hand, *average utilitarianism*, which identifies the value of a world with the *average* welfare of all welfare subjects, is non-additive.

As these examples suggest, we will assume the context of *welfarist population axiology*, meaning that we take the 'value bearers' to be the lives of welfare subjects, and assume that 'value' is a function of their welfare—although, unsurprisingly, our formal results will not depend on this interpretation.

Prima facie, the question of additive separability appears to carry considerable practical significance. In particular, according to any additive axiology, the value contributed to the world by all future people depends linearly on how many such people there will be. This means that additive axiologies are likely to assign very great importance to *existential catastrophes* (human extinction or other events that would seriously curtail humanity's future prospects), since these events will generally correspond to very large reductions in future population size (Bostrom 2003; 2013). On an additive axiology, the sheer number of people whose existence is at stake strongly suggests that we should be willing to pay very high costs (e.g., in terms of the welfare of the present generation) for the sake of avoiding existential catastrophe. In contrast, many non-additive axiologies—particularly average utilitarianism and various kindred views—are not sensitive in the same way to population size, and may therefore regard the question of humanity's long-term survival as having much more limited significance in comparison with the welfare of the present generation.

As a stylized illustration: suppose that there are 10^{10} existing people, all with welfare 1. We can either (O_1) leave things unchanged, (O_2) improve the welfare of all the existing people from 1 to 2, or (O_3) create some number n of new people with welfare 1.5. Total utilitarianism, of course, tells us to choose O_3, as long as n is sufficiently large. But average utilitarianism—while agreeing that O_3 is better than O_1 and that the larger n is, the better—nonetheless prefers O_2 to O_3 no matter how astronomically large n may be. Now, additive axiologies can disagree with total utilitarianism here if they claim that adding people with welfare 1.5 makes the world *worse* instead of better; but the broader point is that they will almost always claim that the difference in value between O_3 and O_1 becomes astronomically large (whether positive or negative) as n increases—bigger, for example, than the difference in value between O_2 and O_1. Non-additive axiologies, on the other hand, need not regard O_3 as making a big difference to the value of the world, regardless of n. Again, average utilitarianism agrees with total utilitarian-

ism that O_3 is an improvement over O_1, but regards it as a *smaller* improvement than O_2, even when it affects vastly more individuals.

Thus, additive separability seems to play a crucial role with respect to arguably the most important practical question in population ethics: the relative importance of (i) ensuring the long-term survival of our civilization and its ability to support a very large number of future individuals with lives worth living vs. (ii) improving the welfare of the present population.

In this paper, however, we show that under certain circumstances a wide range of non-additive axiologies 'converge' with additive ones: that is, they have the same practical implications as certain additive axiologies to which they correspond. This convergence between additive and non-additive axiologies has a number of interesting consequences, but perhaps the most important is that non-additive axiologies can inherit the linear sensitivity of their additive counterparts to changes in population size. This makes arguments for the overwhelming importance of avoiding existential catastrophe based on the potentially astronomical scale of the far future less reliant on the controversial assumption of additive separability. It thereby increases the robustness of the practical case for the overwhelming importance of avoiding existential catastrophe and diminishes the practical importance of additive separability as an abstract axiological principle.

Our starting place is the observation that, according to non-additive axiologies, which of two outcomes is better can depend on the welfare of the people unaffected by the choice between them. That is, suppose we are comparing two populations X and Y.[1] And suppose that, besides X and Y, there is some 'background population' Z that would exist either way. (Z might include, for instance, past human or non-human welfare subjects on Earth, faraway aliens, or present/future welfare subjects who are simply unaffected by our present choice.) Non-additive axiologies allow that whether X-and-Z is better than Y-and-Z can depend on facts about Z.[2]

With this in mind, our argument has two steps. First, we prove several results to the effect that, if the background population Z is sufficiently large, then non-additive axiologies converge with additive ones. For example, average utilitarianism converges with critical-level utilitarianism, and various egalitarian

1. We follow the tradition in population ethics that 'populations' are individuated not only by which people they contain, but also by what their welfare levels would be. (However, in the formalism introduced in Section 2, the populations we'll consider are *anonymous*, i.e., the identities of the people are not specified.)

2. The role of background populations in non-separable axiologies has received surprisingly little attention, but has not gone entirely unnoticed. In particular, Spears and Budolfson (2021) and Budolfson and Spears (2022) consider the implications of background populations for the theoretical importance of the 'Repugnant Conclusion' and for public policies affecting future population size, respectively. (We discuss the former issue in §8.1 below.) And, as we discovered while revising this paper, an argument very much in the spirit of our own (though without our formal results) was elegantly sketched several years ago in a blog post by Carl Shulman (2014).

theories converge with prioritarianism. Second, we argue that the background populations in real-world choice situations are large—at a minimum, orders of magnitude larger than the present and near-future human population, and plausibly orders of magnitude larger than the entire population of our future light cone. This provides some *prima facie* reason to believe that non-additive axiologies of the types we survey will agree closely with their additive counterparts in practice. More specifically, we argue that real-world background populations are large enough to substantially increase the importance that average utilitarianism assigns to avoiding existential catastrophe.

The paper proceeds as follows: Section 2 introduces some formal concepts and notation. Section 3 formally defines additive separability and describes some important classes of additive axiologies. Sections 4–5 survey several classes of non-additive axiologies and show that they become additive in the large-background-population limit. Section 6 argues that real-world background populations are large, and briefly considers what their welfare distributions might look like. Section 7 illustrates the implications of the preceding arguments by examining how realistic background populations affect the importance of avoiding existential catastrophe according to average utilitarianism. Section 8 considers (without endorsing) two ways in which our results might be taken as arguments against the non-additive views to which they apply. Section 9 is the conclusion.

2. Formal Setup

All of the axiologies we will consider evaluate worlds based only on the number of welfare subjects at each level of lifetime welfare. We will consider only worlds containing a finite *total* number of welfare subjects. We will also set aside worlds that contain *no* welfare subjects, simply because some population axiologies, like average utilitarianism, do not evaluate such empty worlds.

Thus, for formal purposes, a *population* is a function from the set $\mathcal{W}$ of all possible welfare levels to the set $\mathbb{Z}_+$ of all non-negative integers, specifying the number of welfare subjects at each level; we require it to be finitely supported, and not everywhere equal to zero.[3] Despite this formalism, we'll say that a welfare level w *occurs* in a population X if $X(w) \neq 0$. An *axiology* $\mathcal{A}$ is a strict partial order $\succ_{\mathcal{A}}$ on the set $\mathcal{P}$ of all populations, with '$X \succ_{\mathcal{A}} Y$' meaning that population X is better than population Y according to $\mathcal{A}$.[4]

3. We also use the standard notation of $\mathbb{R}$ for the set of real numbers, $\mathbb{R}_+$ for the set of non-negative real numbers, and $\mathbb{N}$ for the set of natural numbers (starting from 1).

4. A *strict partial order* is a transitive, irreflexive binary relation. We won't need the relation $\approx$ of equal goodness, but (following Fishburn 1970: 1.2) it is usually possible to recover $\approx$ from betterness: $X \approx Y$ if and only if, for all Z, $(Z \succ X \leftrightarrow Z \succ Y)$ and $(X \succ Z \leftrightarrow Y \succ Z)$.

Almost all the axiologies we will consider in this paper are defined in terms of a *value function* $V_{\succ} : \mathcal{P} \to \mathbb{R}$, which *represents* the axiology's ranking of worlds in the sense that $X \succ_{\mathcal{A}} Y$ if and only if $V_{\mathcal{A}}(X) > V_{\mathcal{A}}(Y)$.[5] When an axiology $\mathcal{A}$ is defined in this way, it is natural (though not obligatory) to think of $V_{\mathcal{A}}$ as encoding not only the 'ordinal' facts about which populations are better than which others, but also the 'cardinal' facts about *how much* better they are. We will state our results in both ordinal and cardinal terms. The cardinal facts may be especially important when evaluating populations in the face of uncertainty, an issue we will mainly set aside until Section 7.2.

To illustrate this formalism, the *size* of a population X, denoted $|X|$, is simply the total number of welfare subjects:

$$|X| := \sum_{w \in \mathcal{W}} X(w).$$

Similarly, the total welfare is

$$\mathrm{Tot}(X) := \sum_{w \in \mathcal{W}} X(w)w.$$

Of course, the definition of $\mathrm{Tot}(X)$ only makes sense on the assumption that we can add together welfare levels, and in this connection we generally assume that $\mathcal{W}$ is given to us as a set of real numbers. (In common terminology, we assume that welfare is 'measurable on a ratio scale'.) With that in mind, the average welfare

$$\overline{X} := \mathrm{Tot}(X)/|X|$$

is also well-defined.

3. Additivity

We can now give a precise definition of additive separability.

5. The use of a value function primarily rules out *incompleteness*, i.e., cases of two populations that are not equally good, but neither of which is better than the other. (See fn. 4 on equal goodness.) Allowing for some incompleteness is quite common. To keep things simple, we will not consider any incomplete axiologies. But it is often possible to represent an incomplete axiology by a *set* $\mathcal{V}_{\mathcal{A}}$ of value functions—in the sense that $X \succ_{\mathcal{A}} Y$ if and only if $V(X) > V(Y)$ for all $V \in \mathcal{V}_{\mathcal{A}}$—and then to apply our results one value function at a time. Another possible strategy is to argue that apparent cases of incompleteness are really cases of vagueness (Broome 1997); one can easily combine our discussion with, e.g., a supervaluationist or epistemicist account of vagueness.

If X and Y are populations, then let $X+Y$ be the population obtained by adding together the number of welfare subjects at each welfare level in X and Y. That is, for all $w \in \mathcal{W}$, $(X+Y)(w) = X(w)+Y(w)$. An axiology is *separable* if, for any populations X, Y, and Z,

$$X+Z \succ Y+Z \Leftrightarrow X \succ Y.$$

This means that in comparing $X+Z$ and $Y+Z$, one can ignore the shared sub-population Z. Separability is entailed by the following more concrete condition:

Additivity
An axiology $\mathcal{A}$ is *additively separable* (or *additive* for short) iff it can be represented by a value function of the form

$$V_{\mathcal{A}}(X) = \sum_{w \in \mathcal{W}} X(w) f(w)$$

with $f : \mathcal{W} \to \mathbb{R}$. Thus the value of X is given by transforming the welfare of each welfare subject by the function f and then adding up the results.

In the following discussion, we will sometimes want to focus on the distinction between additive and non-additive axiologies, and sometimes on the distinction between separable and non-separable axiologies. While an axiology can be separable but non-additive, none of the views that we focus on will have this feature. So for our purposes, the additive/non-additive and separable/non-separable distinctions are more or less extensionally equivalent.[6]

We will consider three categories of additive axiologies in this paper, which we now introduce in order of increasing generality. First, there is *total utilitarianism*, which identifies the value of a population with its total welfare.[7]

6. We say 'more or less' because we briefly consider one view, 'critical-level leximin', that is separable but non-additive according to our definitions, although it is additive in a more general sense—see Section 5.2 and appendix A.

For a detailed discussion of separability principles in population ethics, see Thomas (2022). The main difference between separability and additivity is that the latter, but not the former, entails completeness (see fn. 5) and the *Archimedean condition* (if $X \succ Y \succ Z$ then, for some integer $n > 0$, $nY + Z \succ X + nZ$). Failures of either one of these conditions can complicate, but don't necessarily block, arguments for the overwhelming importance of existential catastrophe based on the astronomical size of the potential far-future population.

7. Total utilitarianism is arguably endorsed (with varying degrees of clarity and explicitness) by classical utilitarians like Hutcheson (1725/1738), Bentham (1789), Mill (1863), and Sidgwick (1874/1907), and has more recently been defended by Hudson (1987), de Lazari-Radek and Singer (2014), and Gustafsson (2020), among others.

Total Utilitarianism (TU)

$$V_{\text{TU}}(X) = \text{Tot}(X) = \sum_{w \in \mathcal{W}} X(w)w = \bar{X}\,|X|.$$

An arguable drawback of TU is that it implies the so-called 'Repugnant Conclusion' (Parfit 1984), that for any two positive welfare levels $w_1 < w_2$, for any population in which everyone has welfare w_2, there is a better population in which everyone has welfare w_1. The desire to avoid the Repugnant Conclusion is one motivation for the next class of additive axiologies, *critical-level* theories.[8]

Critical-Level Utilitarianism (CL)

$$V_{\text{CL}}(X) = \sum_{w \in \mathcal{W}} X(w)(w - c) = \text{Tot}(X) - c\,|X| = (\bar{X} - c)\,|X|$$

for some constant $c \in \mathcal{W}$ (representing the 'critical level' of welfare above which adding an individual to the population constitutes an improvement), generally but not necessarily taken to be positive.

We sometimes write 'CL_c' rather than merely 'CL' to emphasize the dependence on the critical level. TU is a special case of CL, namely, the case with $c = 0$. But as long as c is positive, CL avoids the Repugnant Conclusion since adding lives with very low positive welfare makes things worse rather than better.[9]

Another arguable drawback of both TU and CL is that they give no priority to the less well off—that is, they assign the same marginal value to a given improvement in someone's welfare, regardless of how well off they were to begin with. We might intuit, however, that a one-unit improvement in the welfare of a very badly off individual has greater moral value than the same welfare improvement for someone who is already very well off. This intuition is captured by *prioritarian* theories.[10]

Prioritarianism (PR)

$$V_{\text{PR}}(X) = \sum_{w \in \mathcal{W}} X(w) f(w)$$

8. Critical-level views have been defended by Blackorby, Bossert, and Donaldson (1997; 2005), among others.

9. But a positive critical level also brings its own, arguably greater drawbacks—e.g., the Strong Sadistic Conclusion (Arrhenius 2000), to which we return in Section 8.1.

10. Versions of prioritarianism have been defended by Weirich (1983), Parfit (1997), Arneson (2000), and Adler (2009; 2011), among others. *Sufficientarianism*, which by our definition will count as a special case of prioritarianism, has been defended by Frankfurt (1987) and Crisp (2003), among others.

for some function $f : \mathcal{W} \to \mathbb{R}$ (the 'priority weighting' function) that is concave and strictly increasing.

CL_c is the special case of PR with $f(w) = w - c$, and TU is the special case with $f(w) = w$.[11] Note also that our definition of the prioritarian family of axiologies is very close to our definition of additive separability, just adding the conditions that f is concave and strictly increasing.

4. Averagist and Asymptotically Averagist Views

In this section and the next, we consider a variety of non-additive axiologies, and show that each one gives the same verdicts as some additive axiology when there is a large enough background population. In this sense, non-additive axiologies 'converge' with additive ones. In this section, we show that average utilitarianism and related views converge with CL, where the critical level is the average welfare of the background population. In the next section, we show that various non-additive egalitarian views converge with PR.

4.1. Convergence

First, though, let us make the notion of 'convergence' more precise. Informally, we say that one axiology, $\mathcal{A}$, converges with another, $\mathcal{A}'$, if the verdicts of $\mathcal{A}$ approximate the verdicts of $\mathcal{A}'$ to arbitrary precision, as the size of the background population increases. In spelling this out, we will restrict attention to background populations of a given *type*, for example, all those having a certain average level of welfare. Here is the basic formal definition.

Ordinal Convergence
Axiology $\mathcal{A}$ *converges ordinally* with $\mathcal{A}'$ relative to background populations of type T if and only if, for any populations X and Y, if Z is a sufficiently large population of type T, then

$$X + Z \succ_{\mathcal{A}'} Y + Z \Rightarrow X + Z \succ_{\mathcal{A}} Y + Z.$$

Of course, if $\mathcal{A}'$ is additive, the last implication is equivalent to

$$X \succ_{\mathcal{A}'} Y \Rightarrow X + Z \succ_{\mathcal{A}} Y + Z.$$

11. More generally, V_{PR} represents CL_c if f has the form $f(w) = a(w - c)$, with $a > 0$.

We can, in other words, compare $X+Z$ and $Y+Z$ with respect to $\mathcal{A}$ by comparing X and Y with respect to $\mathcal{A}'$—if we know that Z is a sufficiently large population of the right type.

Note two ways in which this notion of convergence is fairly weak. First, what it means for Z to be 'sufficiently large' can depend on X and Y. Second, the displayed implications need not be a biconditional; thus, when $\mathcal{A}'$ does not have a strict preference between $X+Z$ and $Y+Z$ (e.g., when it is indifferent between them), convergence with $\mathcal{A}'$ does not imply anything about how $\mathcal{A}$ ranks those two populations.[12] Because of this, every axiology converges with the trivial axiology according to which no population is better than any other. Of course, such a result is uninformative, and we are only interested in convergence with more discriminating axiologies. Specifically, we will only ever consider axiologies that satisfy the Pareto principle (which we discuss in Section 5.1).

Ordinal convergence is 'ordinal' because it only concerns the way in which the two axiologies rank populations. As we noted in Section 2, one could interpret the value function used to define an axiology as conveying 'cardinal' information about the relative values of different populations. There is, correspondingly, a different notion of convergence that we will call *cardinal convergence*. Specifically, if $V_{\mathcal{A}}$ is the value function for $\mathcal{A}$, then one could interpret ratios like

$$\frac{V_{\mathcal{A}}(X_1)-V_{\mathcal{A}}(Y_1)}{V_{\mathcal{A}}(X_2)-V_{\mathcal{A}}(Y_2)}$$

as measuring how much better X_1 is than Y_1, compared to how much better X_2 is than Y_2.[13] Cardinal convergence occurs when two axiologies agree about these cardinal facts to arbitrary precision as the background population becomes large.

Cardinal Convergence
Axiology $\mathcal{A}$ (with value function $V_{\mathcal{A}}$) *converges cardinally* with $\mathcal{A}'$ (with value function $V_{\mathcal{A}'}$) relative to background populations of type T if and only if, for any four populations X_1, Y_1, X_2, Y_2 and any margin of error $\epsilon > 0$, if Z is a sufficiently large population of type T, then

12. The basic reason for this asymmetry in our treatment of $\mathcal{A}$ and $\mathcal{A}'$ is that the verdicts of $\mathcal{A}$ depend on Z, while the verdicts of an additive axiology $\mathcal{A}'$ do not. Thus, we should think of the verdicts of $\mathcal{A}$ approaching those of $\mathcal{A}'$, with the latter held fixed. And when $\mathcal{A}'$ says that $X+Z$ and $Y+Z$ are equally good, $\mathcal{A}$ may approximate this verdict equally well by saying that $X+Z$ is slightly better than $Y+Z$ or the other way around.

13. In standard terminology, this is the information encoded by $\mathcal{V}_{\mathcal{A}}$ if it 'measures value on an interval scale'. It is possible, of course, that the value function encodes even more information (perhaps measuring value on a 'ratio scale') but we will make no use of that in this paper.

$$\frac{V_{\mathcal{A}}(X_1+Z)-V_{\mathcal{A}}(Y_1+Z)}{V_{\mathcal{A}}(X_2+Z)-V_{\mathcal{A}}(Y_2+Z)} \text{ is within } \epsilon \text{ of } \frac{V_{\mathcal{A}'}(X_1+Z)-V_{\mathcal{A}'}(Y_1+Z)}{V_{\mathcal{A}'}(X_2+Z)-V_{\mathcal{A}'}(Y_2+Z)}$$

(assuming the denominator $V_{\mathcal{A}'}(X_2+Z)-V_{\mathcal{A}'}(Y_2+Z)\neq 0$).

As we have defined it, cardinal convergence does not quite imply ordinal convergence. Thus our main results will assert both ordinal and cardinal convergence. And we will often just speak of 'convergence' to cover both kinds.

4.2. Average Utilitarianism

Average utilitarianism identifies the value of a population with its average welfare level.[14]

Average Utilitarianism (AU)

$$V_{\text{AU}}(X)=\overline{X}=\sum_{w\in\mathcal{W}}\frac{X(w)}{|X|}w.$$

Our first result describes the behavior of AU as the size of the background population tends to infinity.

Theorem 1. *Average utilitarianism converges ordinally and cardinally to CL_c, relative to background populations with average welfare c. In fact, for any populations X, Y, Z, if $\overline{Z}=c$ and*

$$|Z|>\frac{|X|V_{\text{CL}_c}(Y)-|Y|V_{\text{CL}_c}(X)}{V_{\text{CL}_c}(X)-V_{\text{CL}_c}(Y)}, \tag{1}$$

then $V_{\text{CL}_c}(X)>V_{\text{CL}_c}(Y)\Rightarrow V_{\text{AU}}(X+Z)>V_{\text{AU}}(Y+Z)$.

14. Average utilitarianism is often discussed but rarely endorsed. It has its defenders, however, including Hardin (1968), Harsanyi (1977), and Pressman (2015). Mill (1863) can also be read as an average utilitarian (see fn. 2 in Gustafsson 2022b), though the textual evidence for this reading is not entirely conclusive.

As with all evaluative or normative theories—but perhaps more so than most—average utilitarianism confronts a number of choice points that generate a minor combinatorial explosion of possible variants. Hurka (1982a; 1982b) identifies three such choice points which generate at least twelve different versions of averagism. The view we have labeled AU (which Hurka calls A1) strikes us as the most plausible, but our main line of argument could be applied to many other versions. Versions of averagism that only care about the *future* population do present us with a challenge, which we discuss in §6.1.2.

Proofs of all theorems are given in the appendix.

Discussion of the normative implications of this and other results is deferred to the second half of the paper (§§6–9).

4.3. 'Variable Value' Views

Some philosophers have sought an intermediate position between total and average utilitarianism, acknowledging that increasing the size of a population (without changing its average welfare) can count as an improvement, but holding that additional lives have *diminishing marginal value*. The most widely discussed version of this approach is the *variable value* view.[15] It is useful to distinguish two types of this view, the second more general than the first.

Variable Value I (VV1)

$$V_{\text{VV1}}(X) = \overline{X}g(|X|)$$

where $g : \mathbb{Z}_+ \to \mathbb{R}_+$ is a non-zero function that is weakly increasing, concave, and bounded above.

Recall that the total welfare of a population X is equal to $\overline{X}|X|$; roughly speaking, VV1 says that changes in the second factor, the size of X, are less important when X is already large. The next view also gives varying marginal importance to average welfare:

Variable Value II (VV2)

$$V_{\text{VV2}}(X) = f(\overline{X})g(|X|)$$

where $f : \mathbb{R} \to \mathbb{R}$ is differentiable and strictly increasing, and $g : \mathbb{N} \to \mathbb{R}_+$ is a non-zero function that is weakly increasing, concave, and bounded above.

Sloganistically, variable value views can be 'totalist for small populations' (where g may be nearly linear), but must become 'averagist for large populations' (as g approaches its upper bound). It is therefore not entirely surprising that, in the large-background-population limit, VV1 and VV2 display the same

15. These views were introduced by Hurka (1983). Variable Value I is also discussed by Ng (1989) under the name 'Theory X''.

behavior as AU, converging with a critical-level view with the critical level given by the average welfare of the background population.

Theorem 2. *Variable value views converge ordinally and cardinally to* CL_c*, relative to background populations with average welfare c.*

For the broad class of variable value views, we cannot give the sort of threshold for $|Z|$ that we gave for AU, above which the ranking of $X+Z$ and $Y+Z$ must agree with the ranking given by $\mathrm{CL}_{\overline{Z}}$. For instance, because g can be *any* non-zero function that is weakly increasing, concave, and bounded above, variable value views can remain in arbitrarily close agreement with totalism for arbitrarily large populations, so if TU prefers one population to another, there will always be *some* variable value theory that agrees. In the case of VV1, we can say that if *both* TU and AU prefer X to Y, then all VV1 views will as well (see Proposition 1 in appendix B), and so whenever TU and $\mathrm{CL}_{\overline{Z}}$ have the same strict preference between X and Y, the threshold given in Theorem 1 holds for VV1 as well. For VV2, we cannot even say this much.[16]

5. Non-Additive Egalitarian Views

A second category of non-additive axiologies are motivated by egalitarian considerations. Does adding an individual to a population, or increasing the welfare of an existing individual, increase or decrease equality? The answer depends on the welfare of other individuals in the population, so it is easy to see why concern with equality might motivate separability violations.

Egalitarian views have been widely discussed in the context of distributive justice for fixed populations, but relatively little has been said about egalitarianism in a variable-population context. We are therefore somewhat in the dark as to which egalitarian views are most plausible in that context. But we will consider a few possibilities that seem especially promising, trying to consider each fork of two major choice points for variable-population egalitarianism.

The most important choice point is between (i) 'two-factor'/'pluralistic' egalitarian views, which treat the value of a population as the sum of two (or more)

16. What we can say about VV2 is the following: when $\overline{X} > \overline{Y}$, $|X| \geq |Y|$, and $f(\overline{X}) \geq 0$, VV2 is guaranteed to prefer X to Y. Similarly, when $\overline{X} > \overline{Y}$, $|Y| \geq |X|$, and $f(\overline{Y}) \leq 0$, VV2 is guaranteed to prefer X to Y. (These claims depend only on the fact that f is strictly increasing and g is weakly increasing.) So in any case where the population preferred by $\mathrm{CL}_{\overline{Z}}$ is larger and has average welfare to which VV2 assigns a non-negative value, or the population dispreferred by $\mathrm{CL}_{\overline{Z}}$ is larger and has average welfare to which VV2 assigns a non-positive value, VV2 will agree with $\mathrm{CL}_{\overline{Z}}$ whenever AU does.

terms, one of which is a measure of inequality, and (ii) 'rank-discounting' views, which give less weight to the welfare of individuals who are better off relative to the rest of the population. These two categories of views are extensionally equivalent in the fixed-population context, but come apart in the variable-population context (Kowalczyk 2020: ch. 3).

5.1. Two-Factor Egalitarianism

Among two-factor egalitarian theories, there is another important choice point between 'totalist' and 'averagist' views.

Totalist Two-Factor Egalitarianism

$$V(X) = \mathrm{Tot}(X) - I(X)|X|$$

where I is some measure of inequality in X.

Averagist Two-Factor Egalitarianism

$$V(X) = \overline{X} - I(X)$$

where I is some measure of inequality in X.[17]

Here, in each case, the second term of the value function can be thought of as a penalty representing the badness of inequality. Such a penalty could have any number of forms, but for the purposes of illustration we stipulate that $I(X)$ depends only on the *distribution* of X, where this can be understood formally as the function $X/|X|: \mathcal{W} \to \mathbb{R}$ giving the proportion of the population in X having each welfare level. The *degree* of inequality is indeed plausibly a matter of the distribution in this sense, and the *badness* of inequality is then plausibly a function of the degree of inequality and the size of the population. The more substantial assumption is that the badness of inequality either scales linearly with the size of the population (for the totalist version of the view) or does not depend on population size (for the averagist version).

Now, we want to know what these theories do as $|Z| \to \infty$. In the last section, we had to hold one feature of Z constant as $|Z| \to \infty$, namely, $\overline{Z}$. Egalitarian theories, however, are potentially sensitive to the whole distribution of welfare

17. One could also imagine variable-value two-factor theories (and two-factor theories that incorporate critical levels, priority weighting, etc., into their value functions), but we will set these possibilities aside for simplicity.

levels in the population, and so to obtain limit results it is useful to hold fixed the whole distribution of welfare in the background population, that is, $D := Z/|Z|$.

We'll state the general result, explain some of the terminology it uses, and then give some examples.

Theorem 3. *Suppose V is a value function of the form $V(X) = \mathrm{Tot}(X) - I(X)|X|$, or else $V(X) = \overline{X} - I(X)$, where I is a differentiable function of the distribution of X. Then the axiology $\mathcal{A}$ represented by V converges ordinally and cardinally with an additive axiology, relative to background populations with any fixed distribution D; specifically, it converges with the additive axiology with weighting function given by*[18]

$$f(w) = \lim_{t \to 0^+} \frac{V(D + t1_w) - V(D)}{t}.$$

If the Pareto principle holds with respect to $\mathcal{A}$, then f is weakly increasing, and if Pigou-Dalton transfers are weak improvements, then f is concave.

A few points in the theorem require further explanation. Informally, the function I is *differentiable* if $I(X)$ varies smoothly with X; we will give the formal definition when it comes to the proof (see Remark 1 in the appendix), but at any rate all proposed measures of inequality that we're aware of are differentiable, including the two we discuss below. The *Pareto principle* holds that increasing anyone's welfare increases the value of the population. This principle clearly holds for prioritarian views (because the priority-weighting f is strictly increasing), but it need not in principle hold for egalitarian views: conceptually, increasing someone's wellbeing might contribute so much to inequality as to be on net a bad thing. Still, the Pareto principle is generally held to be a desideratum for egalitarian views. Finally, a *Pigou-Dalton transfer* is a total-preserving transfer of welfare from a better-off person to a worse-off person that keeps the first person better-off than the second. The condition that Pigou-Dalton transfers are at least weak improvements (they do not make things worse) is often understood as a minimal requirement for egalitarianism.

To illustrate Theorem 3, let's consider two more specific families of egalitarian axiologies that instantiate the schemata of totalist and averagist two-factor egalitarianism respectively.

For the first, we'll use a measure of inequality based on the *mean absolute difference* (MD) of welfare, defined for any population X as follows:

$$\mathrm{MD}(X) := \sum_{v,w \in \mathcal{W}} \frac{X(w)X(v)}{|X|^2} |w - v|.$$

18. Here $1_w \in \mathcal{P}$ is the population with a single welfare subject at level w, and we use the fact that value functions of the assumed form can be evaluated directly on any finitely supported, non-zero function $\mathcal{W} \to \mathbb{R}_+$, such as, in particular, D and $D + t1_w$.

$\mathrm{MD}(X)$ represents the average welfare inequality between any two individuals in X.$\mathrm{MD}(X)|X|$, which scales with population size, can be understood as taking each *individual's* average welfare inequality with the members of X (including herself), then summing across individuals. Consider, then, the following totalist two-factor view:

Mean Absolute Difference Total Egalitarianism (MDT)

$$V_{\mathrm{MDT}}(X) = \mathrm{Tot}(X) - \alpha \mathrm{MD}(X)|X|$$

where $\alpha \in (0, 1/2)$ is a constant that determines the relative importance of inequality.[19]

Second, consider the following averagist two-factor view, which identifies overall value with a quasi-arithmetic mean of welfare:[20]

Quasi-Arithmetic Average Egalitarianism (QAA)

$$V_{\mathrm{QAA}}(X) = \mathrm{QAM}(X) = g^{-1}\Big(\sum_{w \in \mathcal{W}} \frac{X(w)}{|X|} g(w)\Big)$$

for some strictly increasing, concave function $g : \mathcal{W} \to \mathbb{R}$.

Implicitly, the measure of inequality in QAA is $I(X) = \bar{X} - \mathrm{QAM}(X)$, which one can show is a positive function, weakly decreasing under Pigou-Dalton transfers. In the limiting case where g is linear, $\mathrm{QAM}(X) = \bar{X}$.

Theorem 4. MDT *converges ordinally and cardinally to* PR, *relative to background populations with a given distribution* D. *Specifically*, MDT_α *converges with* PR_f, *the prioritarian axiology whose weighting function is*

$$f(w) = w - 2\alpha\ \mathrm{MD}(w, D) + \alpha\ \mathrm{MD}(D).$$

Here $\mathrm{MD}(w, D) := \sum_{x \in \mathcal{W}} D(x)|x - w|$ *is the average distance between* w *and the welfare levels occurring in* D.

19. For $\alpha \geq 1/2$, equality would be so important that the Pareto principle would fail, i.e., it would no longer be true in general that increasing someone's welfare level increases the value of the population.

20. See Fleurbaey (2010) and McCarthy (2015: Theorem 1) for axiomatizations of this type of egalitarianism, at least in fixed-population cases where the totalist/averagist distinction is irrelevant.

Theorem 5. QAA *converges ordinally and cardinally to* PR*, relative to background populations with a given distribution* D*. Specifically,* QAA_g *converges with* PR_f*, the prioritarian axiology whose weighting function is*

$$f(w) = g(w) - g(\text{QAM}(D)).$$

5.2. Rank Discounting

Another family of population axiologies that is often taken to reflect egalitarian motivations is *rank-discounted utilitarianism* (RDU). The essential idea of rank-discounting is to give different weights to marginal changes in the welfare of different individuals, not based on their absolute welfare level (as prioritarianism does), but rather based on their welfare *rank* within the population. One potential motivation for RDU over two-factor views is that, because we are simply applying different positive weights to the marginal welfare of each individual, we clearly avoid any charge of 'leveling down': unlike on two-factor views, there is nothing even *pro tanto* good about reducing the welfare of a better-off individual—it is simply *less bad* than reducing the welfare of a worse-off individual.[21]

Versions of rank-discounted utilitarianism have been discussed and advocated under various names in both philosophy and economics, for example, by Asheim and Zuber (2014) and Buchak (2017). In these contexts, the RDU value function is generally taken to have the following form:

$$V(X) = \sum_{k=1}^{|X|} f(k)X_k \tag{2}$$

where X_k denotes the welfare of the kth worst off welfare subject in X, and $f : \mathbb{N} \to \mathbb{R}$ is a positive but weakly decreasing function.[22]

However, these discussions often assume a context of fixed population size, and there are different ways one might extend the formula when the size is not fixed.

21. It is important to remember, however, that two-factor views with an appropriately chosen I, like those we considered in the last section, can avoid *all-things-considered* leveling down: that is, while they may suggest that there is *something good* about making the best off worse off, they never claim that it would be an all-things-considered improvement.

22. To connect this to the standard notation in this paper, one can alternatively write

$$V(X) = \sum_{w \in \mathcal{W}} \Big(g\big(\sum_{v \leq w} X(v)\big) - g\big(\sum_{v < w} X(v)\big) \Big) w$$

for some weakly increasing, concave function $g : \mathbb{R} \to \mathbb{R}$ with $g(0) = 0$. The two presentations are equivalent if $g(k) = \Sigma_{i=1}^{k} f(k)$ or conversely $f(k) = g(k) - g(k-1)$.

We will consider the most obvious approach, simply taking equation (2) as a definition regardless of the size of X.[23] A view of this type, explicitly designed for a variable-population context, is set out in Asheim and Zuber (2014). Simplifying slightly to set aside features irrelevant for our purposes, their view is as follows:

Geometric Rank-Discounted Utilitarianism (GRD)

$$V_{\mathrm{GRD}}(X) = \sum_{k=1}^{|X|} \beta^k X_k$$

for some $\beta \in (0, 1)$.

Here, the rank-weighting function is $f(k) = \beta^k$. In general, since f is assumed to be weakly decreasing and positive, $f(k)$ must asymptotically approach some limit L as k increases. For GRD, $L = 0$. But a simpler situation arises when $L > 0$ (so that f is bounded away from zero):

Bounded Rank-Discounted Utilitarianism (BRD)

$$V_{\mathrm{BRD}}(X) = \sum_{k=1}^{|X|} f(k) X_k$$

for some weakly decreasing, positive function $f : \mathbb{R} \to \mathbb{R}$ that is eventually convex[24] with asymptote $L > 0$.

We will state formal results about both GRD and BRD in Appendix A; they involve a slightly more restricted notion of convergence than we have considered so far. The case of BRD is relatively simple: it converges with total utilitarianism. This is because, when the background population is very large, each life in the foreground population with welfare level w contributes approximately Lw to the overall value of the population (at least assuming that w is higher than some level in the background population). So the overall contribution of the foreground population is approximately equal to its total welfare times L.

When, as in GRD, the asymptote of the weighting function f is at $L = 0$, the situation is subtler and appears to depend on the exact rate at which f decays. We will consider only GRD, as it is the best-motivated example in the literature.

23. An alternative approach would be to extend to variable populations the 'veil of ignorance' motivation for rank-discounting described by Buchak (see also McCarthy, Mikkola, & Thomas 2020: Example 2.9). However, on the most obvious way of doing this, the resulting view is coextensive with a two-factor egalitarian view and so falls under the purview of Theorem 3 (even if it is conceptually different in important ways).

24. That is, there is some k such that f is convex on the interval (k, ∞). The assumption of eventual convexity is often satisfied, but is primarily a technical assumption to be used in Theorem 6 below.

Uniquely among the axiologies we consider, GRD does *not* converge with an additive, Paretian axiology on any interesting range of populations. Roughly speaking, this is because, as the background population gets larger, the weight given to the best-off individual in X becomes arbitrarily small relative to the weight given to the worst-off—smaller than the relative weight given by any particular additive, Paretian axiology.

Nonetheless, it turns out that GRD *does* converge with a *separable*, Paretian axiology, which we call critical-level leximin. This is an extreme form of prioritarianism in which infinite priority is always given to the less well-off. We'll explain this carefully in Appendix A, but perhaps the most important take-away is that (because critical-level leximin is so extreme) GRD leads to some very strange and counterintuitive results when the background population is sufficiently large.

For example, tiny benefits to worse-off individuals will often be preferred over astronomical benefits to even slightly better-off individuals; moreover, adding an individual to the population with anything less than the maximum welfare level in the background population will often make things worse overall.[25] In fact, GRD implies what we might call the 'Snobbish Conclusion':

> **Snobbish Conclusion**
> In some circumstances, given a very high welfare level w_1 just slightly below the best in the background population, and an even higher welfare level w_2 greater than any in the background population, adding even one life at w_1 makes things so much worse that it cannot be compensated by *any* number of lives at w_2.

This seems crazy to us. We could just about understand the Snobbish Conclusion in the context of an anti-natalist view, according to which adding lives *invariably* has negative value; but, according to GRD, there are many possible background populations (for instance, any in which the highest welfare level is less than w_1) to which the addition described above would constitute an improvement. We could also understand the view that adding good lives can make things worse if it lowers average welfare or increases inequality (e.g., as measured by mean absolute difference or standard deviation). But, again, that's not what's going on here. Instead, GRD implies that adding excellent lives makes things worse if

25. A toy example illustrates these phenomena, which are somewhat more general than the theorem entails. Suppose the background population consists of N people at level 100. Let X consist of two people at level 99; let Y consist of one person at level 98 and one at level 1000; and let Z consist of two people at level 99 and one at 99.9. We have $V_{\text{GRD}}(X) - V_{\text{GRD}}(Y) = \beta - \beta^2 - 900\beta^{N+2}$, which is positive if N is large enough, in which case $X \succ_{\text{GRD}} Y$, illustrating the first claim. On the other hand, $V_{\text{GRD}}(X) - V_{\text{GRD}}(Z) = 0.1\beta^3 - 100\beta^{N+3}$, again positive for N large enough; then $X \succ_{\text{GRD}} Z$, illustrating the second claim.

the number of even slightly better lives already in existence happens to be sufficiently great, regardless of the other facts about the distribution. In some cases, it makes things so much worse that it cannot be compensated by adding any number of even better lives.

To sum up, many forms of egalitarianism, including many forms of rank-discounted utilitarianism, converge with interesting additive axiologies. Geometric Rank-Discounted Utilitarianism provides one counterexample, although it does converge with an interesting *separable* axiology. Moreover, our general methodology of thinking about large background populations draws out some features that make GRD seem especially implausible.

6. Real-World Background Populations

In the rest of the paper, we explore the implications of the preceding results, and especially their practical implications for morally significant real-world choices. To apply our limit results in this way, there are two basic things one would like to know, which we investigate in this section.

First, one would like to know that the real-world background population is large enough that non-additive axiologies of the types we investigate give the same verdicts as the additive axiologies with which they converge. This is the topic of Section 6.1. The background population will be large if there are many welfare subjects whose lives are unaffected by our choices (although it may also be larger than the number of unaffected welfare subjects, as we explain in a moment). Many readers will already grant that the background population is extremely large, given the enormous number of welfare subjects in Earth's past, to say nothing of life elsewhere. However, we think it is nonetheless useful to develop some numerical estimates. After all, what counts as 'large enough' in the mathematical sense required to apply the limit results will depend on the specific axiology and the choice situation in question; being enormous by ordinary standards need not suffice. Indeed, while the background population will obviously be large compared to the foreground population in many ordinary or toy cases, this is much less obvious in other cases, where (for example) the future of life on Earth is at stake. The estimates developed in this section will allow us, in Section 7, to reach firmer conclusions about a stylized but basically realistic case of that type. Moreover, as we'll explain in Sections 6.1.1 and 6.1.2, there are some subtle ways in which advocates of non-additive axiologies might try to limit the size of the background population. (Mightn't non-human animals count less than humans? Shouldn't we simply set aside the past?) To evaluate these moves, it will again be useful to have some actual numerical estimates and the justifications for them in mind.

Second, we have seen that *which* additive axiology is relevant depends on the average welfare of the background population, and perhaps on its entire welfare distribution. Thus one would like to know something about this distribution. Here, unfortunately, it is very difficult to go beyond speculation, but we will still make some tentative remarks that will guide our discussion, as well as, we hope, providing a starting point for future research.

We have so far been informal about the distinction between 'background' and 'foreground' populations, but it will now be helpful to make these notions more precise. If we are interested in evaluating populations $X_1, X_2, \ldots X_n$, the population Z that can be treated as background is defined by $Z(w) = \min_i X_i(w)$. That is, the background population consists of the minimum feasible number of welfare subjects at each welfare level. For this Z and for each X_i, there is then a population X_i^* such that $X_i = X_i^* + Z$. A choice between $X_1, X_2, \ldots X_n$ can therefore be understood as a choice between the foreground populations $X_1^*, X_2^*, \ldots, X_n^*$, in the presence of background population Z.

As we noted above, welfare subjects will contribute to the size of the background population if they are unaffected by the choice at hand. However, it is important to realize that the size of the background population can exceed the number of unaffected individuals. This is because the background population depends on the *number* of welfare subjects guaranteed to exist at each level, not on their identities. As a result, for instance, future welfare subjects might contribute to the background population even if their identities are entirely dependent on our present choices (as argued by Parfit 1984: ch. 16, among others).

Having said that, in this section we will focus mainly on welfare subjects whose lives are entirely outside of our causal future, and thus would count as background for *any* choice we could realistically face. We will return to the possibility of affectable individuals contributing to the background population at the end of Section 6.1.2.

6.1. Population Size

We will make two claims about the size of the background populations that are relevant to real-world choices, with different degrees of confidence.

First, with high confidence, these populations are much larger (at least multiple orders of magnitude) than the present human population. Concretely, while there are fewer than 10^{10} humans alive today, we conservatively estimate that there have been at least 10^{17} welfare subjects in Earth's past, with estimates of 10^{20} or more being plausible. Informally, this suggests that our limit results should at least be relevant when comparing options that only affect present and near-future humans (though a background population of this size can also

substantially affect the evaluation of choices affecting the far future, as we will see in §7).

Second, with much lower confidence, real-world background populations may well be much larger (again, by multiple orders of magnitude) than the entire population in our future light cone. If this is right then our limit results are likely to be relevant to essentially any real-world choice.

Let's start by establishing the first claim, which only requires us to consider past welfare subjects on Earth. Estimates of the number of human beings who have ever lived are on the order of 10^{11} (Kaneda & Haub 2018), already an order of magnitude larger than the present human population. But past welfare subjects include a vast number of non-human animals, and especially wild animals over many millions of years. There are today at least 10^{11} wild mammals; for vertebrates in general, the number is far higher, with a conservative lower bound of 10^{13} (dominated by fish).[26] Prehistoric wild animal populations were presumably similarly large or larger, given the significant decline in wild animal populations as a result of human encroachment.[27] Inferring the total number of animals from the number alive at a given time requires assumptions about mortality rates. We will use a very conservative estimate of 0.1 deaths per individual per year in wild animal populations (roughly corresponding to a life expectancy of 10 years). The actual rates are almost certainly much higher for most species (especially given high infant mortality), implying larger total past populations. Being extremely conservative, then, we find that there have been at least 6.6×10^{17} mammals since the extinction of the dinosaurs 66 million years ago.[28] This gives our basic lower bound for the size of the background population. If we less conservatively allow that all vertebrates are welfare subjects, then a similar calculation gives a lower bound of 5×10^{20} individuals over the last 500 million years. And of course some invertebrates may be welfare subjects too.

While these background populations are large compared to the present population, they may not be large compared to the entire affectable future population. If our civilization survives for a very long time, the number of future individuals might be truly astronomical, and actions that affect the long-term future (for instance by causing or preventing existential catastrophes) might affect this entire population. If we can sustain just the size of the present human population until the Earth becomes uninhabitable, this would yield a future population

26. For useful surveys of evidence on present animal population sizes, see Tomasik (2019) and Bar-On, Phillips, and Milo (2018) (especially 61–4 and Table S1 in the supplementary appendix).

27. For instance, Smil (2013: 228) estimates that wild mammalian biomass has declined by 50% in the period 1900–2000 alone.

28. In detail, this is our low estimate of 10^{11} for the number of mammals at any given time, multiplied by our low estimate of 0.1 for the mortality rate, to obtain a low estimate for the number of mammals that died in any given year; and then multiplied by the number of years we are considering (here 6.6×10^7).

size on the order of 10^{17}, even ignoring non-humans.[29] This is roughly on a par with our lower-bound estimate of the number of past animals on Earth (though still much smaller than the more generous estimate that includes all vertebrates). Less conservatively, if humanity someday settles the cosmos and creates digital minds on a mass scale, far larger populations become possible—for instance, Bostrom (2013) estimates that such an interstellar civilization could support 10^{54} subjective life-years of human-like experience, perhaps in the form of 10^{52} lives with a subjective duration of 100 years each. Any plausible estimate of past animals on Earth will pale in comparison with these numbers.

There may nevertheless be unaffectable background populations even larger than these astronomical potential future populations. The crucial point is that the universe as a whole appears to be at least 100 times larger, and perhaps vastly larger, than the accessible universe (the portion of the universe that it is possible in principle for us to reach).[30] So, if life arises independently in many places, we would expect at least 99% of it to be outside the accessible universe and thus necessarily part of the background population. Similarly, if the universe contains many spacefaring civilizations, at least 99% of them should be inaccessible. However large the population of future human-originating civilization, this background population (consisting of many similar civilizations) will be orders of magnitude larger. But, of course, the hypothesis of extraterrestrial civilizations is entirely speculative, and deserves significantly less confidence than our lower-bound estimate of 6.6×10^{17} for the size of the background population.[31] We next consider two common objections to this lower-bound estimate.

29. This assumes 10^{10} individuals alive at a time, living for a century each, for the next billion years.

30. According to our best present understanding, the accessible universe contains about 20 billion galaxies, while the observable universe (the portion of the universe from which light has had time to reach us since the Big Bang) contains approximately 400 billion (Ord 2021). Moreover, our failure to detect positive curvature in the observable universe indicates that the universe as a whole must be *at least* 7.7 times larger than the observable universe (Vardanyan, Trotta, & Silk 2011), or 154 times larger than the accessible universe. Indeed, there is no known upper bound on the size of the universe as a whole, even assuming that it is finite. Greene notes that in many inflationary models, the universe is so large that '[i]f the entire cosmos were scaled down to the size of earth, the part accessible to us would be much smaller than a grain of sand' (2004: 285).

31. The preceding discussion sets aside the very real possibility that the universe is *infinite*, in such a way as to contain infinitely many civilizations and welfare subjects outside our future light cone (Knobe, Olum, & Vilenkin 2006; Vardanyan, Trotta, & Silk 2009). Intuitively, this possibility seems to bolster the practical import of our results, since if we can only affect a finite part of an infinite universe, then we are actually 'in' the limit case of an infinitely large background population, and not merely 'tending toward it'. That is, it's natural to think that in this case the non-additive views covered by our results should agree *exactly* with their additive limit theories. (In this way, our limit results also suggest a way of partially extending the non-additive views we consider to the context of infinite populations—specifically, that they should compare infinite populations that differ only finitely by applying the appropriate additive limit theory to the finite foreground populations.) However, the infinite context raises further complications that we don't have space to consider—in particular, how to define the *average welfare* and *welfare distribution* of an infinite background population.

6.1.1. Counting Some for Less than One?

Our basic estimate involves a large number of small and relatively simple animals. Several readers have suggested that, although such simple animals are still welfare subjects, perhaps they should receive less weight when we calculate the 'size' of a population for axiological purposes: perhaps, when evaluating outcomes, a typical mouse should effectively count as only (say) one fiftieth of a welfare subject, given its cognitive simplicity. In fact, a view along these lines is developed by Kagan (2019: see especially §4.5.). This way of accounting could, in principle, dramatically reduce the size of the background population.[32]

We have three responses to this suggestion. First, of course, one might lodge straightforward ethical objections to assigning different weights to different animals, since it seems to contradict the ideals of impartiality and equal consideration that are often seen as central to ethics in general and axiology in particular. Second, it seems that any plausible way of assigning weights is likely to leave a background population several orders of magnitude larger than the present human population. Let us take mice as representative of the background mammalian population. Adopting Kagan's suggestion of an axiological weight of 1/50 for mice (2019: 109) would only lower our estimate of the background population to ~10^{16}. Alternatively, it might be plausible to adopt weights that are proportional to cortical neuron count or lifespan.[33] But even weighting by both cortical neuron count *and* lifespan would only cut our lower-bound estimate of the size of the background population down to ~10^{13}, three orders of magnitude larger than the present human population. And this, of course, still only counts mammals since the extinction of the dinosaurs, ignoring all other animals.

Perhaps, however, there is some other rationale on which one would assign even tinier weights to practically all non-human animals. This brings us to our third response: even if we entirely ignore non-humans we may still find that background populations are large relative to foreground populations in most

32. Thanks to Tomi Francis and Toby Ord, who each separately suggested this objection. To accommodate axiological weights in our formal framework, we can allow populations to be *real*-valued functions on the set of welfare levels. For example, a population X consisting of one mouse at welfare level 10 would have $X(10) = 1/50$, using the weight mentioned in the text.

33. Weighting by lifespan seems particularly natural if we think that our ultimate objects of moral concern are *stages*, rather than complete, temporally extended individuals. Weighting by brain size or neuron count may seem natural if we believe that, in some sense, morally significant properties like sentience 'scale with' these measures of size. To arrive at the estimates in the text, we use the fact that humans have roughly 2875 times as many cortical neurons as mice (Roth & Dicke 2005: 251), and we generously assume a lifespan of 100 years for present humans. Note that weighting by lifespan means that our estimates no longer rely on our earlier assumption about mortality rates in past animal populations.

present-day choice situations. Past humans outnumber present humans by more than an order of magnitude, as we saw above. And, as we'll argue at the end of §6.1.2, it seems plausible that the large majority even of the present and near-future human population is approximately background in most choice situations.

6.1.2. A Causal Domain Restriction?

Here is another way in which proponents of non-additive axiologies might limit the size of the background population, at least for practical purposes. They could claim that, when it comes to decision making, we should apply our axiology to the population of welfare subjects who exist in our causal future (presumably, our future light cone), rather than to the universe as a whole. Such a *causal domain restriction* (Bostrom 2011) would simply exclude the kinds of large background populations we have considered so far. It could also be seen as a somewhat principled way for proponents of non-additive views to explain the common intuition that facts about the welfare levels of ancient humans simply can't be practically relevant today; or to mitigate the difficulty of applying non-additive views given our deep uncertainties about life outside the accessible universe.

We have three replies to this suggestion. First, to adopt a causal domain restriction is to abandon a central and deeply appealing feature of consequentialism, namely, the idea that we have reason *to make the world a better place*, from an impartial and universal point of view. That some act would make the world a better place, *full stop*, is a straightforward and compelling reason to do it. It is much harder to explain why the fact that an act would make *your future light cone* a better place (e.g., by maximizing the average welfare of its population), while making the world as a whole worse, should count in its favor.[34]

Second, the combination of a causal domain restriction with a non-separable axiology can generate counterintuitive inconsistencies between agents (and agent-stages) located at different times and places, with resulting inefficiencies. As a simple example, suppose that A and B are both agents who evaluate their options using causal-domain-restricted average utilitarianism. At t_1, A must choose between a population of one individual with welfare 0 who will live from t_1 to t_2 (population X) or a population of one individual with welfare –1 who will live from t_2 to t_3 (population Y). At t_2, B must choose between a population of three individuals with welfare 5 (population Z) or a population of one individual with welfare 6 (population W), both of which will live from t_2 to t_3. If A chooses X, then B will choose W (yielding an average welfare of 6 in B's future light cone), but if A chooses Y, then B will choose Z (since $Y + Z$ yields average welfare 3.5

34. This point goes back to Broad (1914); see Carlson (1995) for a detailed discussion of this area.

in B's future light cone, while $Y+W$ yields only 2.5). Since A prefers $Y+Z$ to $X+W$ (which yield averages of 3.5 and 3 respectively in A's future light cone), A will choose Y. Thus we get $Y+Z$, even though $X+Z$ would have been better from both A's and B's perspectives. That two agents who accept exactly the same normative theory and have exactly the same, perfect information can find themselves in such pointless squabbles is surely an unwelcome feature of that normative theory, though we leave it to the reader to decide just how unwelcome.[35]

Third, a causal domain restriction might not be enough to avoid the limit behaviors described in §§4–5, if there are large populations inside our future light cones that are background (at least, to a good approximation) with respect to most real-world choice situations. For instance, it seems likely that most choices we face will have little effect on wild animal populations over the next 100 years. More precisely, our choices might determine the *identities* of wild animals born in the next century (in the standard ways in which our choices are generally supposed to be identity-affecting with respect to most of the future population), while having little if any effect on the *number* of individuals at each welfare level in that population. And this alone would supply quite a large background population—conservatively, 10^{12} mammals and 10^{14} vertebrates. Indeed, it is plausible that with respect to most choices (even comparatively major, impactful choices), the vast majority of the present and near-future *human* population can be treated as background. For instance, if we are choosing between spending $1 million on anti-malarial bednets or on efforts to mitigate long-term existential risks, even the intervention that more directly impacts the near future (bednets) may have only a comparatively tiny effect on the number of individuals at each welfare level in the present- and near-future human population, so that most of that population can be treated as background.[36]

6.2. The Distribution of Welfare

What about the distribution of welfare in the background population? Anything we say about this will of course be enormously speculative. However, since it is—according to non-additive views!—an important topic, it seems worth making a few brief remarks.

35. This argument is essentially due to Rabinowicz (1989); see also the cases of intertemporal conflict for future-biased average utilitarianism in Hurka (1982b: 118–19).

Of course, cases like these also create potential time-inconsistencies for individual agents, as well as conflicts between multiple agents. But these inconsistencies might be avoidable by standard tools of diachronic rationality like 'resolute choice'.

36. For further discussion of, and objections to, causal domain restrictions in the context of infinite ethics, see Bostrom (2011) and Arntzenius (2014).

With respect to average welfare in the background population, two hypotheses seem particularly plausible.

Hypothesis 1
The background population consists mainly of small animals (whether terrestrial or extraterrestrial). Most of these animals have short natural lifespans, with high rates of infant mortality, so the average welfare level of the background population is likely close to zero. If the capacity for welfare scales with brain size or something similar, this would reinforce the same conclusion. Moreover, it seems plausible that average welfare in these populations will be negative, at least on a hedonic view of welfare (Horta 2010; Ng 1995). These assumptions imply, for instance, that AU, VV1 and VV2 converge with a version of CL with a slightly negative critical level.

Hypothesis 2
The background population consists mainly of the members of advanced alien civilizations, with astronomically large population sizes driven by space settlement or other technological advances. Under this hypothesis, given the limits of our present knowledge, all bets are off: average welfare in the background population could be very high (Ord 2020: 235–39), very low (Sotala & Gloor 2017), or anything in between.

With respect to the distribution of welfare more generally, we have even less to say. There is clearly a wide range of welfare levels in the background population, leading to significant inequality within specific groups.[37] However, it could still turn out that the background population as a whole is dominated by welfare subjects who lead fairly similar lives—for example, by small animals who almost always experience lifetime welfare close to 0, or by members of a highly egalitarian alien civilization. This would lead to a low level of inequality, at least by standard measures.

7. The Importance of Existential Catastrophe

If, as we have just argued, real-world background populations are indeed large relative to foreground populations, this provides some *prima facie* reason to believe that our limit results are practically significant: many plausible non-

37. For example, there is significant welfare inequality among contemporary humans (and so, presumably, among humans in the recent past), as indicated by self-reports (Helliwell, Layard, & Sachs 2019: ch. 2). Some literature on farm animal welfare also suggests significant inter-species welfare inequalities (e.g., Norwood & Lusk 2011: 224–29; Browning 2020).

additive views will agree closely with their additive counterparts. So, even if we don't accept additivity as a fundamental axiological principle, it may nevertheless be a useful heuristic for real-world decision-making purposes, and arguments in practical ethics that rely on separability assumptions may still succeed in practice.

In this section we give a concrete illustration of this point. As we suggested in §1, perhaps the most important practical question at stake in debates over additive separability is the relative importance of

(i) ensuring the existence of a large future population; versus
(ii) improving the welfare of the present generation.

For example, what sacrifice by the present generation would be worth it to forestall an 'existential catastrophe' that drastically reduces future population sizes?[38] On additive views, the amount of present welfare we should be willing to sacrifice to ensure the existence of a future population F scales linearly with $|F|$.[39] Thus, insofar as future populations would be astronomically larger than the present human population, it would be worth very large sacrifices on the part of the present generation to ensure their existence. But non-additive views need not endorse this sort of reasoning—in particular, AU and other similar views do not.

We therefore present a deeper analysis of how real-world background populations affect the relative importance of these two objectives according to AU. We focus on AU to keep the discussion manageable, and because AU exhibits the central relevant feature of insensitivity to population size, without the essentially orthogonal feature of inequality aversion.[40] Moreover, we will assume that the future generations that would exist if we avoid existential catastrophe would have higher-than-average welfare; in this case, AU assigns positive value to avoiding existential catastrophe. But most of what we say also applies, *mutatis mutandis*, to the *dis*value of avoiding existential catastrophe on the opposite assumption that the potential future population would have lower-than-average welfare.

38. An existential catastrophe, in our sense, might involve human extinction, but it might not. Our usage is slightly different from the one common in the philosophical literature, according to which an existential catastrophe is roughly 'any event that would permanently curtail humanity's long-term potential for value' (see, for instance, Bostrom 2013: 15; Ord 2020: 37).

39. Here we assume that the members of F would have high enough welfare that F contributes positive value to the world—*some* sacrifice would be worthwhile!—and we keep the distribution of welfare within F fixed while we scale up its size.

40. For example, while totalist two-factor egalitarianism is not additive, it is relatively clear that it can give great value to avoiding existential catastrophe, since the value of a population scales with its size.

Given Theorem 1, our basic conclusion will be unsurprising: if the background population is indeed as large as we have suggested in Section 6, then even AU gives great importance to existential catastrophe. However, there are a number of subtleties that we think are worth drawing out. In particular, we will take into account two points that complicate the application of our theorems. First, we can at best hope to affect the *probability* of existential catastrophe—a topic our theorems say nothing about. And, second, our more conservative estimates of the background population suggest that it may be much *smaller* than the size of the affected future population, making it less clear that AU will give verdicts similar to the additive axiology $\mathrm{CL}_{\bar{Z}}$.

The results of our analysis are summarized in Section 7.5.

7.1. Making the Question Precise: Existential Risk

We almost never face a choice between a certainty of catastrophe and a certainty of non-catastrophe. So, we suggest, the best way to understand the tradeoff between (i) and (ii) is in terms of the sacrifice the current generation might make in order to reduce existential *risk*, that is, the *probability* of existential catastrophe.[41]

To make this precise, let Z denote a background population that includes, as usual, any past welfare subjects, as well as any present or future ones who will be unaffected by the choice. Let F denote the future population that will exist only if we avoid existential catastrophe. Let C denote the current generation; more specifically, let C_w be a version of the current generation with average welfare w (and fixed size $|C|$). The following risky prospect then represents a p probability of existential catastrophe:

$$P: C_w + Z \text{ with probability } p;\ F + C_w + Z \text{ otherwise.}$$

From this baseline we can consider reducing the probability p of catastrophe by an infinitesimal amount δp while also decreasing the average welfare w of the current generation by δw to obtain a new prospect

$$P': C_{w-\delta w} + Z \text{ with probability } p - \delta p;\ F + C_{w-\delta w} + Z \text{ otherwise.}$$

41. To avoid grappling with probabilities, one could consider a straight choice between an outcome where existential catastrophe happens (in the notation about to be introduced, $C_w + Z$) and one in which the current generation sacrifices δw in average welfare to prevent it ($F + C_{w-\delta w} + Z$): how big would δw have to be to make these outcomes equally good? The results of this analysis would be qualitatively very similar to those below, but less relevant to the sorts of choices we actually face.

Suppose we do this in such a way that P and P' are equally good prospects. Then the ratio $\delta w / \delta p$ is an 'exchange rate' telling us how to weigh small changes in the welfare w of the current generation against small changes in the probability p of catastrophe. The higher the exchange rate, the greater the sacrifice that would be compensated by a marginal reduction in risk. So, formally, our question becomes:

> **Question 1.** How important is existential catastrophe as measured by the exchange rate $\delta w / \delta p$? In particular, how does it depend on the relative sizes of C, F, and Z?

Unfortunately, one cannot read the answer to Question 1 directly off of our limit results, which, after all, say nothing about probabilities. The plan for the rest of the section is to explain why our limit results are nonetheless relevant, and then to give a concrete analysis using the estimates for population sizes that we developed in Section 6.

7.2. Expected Value and Cardinal Convergence

To address Question 1, we must adopt some rule for evaluating risky prospects like P. When—as in all our examples—an axiology is defined using a value function, the most obvious rule is to rank prospects by their *expected* value. We will assume that this is the appropriate rule for both AU and for its limiting axiology $\mathrm{CL}_{\overline{Z}}$. Let us call the extended theories $\mathbb{E}$AU and $\mathbb{E}\mathrm{CL}_{\overline{Z}}$, respectively, where the $\mathbb{E}$ stands for *expected*.[42]

What justifies this assumption? When it comes to critical level views, there are foundational arguments supporting the use of expected value (see, e.g., Blackorby et al. 2005). For AU we have less to go on, but maximizing expected average welfare seems like a natural default, and there is no alternative for AU (or for other non-additive axiologies) that has achieved anything like widespread acceptance. Moreover, the use of expected value is closely connected to the idea that the value function V_{AU} represents cardinal facts about value. If $V_{\mathrm{AU}}(X) - V_{\mathrm{AU}}(Y)$ is a measure of how much better it is to get X instead of Y, we

42. Our conclusions will, nonetheless, be somewhat robust with respect to variations on $\mathbb{E}$AU. For example, McCarthy et al. (2020: Example 3.11) argue that the best way to extend AU to handle uncertainty is to evaluate each prospect by its expected total welfare divided by its expected population size. This amounts to applying V_{AU} directly to the 'population' that has the expected number of people at each welfare level. Although this view can behave quite differently from $\mathbb{E}$AU in general, the main qualitative conclusions described below still hold: rough independence from population size in Case 1, dependence on $|Z|/|C|$ in Case 2, and dependence on $|F|/|C|$ in Case 3.

should expect $p \times (V_{\text{AU}}(X) - V_{\text{AU}}(Y))$ to be at least a rough measure of how much better it is to get X instead of Y with probability p. The use of expected value is a systematic development of this idea.

This connection explains why our results about *cardinal* convergence are, after all, relevant to Question 1. We can interpret P and P' as involving three scenarios: (1) with probability $p - \delta p$, the catastrophe happens no matter which option is chosen; (2) with probability δp, choosing P' would successfully prevent the catastrophe; (3) with probability $1 - p$, the catastrophe fails to happen no matter what. In each of these scenarios, P and P' lead to different outcomes, and expected value theory effectively tells us to weigh up how much better or worse the outcome of P would be than the outcome of P' in each scenario, using the probabilities as weights. When the background population is extremely large, Theorem 1 says that V_{AU} and $V_{\text{CL}_{\bar{Z}}}$ will agree to high precision about the relative sizes of these differences in value. They will thus tend to agree about which option has higher expected value. And in particular, they will answer Question 1 in approximately the same way.

Here is the general theoretical lesson, stated somewhat informally. Suppose that axiology $\mathcal{A}$ converges *cardinally* with $\mathcal{A}'$ relative to background populations of type T. For any two prospects, each involving finitely many possible outcomes, if there is certain to be a sufficiently large background population of type T, then $\mathcal{A}$ and $\mathcal{A}'$ will agree about which prospect has higher expected value.

7.3. Analysis

With this set-up in hand, we now compare the answers to Question 1 given by $\mathbb{E}\text{AU}$ and by $\mathbb{E}\text{CL}_{\bar{Z}}$. We will give a general qualitative analysis of how the answers depend on the population sizes, illustrated numerically using some of our estimates from Section 6.

Let's first consider the case of $\text{CL}_{\bar{Z}}$. Again, we will assume that the prospect P is evaluated using its expected value $\mathbb{E}V_{\text{CL}_{\bar{Z}}}(P)$. By definition,

$$\mathbb{E}V_{\text{CL}_{\bar{Z}}}(P) = pV_{\text{CL}_{\bar{Z}}}(C_w + Z) + (1-p)V_{\text{CL}_{\bar{Z}}}(F + C_w + Z).$$

We can think of this expected value as a function of p and w, while holding all other parameters fixed. Then the exchange rate is given by a ratio of derivatives:

$$\frac{\delta w}{\delta p} = -\frac{d\mathbb{E}V_{\text{CL}_{\bar{Z}}}(P)}{dp} \Bigg/ \frac{d\mathbb{E}V_{\text{CL}_{\bar{Z}}}(P)}{dw}.$$

(For example, if the expected value decreases rapidly as p increases, but increases slowly as w increases, then existential catastrophe is relatively important.) As is easy to deduce,

$$\frac{\delta w}{\delta p} = \frac{|F|}{|C|}(\overline{F} - \overline{Z}) \qquad \text{(for } \mathrm{CL}_{\overline{Z}}\text{)}$$

As one would anticipate, this quantity is positive (so *some* sacrifice is warranted) only if $\overline{F}$ is above the critical level $\overline{Z}$, and the importance of existential catastrophe scales linearly with $|F|$.

By contrast, using the value function V_{AU} instead of $V_{\mathrm{CL}_{\overline{Z}}}$, one finds

$$\frac{\delta w}{\delta p} = \frac{|F||Z|(\overline{F} - \overline{Z}) + |F||C|(\overline{F} - \overline{C_w})}{|F||C|\,p + |Z||C| + |C|^2} \qquad \text{(for AU)}$$

This expression is unattractive, but informative, and it simplifies greatly if we make further assumptions about population sizes. Consider the following three cases:

Case 1: $|\mathbf{F}| \gg |\mathbf{C}| \gg |\mathbf{Z}|$. In this case, with a small background population, $\delta w / \delta p$ is approximately $\frac{1}{p}(\overline{F} - w)$.[43] So it is roughly independent of the population sizes, and (in particular) does not scale with $|F|$. Moreover, in this approximation, reducing existential risk is only worthwhile if the future population would have average welfare higher than the current generation.

Case 2: $|\mathbf{F}| \gg |\mathbf{Z}| \gg |\mathbf{C}|$. In this case, where Z is intermediate in size between F and C, $\delta w / \delta p$ is approximately $\frac{1}{p}\frac{|Z|}{|C|}(\overline{F} - \overline{Z})$. In one way, this approximation agrees with $\mathrm{CL}_{\overline{Z}}$: it is worth reducing existential risk insofar as $\overline{F} > \overline{Z}$. Moreover, $\delta w / \delta p$ will tend to be very large, proportional to $|Z|/|C|$. So, in this regime, existential catastrophe may be very important, but its importance is still insensitive to the size of F.

Case 3: $|\mathbf{Z}| \gg |\mathbf{F}| \gg |\mathbf{C}|$. In this case, where the background population is much larger than any of the potential foreground populations, $\delta w / \delta p$ is approximately $\frac{|F|}{|C|}(\overline{F} - \overline{Z})$. This is exactly the value we found using $\mathrm{CL}_{\overline{Z}}$. In particular, for both AU and $\mathrm{CL}_{\overline{Z}}$, the importance of existential catastrophe scales with $|F|$ in this regime.

43. Formally, 'if $a \gg b \gg c$ then x is approximately y' should be interpreted to mean that $\lim_{a/b,\, b/c \to \infty} x / y = 1$.

The most basic qualitative point to take away from this analysis is that $\delta w / \delta p$ increases without bound as we increase *both* $|F|$ and $|Z|$. The fact that possible future and actual background populations are both extremely large suggests that $\delta w / \delta p$ is likely to be large (thus favoring existential risk minimization) for a robust range of the other parameters.

7.4. Numerical Illustration

We now illustrate the preceding qualitative points using plausible numerical estimates of the various population sizes. The results are summarized in Table 1.

For the sizes of the foreground populations, we will suppose that $|C| = 10^{10}$ and $|F| = 10^{17}$. The former is a realistic estimate of the size of the present and near-future human population; the latter is a rough estimate of the potential size of the future human-originating population, supposing that we maintain current population sizes for as long as possible on Earth (see §6.1).

For the size of the background population, we will consider three values, which correspond exactly to Cases 1–3 above. First, just for illustration, we consider $|Z| = 0$ (the case of no background population). Second, more realistically, $|Z| = 10^{13}$, a rounding-down of our most conservative estimate of the number of past mammals, weighted by lifespan and cortical neuron count, from §6.1.1. Finally, a somewhat less conservative estimate of $|Z| = 10^{20}$. This last value corresponds to our lower bound for the number of vertebrates in Earth's past; alternatively, it could correspond to a background population dominated by 1000 alien civilizations, of the same scale that our civilization will achieve if we avoid existential catastrophe.

In terms of average welfare, we have less to go on, but the specific values are also less important for our present purposes. We will assume $\overline{Z} = 0$ (plausible for the case where Z consists mainly of wild animals, somewhat less

Axiology	$\|Z\|$	$\delta w / \delta p$	**Approximation**
$\mathbb{E}$AU	0	1.9999996	$\frac{1}{p}(\overline{F} - w) = 2$
$\mathbb{E}$AU	10^{13}	4.001×10^3	$\frac{1}{p}\frac{\|Z\|}{\|C\|}(\overline{F} - \overline{Z}) = 4 \times 10^3$
$\mathbb{E}$AU	10^{20}	1.999×10^7	$\frac{\|F\|}{\|C\|}(\overline{F} - \overline{Z}) = 2 \times 10^7$
$\mathbb{E}\mathrm{CL}_{\overline{Z}}$	Any	2×10^7	–

Table 1: The importance of avoiding existential catastrophe, as measured by $\delta w / \delta p$, according to $\mathbb{E}$AU or $\mathbb{E}\mathrm{CL}_{\overline{Z}}$ with different background sizes. The other parameters are $\overline{F} = 2$, $|F| = 10^{17}$, $w = 1$, $|C| = 10^{10}$, $\overline{Z} = 0$, and $p = 0.5$. We give just enough significant figures to show disagreement with the approximations developed in Cases 1–3, and stated in the fourth column for comparison.

plausible for the case where it consists mainly of advanced civilizations). If the current generation has positive average welfare, we can then choose units so that $\underline{C}_w = w = 1$. And finally, for simplicity let us suppose $\overline{F} = 2$ (ensuring that reducing existential risk will be worth some sacrifice in all three cases). Finally, we take $p = 0.5$.

Table 1 gives the values of $\delta p / \delta w$ according to $\mathbb{E}\mathrm{AU}$ and $\mathbb{E}\mathrm{CL}_{\overline{Z}}$, under these assumptions, for all three background population sizes. Comparing the values in the third and fourth columns, we see that in this example, with three- or four-order-of-magnitude differences in the population sizes of C, F, and Z, the approximations used in the last subsection are accurate to at least the third significant figure. In particular, in the third case, where $|Z| \gg |F| \gg |C|$, $\mathbb{E}\mathrm{AU}$ agrees with $\mathbb{E}\mathrm{CL}_{\overline{Z}}$ to the third significant figure—preferring even very small reductions in the probability of existential catastrophe over a fairly substantial increase in the welfare of the current generation.

7.5. Conclusions

We have used the standard value functions defining AU and $\mathrm{CL}_{\overline{Z}}$ to analyze the expected value of reducing existential risk. This analysis yields the following conclusions: (1) When the background population Z is small or non-existent, the importance of avoiding existential catastrophe according to AU is approximately independent of population size, depending only on the average welfares of the potential foreground populations. It is therefore unlikely to be astronomically large. (2) When—as suggested by our most conservative estimates—Z is much larger than the current generation but still much smaller than the potential future population, the importance of avoiding existential catastrophe according to AU approximately scales with $|Z|$, and may therefore be extremely large, while still falling well short of its importance according to $\mathrm{CL}_{\overline{Z}}$. (3) Finally, if the background population is much larger even than the potential future population F (as it would be, for instance, if it includes many advanced civilizations elsewhere in the universe), AU agrees closely with $\mathrm{CL}_{\overline{Z}}$ about the importance of avoiding existential catastrophe, treating it as approximately linear in $|F|$.

8. Difficulties for Non-Additive Views

Our primary goal in this paper has been to explore the implications of non-additive axiologies in the context of large background populations. We don't see our results primarily as a reason to reject the views to which they apply, but rather simply as suggesting that their practical implications are more similar to

additive views than one might have thought. However, there are at least two ways in which our results might be taken to support objections to those views, which we briefly explore in this section.

8.1. Repugnant and Sadistic Addition

The Repugnant Conclusion, recall, is the conclusion (implied by TU among other axiologies) that for any two positive welfare levels $w_1 < w_2$, for any population in which everyone has welfare w_2, there is a better population in which everyone has welfare w_1. Avoidance of the Repugnant Conclusion is often seen as a significant desideratum in population axiology. But additive axiologies, as we have defined them, can avoid the Repugnant Conclusion only at the cost of implying the *Strong Sadistic Conclusion*: for any negative welfare level w, for any population in which everyone has welfare w, there is a *worse* population in which everyone has positive welfare (Arrhenius 2000).[44]

One of the motivations for population axiologies with an 'averagist' flavor (like AU and VV) is to avoid these unappealing consequences of additivity. These views do not imply either the Repugnant Conclusion or the Strong Sadistic Conclusion. But a straightforward implication of our results is that they imply *both* of the following, closely related conclusions.

> **Repugnant Addition.** For any positive welfare levels $w_1 < w_2$ and any population X in which everyone has welfare w_2, there is a population Y in which everyone has welfare w_1 and a population Z such that $Y + Z \succ X + Z$.

> **Strong Sadistic Addition.** For any negative welfare level w and any population X in which everyone has welfare w, there is a population Y in which everyone has positive welfare and a population Z such that $X + Z \succ Y + Z$.

Informally, where the Repugnant Conclusion says that for any imaginable utopia, there's a better population in which everyone's life is barely worth living, Repugnant Addition says that it's sometimes better to *add* the latter population

44. Additive axiologies as we define them imply the Repugnant Conclusion if they have a zero or negative critical level, and the Strong Sadistic Conclusion if they have a positive critical level. The broader class of additive views that represent individual welfare levels with vectors rather than real numbers can avoid both conclusions either by allowing incomparability (as 'critical-range' views do) or by treating some positive/negative welfare levels as lexically better/worse than others. But either strategy requires giving up axioms that many find appealing (completeness and 'Archimedean' axioms, respectively).

to a preexisting population. And likewise, where the Strong Sadistic Conclusion says that for any imaginable dystopia, there's a *worse* population in which everyone's life is worth living (if only barely), Strong Sadistic Addition says that it's sometimes *worse* to add the latter population to a preexisting population.

A non-additive view will imply both of these conclusions if it can converge with an additive view with either a positive or a negative critical level. This includes AU, VV1, VV2, and all natural versions of two-factor egalitarianism.[45] AU yields Repugnant Addition, for instance, when there is a large background population with average welfare 0, and yields Strong Sadistic Addition when there is a large background population with positive average welfare.[46] We find these implications nearly as counterintuitive as the original Repugnant and Strong Sadistic Conclusions, though your mileage may vary. And non-additive views that imply *both* Repugnant Addition and Strong Sadistic Addition are, in one respect, worse off than any additive view, which will only imply one of the Repugnant and Strong Sadistic Conclusions.

These observations are not particularly original. Spears and Budolfson (2021) point out the difficulty of avoiding Repugnant Addition, for a broader range of axiologies than we have considered in this paper. And Franz and Spears (2020) show that, under modest assumptions, any view that rejects Mere Addition (including AU and similar views) will imply a weaker version of what we have called Strong Sadistic Addition. But our results provide a systematic and illuminating explanation for the difficulty of avoiding these unpalatable conclusions.

---

45. Specifically, all two-factor theories where adding an individual with welfare 0 can either increase inequality (making things strictly worse, corresponding to a positive critical level) or decrease inequality (making things strictly better, corresponding to a negative critical level). This includes, but isn't limited to, all theories where the measure of inequality I is strictly increasing under Pigou-Dalton transfers.

As before, rank-discounting views present distinctive complications. Since BRD converges with TU, it straightforwardly implies Repugnant Addition. If individual welfare is unbounded above, it also implies Strong Sadistic Addition. If individual welfare is bounded above, then whether it implies Strong Sadistic Addition depends on the shape of the rank-discounting function f (in particular, how fast it approaches its lower limit L). GRD, on the other hand, does not imply Repugnant Addition, and implies Strong Sadistic Addition if and only if individual welfare is unbounded above.

(Many of the above claims depend on the possibility of both positive and negative welfare levels. If positive welfare is impossible, then Repugnant Addition is trivially true and Strong Sadistic Addition trivially false. If negative welfare is impossible, then Strong Sadistic Addition is trivially true.)

46. Indeed, since AU and VV can converge with additive views with *any* critical level, arbitrarily positive or arbitrarily negative, they imply the still more counterintuitive conclusion that for any pair of welfare levels w_1 and w_2 (excluding the minimum and maximum possible welfare levels, if such there be), it is sometimes better to add a population in which everyone has welfare w_1 (e.g., extremely negative welfare) rather than a population in which everyone has welfare w_2 (e.g., extremely positive welfare).

8.2. Exploitability

Our results also show that agents whose choices are guided by non-additive axiologies are vulnerable to a particular kind of exploitation. Suppose, for instance, that we in the Milky Way are all average utilitarians, while the inhabitants of the Andromeda Galaxy are all total utilitarians. And suppose that, the distance between the galaxies being what it is, we can communicate with each other but cannot otherwise interact. Being total utilitarians, the Andromedans would prefer that we act in ways that maximize total welfare in the Milky Way. To bring that about, they might create an astronomical number of welfare subjects with welfare very close to zero—for instance, very small, short-lived animals with mostly bland experiences—and send us evidence that they have done so. We in the Milky Way would then make all our choices under the awareness of a large background population whose average welfare is close to zero. The Andromedans could thus 'force' us to behave like *de facto* total utilitarians, doing the work of total utilitarianism on their behalf.

Agents who accept additive (or, more generally, separable) axiologies, on the other hand, are immune from this sort of exploitation. Thus non-additivists are at a practical disadvantage in strategic interactions with additivists. The *incentive* for others to exploit in this way also makes non-additive views potentially *self-defeating*: adopting and acting on such a view can incentivize other agents to act in ways that make things worse by its lights. For instance, the existence of average utilitarians incentivizes total utilitarians to add individuals with welfare 0 to the population, which makes things worse from the average utilitarian's perspective if the average welfare of the preexisting population is positive.

We ourselves do not see this vulnerability as a particularly weighty reason to reject non-additive views—after all, nearly every agent is vulnerable to *some* forms of exploitation. (On the vulnerabilities of total utilitarians, for instance, see Gustafsson 2022a.) But it is certainly an unwelcome feature, and others may see it as a more severe drawback.

9. Conclusion

We have shown that, in the presence of large enough background populations, a range of non-additive axiologies asymptotically agree with some counterpart additive axiology (either critical-level or, more broadly, prioritarian). And we have argued that the real-world background population is large enough to make these limit results practically relevant. These facts may have important practical implications for tradeoffs between avoiding existential catastrophe and benefit-

ing the current generation: they suggest that AU and kindred axiologies should, in practice, strongly prioritize existential catastrophe avoidance in virtue of the astronomical size of the potential future population, just as additive axiologies seem to do. Thus, arguments for the overwhelming practical importance of avoiding existential catastrophe may not depend on additive separability.

We have left many questions unanswered that might be valuable topics of future research: (1) a more careful characterization of the size and welfare distribution of real-world background populations; (2) how to extend our limit results to the context of risk/uncertainty, including uncertainty about features of the background population; (3) the behavior of a wider range of non-additive axiologies (e.g., incomplete, intransitive, or person-affecting) in the large-background-population limit; and (4) exploring more generally the question of how large the background population needs to be for the limit results to 'kick in', for a wider range of axiologies and choice situations than we considered in §7.

Acknowledgments

For helpful discussion and/or feedback on drafts of this paper, we are grateful to Tomi Francis, Hilary Greaves, Kacper Kowalczyk, Toby Newberry, Toby Ord, Itzhak Rasooly, Dean Spears, Orri Stefánsson, Philip Trammell, and audiences at the University of Oxford, Jagiellonian University, Kansas State University, and the Massachusetts Institute of Technology.

A. Rank-Discounted Utilitarianism

In this appendix, we present two results about rank-discounted utilitarianism that are explained informally in Section 5.2. In stating the results, we will need to restrict the foreground populations under consideration.

Ordinal Convergence on *S*
Axiology $\mathcal{A}$ converges ordinally with $\mathcal{A}'$, relative to background populations of type T, on a set of populations S, if and only if, for any populations X and Y in S, if Z is a sufficiently large population of type T, then

$$X + Z \succ_{\mathcal{A}'} Y + Z \Rightarrow X + Z \succ_{\mathcal{A}} Y + Z.$$

Similarly, we have an obvious notion of *cardinal convergence on S*, where the four populations X_1, Y_1, X_2, Y_2 occurring in the definition of cardinal convergence are restricted to be elements of S.

Having fixed a background distribution $D = Z/|Z|$, say that a population X is *moderate* with respect to D if the lowest welfare level in X is no lower than the the lowest welfare level in D. In other words, for any $x \in \mathcal{W}$ with $X(x) \neq 0$, there is some $z \in \mathcal{W}$ with $z \leq x$ and $D(z) \neq 0$. Then we can state the following result:

Theorem 6. BRD *converges ordinally and cardinally to* TU *relative to background populations with a given distribution D, on the set of populations that are moderate with respect to D.*

Now we turn to GRD. The limiting axiology will be *critical-level leximin*, defined by the following conditions:

> **Critical-Level Leximin** (CLL_c)
> If X and Y have the same size, then $X \succ Y$ if and only if $X \neq Y$ and the least k such that $X_k \neq Y_k$ is such that $X_k \succ Y_k$.
> If X and Y differ only in that Y has additional individuals at welfare level c, then X and Y are equally good.[47]

Although CLL is not additively separable in the narrow sense defined in §2, which requires an assignment of real numbers to each individual, one can check that it is separable, and indeed one can show that it is additively separable in a more general sense, if we allow the contributory value of an individual's welfare to be represented by a vector rather than a single real number.[48]

To state the theorem, fix a set $W \subset \mathcal{W}$ of welfare levels. Say that a population X is *supported* on W if and only if $X(w) = 0$ for all $w \notin W$. And say that W is *covered* by a distribution $D = Z/|Z|$ if and only if there is a welfare level in Z between any two elements of W, a welfare level in Z below every element of W, and welfare level in Z above every element of W.

Theorem 7. *Let $W \subset \mathcal{W}$ be any set of welfare levels, and D a distribution that covers W.* GRD *converges ordinally with* CLL_c *relative to background populations with distribution D, on the set of populations that are supported on W; the critical level c is the highest welfare level occurring in D.*

Here, note, we only consider ordinal convergence, since CLL_c is not defined using a real-valued value function.

47. To compare X and Y in general, use the second condition to find populations X' and Y' that are equally as good as X and Y respectively, but such that $|X'| = |Y'|$, and then compare them using the first condition.

48. See McCarthy et al. (2020: Example 2.7) for details in the constant-population-size case.

B. Proofs

Recall that $\mathcal{W}$ is the set of welfare levels, and $\mathcal{P}$ consists of all non-zero, finitely supported functions $\mathcal{W} \to \mathbb{Z}_+$. By a *type* of population we mean a set $T \subset \mathcal{P}$ that contains populations of arbitrarily large size: for all $n \in \mathbb{N}$ there exists $X \in T$ with $|X| \geq n$.

The following result, while elementary, indicates our general method. Let us say that a function $V : \mathcal{P} \to \mathbb{R}$ is *additive* if $V(X+Y) = V(X) + V(Y)$ for all $X, Y \in \mathcal{P}$. The value functions we have given for additive axiologies are all of this kind.

Lemma 1. *Suppose given $V : \mathcal{P} \to \mathbb{R}$ and a positive function $s : \mathbb{N} \to \mathbb{R}$. Define*

$$V^s(X) := \lim_{|Z| \to \infty} \Big(V(X+Z) - V(Z) \Big) s(|Z|)$$

as Z ranges over populations of some type T. If V^s is an additive function, then the axiology with value function V converges ordinally and cardinally to the axiology with value function V^s, relative to background populations of type T.

Proof. Let X, Y be two populations, and let Z be a background population of type T. To prove ordinal convergence, suppose $V^s(X+Z) > V^s(Y+Z)$. Then, by additivity of V^s, $V^s(X) > V^s(Y)$. Moreover,

$$\lim_{|Z| \to \infty} (V(X+Z) - V(Y+Z)) s(Z) = V^s(X) - V^s(Y) > 0.$$

Therefore, if $|Z|$ is large enough, we must have $V(X+Z) > V(Y+Z)$, as ordinal convergence requires.

For cardinal convergence, consider four populations X_1, Y_1, X_2, Y_2. We have

$$\lim_{|Z| \to \infty} \frac{V(X_1+Z) - V(Y_1+Z)}{V(X_2+Z) - V(Y_2+Z)} = \frac{V^s(X_1) - V^s(Y_1)}{V^s(X_2) - V^s(Y_2)}$$

as long as the denominator $V^s(X_2) - V^s(Y_2) \neq 0$. On the other hand, for any given Z,

$$\frac{V^s(X_1) - V^s(Y_1)}{V^s(X_2) - V^s(Y_2)} = \frac{V^s(X_1+Z) - V^s(Y_1+Z)}{V^s(X_2+Z) - V^s(Y_2+Z)}$$

by additivity. Therefore

$$\lim_{|Z| \to \infty} \frac{V(X_1+Z) - V(Y_1+Z)}{V(X_2+Z) - V(Y_2+Z)} - \frac{V^s(X_1+Z) - V^s(Y_1+Z)}{V^s(X_2+Z) - V^s(Y_2+Z)} = 0$$

as cardinal convergence requires.

Theorem 1. *Average utilitarianism converges ordinally and cardinally to* CL_c*, relative to background populations with average welfare* c*. In fact, for any populations* X, Y, Z*,* if $\overline{Z} = c$ *and*

$$|Z| > \frac{|X|V_{\mathrm{CL}c}(Y) - |Y|V_{\mathrm{CL}c}(X)}{V_{\mathrm{CL}c}(X) - V_{\mathrm{CL}c}(Y)}, \tag{1}$$

then $V_{\mathrm{CL}c}(X) > V_{\mathrm{CL}c}(Y) \Rightarrow V_{\mathrm{AU}}(X+Z) > V_{\mathrm{AU}}(Y+Z)$.

Proof. In this case, a brief calculation shows

$$V_{\mathrm{AU}}(X+Z) - V_{\mathrm{AU}}(Z) = \frac{(\overline{X} - \overline{Z})|X|}{|X|+|Z|} = \frac{V_{\mathrm{CL}c}(X)}{|X|+|Z|}. \tag{3}$$

Setting $s(n) = n$ we find $V^{s}_{\mathrm{AU}}(X) = V_{\mathrm{CL}c}(X)$, in the notation of Lemma 1. That lemma then yields the first statement.

We now verify the more precise second statement directly. Suppose $\overline{Z} = c$, that (1) holds, and that $V_{\mathrm{CL}c}(X) > V_{\mathrm{CL}c}(Y)$. We have to show $V_{\mathrm{AU}}(X+Z) > V_{\mathrm{AU}}(Y+Z)$. Using (3), that desired conclusion is equivalent to

$$\frac{V_{\mathrm{CL}c}(X)}{|X|+|Z|} > \frac{V_{\mathrm{CL}c}(Y)}{|Y|+|Z|}.$$

Cross-multiplying, this is equivalent to

$$V_{\mathrm{CL}c}(X)(|Y|+|Z|) > V_{\mathrm{CL}c}(Y)(|X|+|Z|)$$

or, rearranging,

$$|Z|(V_{\mathrm{CL}c}(X) - V_{\mathrm{CL}c}(Y)) > |X|V_{\mathrm{CL}c}(Y) - |Y|V_{\mathrm{CL}c}(X). \tag{4}$$

Given that $V_{\mathrm{CL}c}(X) - V_{\mathrm{CL}c}(Y) > 0$, the desired conclusion (4) follows from (1).

Theorem 2. *Variable value views converge ordinally and cardinally to* CL_c*, relative to background populations with average welfare* c*.*

Proof. Suppose the variable value view has a value function of the form $V(X) = f(\overline{X})g(|X|)$. Then

$$\begin{aligned} V(X+Z) - V(Z) &= f(\overline{X+Z})g(|X|+|Z|) - f(\overline{Z})g(|Z|) \\ &= f(\overline{X+Z})\big(g(|X|+|Z|) - g(|Z|)\big) + \big(f(\overline{X+Z}) - f(\overline{Z})\big)g(|Z|). \end{aligned}$$

We now apply two lemmas, proved below.

Lemma 2. *We have* $\left(g(|X+Z|)-g(|Z|)\right)|Z| \to 0$ *as* $|Z| \to \infty$.

Lemma 3. *We have* $\left(f(\overline{X+Z})-f(\overline{Z})\right)|Z| \to f'(c)V_{\mathrm{CL}c}(X)$ *as* $|Z| \to \infty$ *with* $\overline{Z}=c$.

Since $f(\overline{X+Z}) \to f(c)$, and $g(|Z|)$ approaches some upper bound L as $|Z| \to \infty$, we find

$$\lim_{|Z|\to\infty}\left(V(X+Z)-V(Z)\right)|Z| = f'(c)V_{\mathrm{CL}c}(X)L$$

as Z ranges over populations with $\overline{Z}=c$. Let $s(n)=\frac{n}{f'(c)L}$. Then we have found

$$\lim_{|Z|\to\infty}\left(V(X+Z)-V(Z)\right)s(|Z|) = V_{\mathrm{CL}c}(X).$$

The result now follows from Lemma 1.

Proof of Lemma 2. Let z be the result of rounding $|Z|/2$ up to the nearest integer. By increasingness and concavity of g, we have[49]

$$0 \le \frac{g(|X+Z|)-g(|Z|)}{|X|} \le \frac{g(|Z|)-g(z)}{|Z|-z} \le \frac{g(|Z|)-g(z)}{|Z|/2}.$$

Cross-multiplying,

$$0 \le \left(g(|X+Z|)-g(|Z|)\right)|Z| \le 2\left(g(|Z|)-g(z)\right)|X|.$$

Since $g(|Z|)$ and $g(z)$ both tend to a common limit L as $|Z| \to \infty$, we find that the right-hand side tends to 0 in that limit. Therefore the expression in the middle also tends to 0.

Proof of Lemma 3. First, if $\overline{X}=c$ then $f(\overline{X+Z})-f(\overline{Z})=0$ and $V_{\mathrm{CL}c}(X)=0$, so the result is trivial in that case. Otherwise, since $X+Z$ tends toward c as $|Z| \to \infty$, we have (by the definition of the derivative)

$$\frac{f(\overline{X+Z})-f(\overline{Z})}{\overline{X+Z}-\overline{Z}} \to f'(c).$$

We have, from (3),

49. The general fact being used about concavity is that, if $x>y>z$, then $\frac{g(x)-g(y)}{x-y} \le \frac{g(y)-g(z)}{y-z}$.

$$\overline{X+Z}-\overline{Z}=\frac{V_{\mathrm{CL}c}(X)}{|X|+|Z|}.$$

Inserting this into the preceding formula, we find

$$(f(\overline{X+Z})-f(\overline{Z}))(|X|+|Z|)\to f'(c)V_{\mathrm{CL}c}(X).$$

Since $(f(\overline{X+Z})-f(\overline{Z}))\,|X|\to 0$, we obtain the desired result.

Proposition 1. *For any populations X and Y, if $X\succ_{\mathrm{TU}} Y$ and $X\succ_{\mathrm{AU}} Y$, then $X\succ_{\mathrm{VV1}} Y$.*

Proof. Since $g:\mathbb{Z}_+\to\mathbb{R}_+$ is non-zero, non-negative, and concave, we must have $g(n)>0$ for all $n>0$ (it is possible, however, that $g(0)=0$). It follows that $V_{\mathrm{VV1}}(X)$ has the same sign as $\overline{X}$. So if $\overline{X}\geq 0\geq\overline{Y}$, it is automatic that $V_{\mathrm{VV1}}(X)>V_{\mathrm{VV1}}(Y)$. (The condition that $X\succ_{\mathrm{AU}} Y$ excludes the case where $\overline{X}=0=\overline{Y}$.) Thus it remains to consider the cases when $\overline{X}$ and $\overline{Y}$ are both positive or both negative. Let us treat the case where both are positive; the case where both are negative is similar, with careful attention to signs.

Suppose $|X|\geq|Y|$. Since g is weakly increasing and $\overline{X}>\overline{Y}$, we find $\overline{X}g(|X|)>\overline{Y}g(|Y|)$. In other words, $V_{\mathrm{VV1}}(X)>V_{\mathrm{VV1}}(Y)$, as required.

Suppose instead $|Y|>|X|$. We have

$$\frac{g(|X|)}{|X|}=\frac{g(|X|)-g(0)}{|X|-0}+\frac{g(0)}{|X|}.$$

Both terms on the right are weakly decreasing in $|X|$ (the first because g is concave). Therefore $g(|X|)/|X|\geq g(|Y|)/|Y|$. This yields

$$\frac{V_{\mathrm{VV1}}(X)}{V_{\mathrm{VV1}}(Y)}=\frac{\overline{X}g(|X|)}{\overline{Y}g(|Y|)}\geq\frac{\overline{X}\,|X|}{\overline{Y}\,|Y|}=\frac{\mathrm{Tot}(X)}{\mathrm{Tot}(Y)}.$$

Since $\mathrm{Tot}(X)>\mathrm{Tot}(Y)$, we conclude that $V_{\mathrm{VV1}}(X)>V_{\mathrm{VV1}}(Y)$.

Theorem 3. *Suppose V is a value function of the form $V(X)=\mathrm{Tot}(X)-I(X)|X|$, or else $V(X)=\overline{X}-I(X)$, where I is a differentiable function of the distribution of X. Then the axiology $\mathcal{A}$ represented by V converges ordinally and cardinally with an additive axiology, relative to background populations with any fixed distribution D; specifically, it converges with the additive axiology with weighting function given by*

$$f(w)=\lim_{t\to 0^+}\frac{V(D+t1_w)-V(D)}{t}.$$

If the Pareto principle holds with respect to $\mathcal{A}$, then f is weakly increasing, and if Pigou-Dalton transfers are weak improvements, then f is concave.

Remark 1. Before proving Theorem 3, we should explain the requirement that 'I is a differentiable function of the distribution of X'. It has two parts. First, let $\mathcal{P}_{\mathbb{R}}$ be the set of finitely-supported, non-zero functions $\mathcal{W} \to \mathbb{R}_+$. Let $\mathcal{D} \subset \mathcal{P}_{\mathbb{R}}$ be the subset of distributions, that is, those functions that sum to 1. The first part of the requirement is that there is a function $\tilde{I} : \mathcal{D} \to \mathbb{R}$ such that $I(X) = \tilde{I}(X/|X|)$. In that sense, $I(X)$ is just a function of the distribution of X. Another way to put this is that I can be extended to a function on all of $\mathcal{P}_{\mathbb{R}}$ that is scale-invariant, that is, $I(nX) = I(X)$ for all reals $n > 0$ and all $X \in \mathcal{P}_{\mathbb{R}}$. The second part of the requirement is that I, so extended, is differentiable, in the following sense:[50] for all $P, Q \in \mathcal{P}_{\mathbb{R}}$, the limit

$$\partial_Q I(P) := \lim_{t \to 0^+} \frac{I(P + tQ) - I(P)}{t}$$

exists and is linear as a function of Q. In effect, $Q \mapsto \partial_Q I(P)$ is the best linear approximation of $I - I(P)$. In practice we only need I to be differentiable at the background distribution D.

Proof. Let Z range over background populations with the given distribution $D = Z/|Z|$. Thus Z is of the form nD for some $n > 0 \in \mathbb{R}$.

Define $s(n) = 1$, in the case of TU-based egalitarianism, and $s(n) = n$ in the case of AU-based egalitarianism. Noting that value functions of the assumed form can be evaluated not only on $\mathcal{P}$ but on the larger set $\mathcal{P}_{\mathbb{R}}$ (see Remark 1), we have

$$V(nX) = (n / s(n))V(X).$$

We can then see that V^s (as defined in Lemma 1) is the directional derivative of V at D:

$$\begin{aligned} V^s(X) &= \lim_{|Z| \to \infty} \big(V(Z + X) - V(Z)\big) s(|Z|) \\ &= \lim_{n \to \infty} \big(V(nD + X) - V(nD)\big) s(n) \\ &= \lim_{n \to \infty} \frac{V(D + \frac{1}{n}X) - V(D)}{1/n} =: \partial_X V(D). \end{aligned}$$

Given that I is differentiable as in Remark 1, this function is linear in X and therefore represents an additive axiology $\mathcal{A}'$. More specifically, for each welfare level w let 1_w be a population with one person at level w. We then have

50. This can also be interpreted as a differentiability requirement directly on $\tilde{I}$: it should have a linear Gâteaux derivative.

$$V^s(X) = \sum_{w \in \mathcal{W}} X(w) f(w) \quad \text{with} \quad f(w) = \partial_{1_w} V(D)$$

as claimed in the statement of the theorem. In particular, for totalist egalitarianism, we find that

$$f(w) = w - \partial_{1_w} I(D) - I(D).$$

Similarly, for averagist egalitarianism,

$$f(w) = w - \partial_{1_w} I(D) - \bar{D}.$$

Now, suppose X^+ differs from X in that one person is better off, say with welfare v instead of w. If the Pareto principle holds with respect to $\mathcal{A}$, then $V(X^+ + Z) > V(X + Z)$ for all Z; by convergence, we cannot have $V^s(X^+) < V^s(X)$. It follows that $f(v) \geq f(w)$; thus f is weakly increasing. By the same logic, Pigou-Dalton transfers do not make things worse with respect to $\mathcal{A}'$, and it follows that f is concave.

Theorem 4. MDT *converges ordinally and cardinally to* PR, *relative to background populations with a given distribution* D. *Specifically,* MDT_α *converges with* PR_f, *the prioritarian axiology whose weighting function is*

$$f(w) = w - 2\alpha \ \text{MD}(w, D) + \alpha \ \text{MD}(D).$$

Here $\text{MD}(w, D) := \sum_{x \in \mathcal{W}} D(x)|x - w|$ *is the average distance between w and the welfare levels occurring in* D.

Proof. Define $\langle X, Y \rangle = \Sigma_{x,y \in \mathcal{W}} X(x) Y(y) |x - y|$. Then $\text{MD}(Z) = \langle Z, Z \rangle / |Z|^2$. It is easy to check that $\partial_X \langle Z, Z \rangle = 2\langle X, Z \rangle$ and therefore

$$\partial_X \ \text{MD}(Z) = 2\frac{\langle X, Z \rangle}{|Z|^2} - 2\frac{\langle Z, Z \rangle}{|Z|^3}|X|.$$

In particular, MD is differentiable and Theorem 3 applies. We know from equation (5) in the proof of Theorem 3 that MDT converges with the additive axiology $\mathcal{A}'$ with weighting function

$$\begin{aligned} f(w) &= w - \alpha \partial_{1_w} \text{MD}(D) - \alpha \text{MD}(D) \\ &= w - 2\alpha \langle 1_w, D \rangle - \alpha \ \text{MD}(D) \\ &= w - 2\alpha \ \text{MD}(w, D) + \alpha \ \text{MD}(D). \end{aligned}$$

Theorem 5. QAA *converges ordinally and cardinally to* PR*, relative to background populations with a given distribution* D*. Specifically,* QAA_g *converges with* PR_f*, the prioritarian axiology whose weighting function is*

$$f(w) = g(w) - g(\mathrm{QAM}(D)).$$

Proof. Theorem 3 applies, with $I(X) = \overline{X} - \mathrm{QAM}(X)$. (We omit the proof that this I is differentiable.) We have, then, convergence with prioritarianism with a priority weighting function

$$f(w) = \partial_{1_w} \mathrm{QAM}(D) = \frac{g(w) - \sum_{x \in \mathcal{W}} D(x)g(x)}{g'(\mathrm{QAM}(D))}.$$

Since the background distribution D is fixed, this differs from the stated priority weighting function only by a positive scalar (i.e., the denominator), which does not affect which axiology the value function represents.

Theorem 6. BRD *converges ordinally and cardinally to* TU *relative to background populations with a given distribution* D*, on the set of populations that are moderate with respect to* D*.*

Proof. Suppose that the weighting function f has a horizontal asymptote at $L > 0$. As in Lemma 1 it suffices to show that $\lim_{|Z| \to \infty} V(X+Z) - V(Z) = L\mathrm{Tot}(X)$, as Z ranges over populations with distribution D, and on the assumption that X is moderate with respect to D.

Write $X_{\leq w} = \Sigma_{x \leq w} X(w)$ for the number of people in X with welfare at most w, and similarly $X_{<w} = \Sigma_{x<w} X(w)$. Separating out contributions from X and contributions from Z, we have

$$\begin{aligned} V(X+Z) - V(Z) = &\sum_{w \in \mathcal{W}} \sum_{i=1}^{X(w)} f(Z_{\leq w} + X_{<w} + i)w \\ &+ \sum_{w \in \mathcal{W}} \sum_{i=1}^{Z(w)} f(Z_{<w} + X_{<w} + i) - f(Z_{<w} + i)\big)w. \end{aligned}$$

The assumption that X is moderate means that, in those cases where $X(w) \geq 1$, so that the first inner sum is non-trivial, we also have $Z_{\leq w} \to \infty$. Therefore each summand in the first double-sum tends to Lw. The first double sum then converges to $\Sigma_{w \in \mathcal{W}} X(w)Lw = L\mathrm{Tot}(X)$. It remains to show that the second double sum converges to 0. Call the summand in that double sum $S(w, i)$.

Since there are finitely many w for which $Z(w) \geq 1$ (i.e., for which the inner sum is non-trivial), it suffices to show that, for each such w, the inner sum con-

verges to 0. If $X_{<w} = 0$, then the inner sum is identically zero, so we can assume $X_{<w} \geq 1$. We can also assume that $Z_{<w}$ is large enough that f is convex in the relevant range; then

$$0 \leq |S(w,i)| \leq |f(Z_{<w} + X_{<w}) - f(Z_{<w})||w|.$$

Moreover, the number of terms in the inner sum, $Z(w)$, is proportional to $Z_{<w}$. It remains to apply the following elementary lemma with $n = Z_{<w}$ and $m = X_{<w}$.

Lemma 4. *If f is an eventually convex function decreasing to a finite limit, then $n(f(n+m) - f(n)) \to 0$ as $n \to \infty$.*

This is just a small variation on Lemma 2, and we omit the proof.

Theorem 7. *Let $W \subset \mathcal{W}$ be any set of welfare levels, and D a distribution that covers W.* GRD *converges ordinally with* CLL_c *relative to background populations with distribution D, on the set of populations that are supported on W; the critical level c is the highest welfare level occurring in D.*

Proof. Suppose X and Y are supported on W, and $X \succ_{\text{CLL}} Y$. Let Z be a population with distribution D, so $Z = nD$ for some $n > 0$. We have to show that $X + Z \succ_{\text{GRD}} Y + Z$ for all n large enough.

Let X' and Y' be populations of equal size, obtained from X and Y by adding people at the critical level c. By the second condition characterizing CLL, X' is just as good as X, and Y' just as good as Y. Therefore, the assumption that $X \succ_{\text{CLL}} Y$ implies that $X' \succ_{\text{CLL}} Y'$. According to the first condition characterizing CLL, we have $X'_k > Y'_k$ for the first k such that $X'_k \neq Y'_k$. Now, since D covers W, no welfare level occurring in X' or Y' is higher than c. Thus $c \geq X'_k > Y'_k$, and it follows that $Y'_k = Y_k$. For brevity define $w := Y_k$.

Let v be the next welfare level occurring in $X + Y$ above w. If there is no such welfare level, then define $v = c + 1$.

We can decompose Z (and similarly other populations) as $Z = Z_- + Z_w + Z_0 + Z_+$, where Z_- only involves welfare levels in the interval $(-\infty, w)$, Z_w involves only w, Z_0 only involves welfare levels in (w, v), and Z_+ only involves those in $[v, \infty)$. Note that $X_0 = Y_0 = 0$, because of the way v was chosen. We can therefore write

$$X + Z = (X_- + Z_- + Z_w) + X_w + Z_0 + (X_+ + Z_+).$$

The populations on the right are written in rank-order; that is, every welfare level in $(X_- + Z_- + Z_w)$ is below every welfare level in X_w, and so on. This makes it easy to apply the value function $V = V_{\text{GRD}}$:

$$\begin{aligned} V(X+Z) = V(X_- + Z_- + Z_w) &+ \beta^{|X_- + Z_- + Z_w|} V(X_w) \\ &+ \beta^{|X_- + Z_- + Z_w + X_w|} V(Z_0) \\ &+ \beta^{|X_- + Z_- + Z_w + X_w + Z_0|} V(X_+ + Z_+). \end{aligned}$$

A similar expression holds for Y in place of X. Note that $X_- = Y_-$ because of the way w was chosen. Combining expressions for $V(X+Z)$ and $V(Y+Z)$, and dividing by a common factor, we find

$$\frac{V(X+Z) - V(Y+Z)}{\beta^{|X_- + Z_- + Z_w|}} = V(X_w) - V(Y_w) + (\beta^{|X_w|} - \beta^{|Y_w|}) V(Z_0) + R \tag{6}$$

where the remainder R is given by

$$R = \beta^{|Z_0|}\big(\beta^{|X_w|} V(X_+ + Z_+) - \beta^{|Y_w|} V(Y_+ + Z_+)\big).$$

Our goal is to show that the right-hand side of (6) is positive when n is sufficiently large, for if it is positive then $V(X+Z) > V(Y+Z)$ and thus $X + Z \succ_{\text{GRD}} Y + Z$.

To simplify (6), we use the standard fact that $\Sigma_{i=1}^{m} \beta^i = (1-\beta^m)\frac{\beta}{1-\beta}$. Since $V(X_w) = \Sigma_{i=1}^{|X_w|} \beta^i w$, and similarly for $V(Y_w)$, we find

$$V(X_w) - V(Y_w) = (\beta^{|Y_w|} - \beta^{|X_w|}) \frac{\beta w}{1-\beta}.$$

Substituting this into (6) and rearranging, we find

$$\frac{V(X+Z) - V(Y+Z)}{\beta^{|X_- + Z_- + Z_w|}} = (\beta^{|X_w|} - \beta^{|Y_w|})(V(Z_0) - \frac{\beta w}{1-\beta}) + R.$$

To conclude that $V(X+Z) > V(Y+Z)$ for all n large enough, it suffices to show

$$\beta^{|X_w|} - \beta^{|Y_w|} > 0, \qquad \lim_{n\to\infty} V(Z_0) > \frac{\beta w}{1-\beta}, \quad \text{and} \quad \lim_{n\to\infty} R = 0.$$

For the first of these conditions, note that $|X_w| < |Y_w|$, by the way w was chosen; therefore $\beta^{|X_w|} - \beta^{|Y_w|} > 0$.

For the second, we claim that $D_0 \neq 0$: that is, some welfare level in (w, v) occurs in D. There are two cases. First, if $w, v \in W$, some welfare level in (w, v) occurs in D, because D covers W. Otherwise, $w \in W$ but $v = c + 1$. Then $c \in (w, v)$, and c occurs in D. Having proved the claim, let v' be the *lowest* welfare level occurring in D_0. Since Z has distribution D, v' is also the lowest welfare level in Z_0. Then $\lim_{n\to\infty} V(Z_0) = \Sigma_{i=1}^{\infty} \beta^i v' = \frac{\beta v'}{1-\beta} > \frac{\beta w}{1-\beta}$.

Finally, we will have $R \to 0$ as long as $\beta^{|Z_0|} \to 0$, since the second, complicated factor in the definition of R is bounded as $n \to \infty$. And since $|Z_0| = n|D_0|$, it suffices that $D_0 \neq 0$, as we already showed.

References

Adler, Matthew (2009). Future Generations: A Prioritarian View. *George Washington Law Review*, *77* (5–6), 1478–520.

Adler, Matthew (2011). *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis*. Oxford University Press.

Arneson, Richard J. (2000). Luck Egalitarianism and Prioritarianism. *Ethics*, *110* (2), 339–49. https://doi.org/10.1086/233272

Arntzenius, Frank (2014). Utilitarianism, Decision Theory and Eternity. *Philosophical Perspectives*, *28* (1), 31–58.

Arrhenius, Gustaf (2000). An Impossibility Theorem for Welfarist Axiologies. *Economics and Philosophy*, *16* (2), 247–66. https://doi.org/10.1017/s0266267100000249

Asheim, Geir B. and Stéphane Zuber (2014). Escaping the Repugnant Conclusion: Rank-Discounted Utilitarianism with Variable Population. *Theoretical Economics*, *9* (3), 629–50.

Bar-On, Yinon M., Rob Phillips, and Ron Milo (2018). The Biomass Distribution on Earth. *Proceedings of the National Academy of Sciences*, *115* (25), 6506–11.

Bentham, Jeremy (1789). *An Introduction to the Principles of Morals and Legislation*. T. Payne and Son.

Blackorby, Charles, Walter Bossert, and David Donaldson (1997). Critical-Level Utilitarianism and the Population-Ethics Dilemma. *Economics and Philosophy*, *13* (2), 197–230. https://doi.org/10.1017/S026626710000448X

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

Bostrom, Nick (2003). Astronomical Waste: The Opportunity Cost of Delayed Technological Development. *Utilitas*, *15* (3), 308–14.

Bostrom, Nick (2011). Infinite Ethics. *Analysis and Metaphysics*, *10*, 9–59.

Bostrom, Nick (2013). Existential Risk Prevention as Global Priority. *Global Policy*, *4* (1), 15–31.

Broad, C. D. (1914). The Doctrine of Consequences in Ethics. *International Journal of Ethics*, *24* (3), 293–320.

Broome, John (1997). Is Incommensurability Vagueness? In Ruth Chang (Ed.), *Incommensurability, Incomparability and Practical Reason* (67–89). Harvard University Press.

Browning, Heather (2020). *If I Could Talk to the Animals: Measuring Animal Welfare* (Doctoral dissertation). Australian National University

Buchak, Lara (2017). Taking Risks Behind the Veil of Ignorance. *Ethics*, *127* (3), 610–44. https://doi.org/10.1086/690070

Budolfson, Mark and Dean Spears (2022). Does the Repugnant Conclusion Have Important Implications for Axiology or for Public Policy? In Gustaf Arrhenius, Krister Bykvist, and Tim Campbell (Eds.), *Oxford Handbook of Population Ethics* (350–68). Oxford University Press.

Carlson, Erik (1995). *Consequentialism Reconsidered*. Kluwer.

Crisp, Roger (2003). Equality, Priority, and Compassion. *Ethics*, *113* (4), 745–63. https://doi.org/10.1086/373954

de Lazari-Radek, Katarzyna and Peter Singer (2014). *The Point of View of the Universe: Sidgwick and Contemporary Ethics*. Oxford University Press.

Fishburn, Peter C. (1970). Intransitive Indifference in Preference Theory: A Survey. *Operations Research*, *18* (2), 207–28.

Fleurbaey, Marc (2010). Assessing Risky Social Situations. *Journal of Political Economy*, *118* (4), 649–80.

Frankfurt, Harry (1987). Equality as a Moral Ideal. *Ethics*, *98* (1), 21–43. https://doi.org/10.1086/292913

Franz, Nathan and Dean Spears (2020). Mere Addition is Equivalent to Avoiding the Sadistic Conclusion in All Plausible Variable-Population Social Orderings. *Economics letters*, *196*, 109547.

Greene, Brian (2004). *The Fabric of the Cosmos: Space, Time and the Texture of Reality*. Random House.

Gustafsson, Johan E. (2020). Population Axiology and the Possibility of a Fourth Category of Absolute Value. *Economics and Philosophy*, *36* (1), 81–110.

Gustafsson, Johan E. (2022a). Bentham's Mugging. *Utilitas*, *34* (4), 386–91.

Gustafsson, Johan E. (2022b). Our Intuitive Grasp of the Repugnant Conclusion. In Gustaf Arrhenius, Krister Bykvist, and Tim Campbell (Eds.), *The Oxford Handbook of Population Ethics* (371–89). Oxford University Press.

Hardin, Garrett (1968). The Tragedy of the Commons. *Science*, *162* (3859), 1243–48.

Harsanyi, John C. (1977). Morality and the Theory of Rational Behavior. *Social Research*, *44* (4), 623–56.

Helliwell, John F., Richard Layard, and Jeffrey D. Sachs (2019). *World Happiness Report 2019*. Sustainable Development Solutions Network.

Horta, Oscar (2010). Debunking the Idyllic View of Natural Processes: Population Dynamics and Suffering in the Wild. *Telos: Revista Iberoamericana de Estudios Utilitaristas*, *17* (1), 73–90.

Hudson, James L. (1987). The Diminishing Marginal Value of Happy People. *Philosophical Studies*, *51* (1), 123–37. https://doi.org/10.1007/BF00353967

Hurka, Thomas (1982a). Average Utilitarianisms. *Analysis*, *42* (2), 65–69. https://doi.org/10.1093/analys/42.2.65

Hurka, Thomas (1982b). More Average Utilitarianisms. *Analysis*, *42* (3), 115–19. https://doi.org/10.1093/analys/42.3.115a

Hurka, Thomas (1983). Value and Population Size. *Ethics*, *93* (3), 496–507. https://doi.org/10.1086/292462

Hutcheson, Francis (1738). *An Inquiry into the Original of our Ideas of Beauty and Virtue, In Two Treatises* (4th ed.). D. Midwinter, A. Bettesworth, and C. Hitch, J. and J. Pemberton, R. Ware, C. Rivington, F. Clay, A. Ward, J. and P. Knap. (Original work published 1725)

Kagan, Shelly (2019). *How to Count Animals, More or Less*. Oxford University Press.

Kaneda, Toshiko and Carl Haub. How Many People Have Ever Lived on Earth? Population Reference Bureau. Retrieved from https://www.prb.org/howmanypeoplehaveeverlivedonearth/

Knobe, Joshua, Ken D. Olum, and Alexander Vilenkin (2006). Philosophical Implications of Inflationary Cosmology. *The British Journal for the Philosophy of Science*, *57* (1), 47–67.

Kowalczyk, Kacper (2020). *Persons, Populations, and Value* (Doctoral dissertation). University of Oxford.

McCarthy, David (2015). Distributive Equality. *Mind*, *124* (496), 1045–109. https://doi.org/10.1093/mind/fzv028

McCarthy, David, Kalle Mikkola, and Teruji Thomas (2020). Utilitarianism With and Without Expected Utility. *Journal of Mathematical Economics*, *87*, 77–113.

Mill, John Stuart (1863). *Utilitarianism*. Parker, Son, and Bourne.

Ng, Yew-Kwang (1989). What Should We Do About Future Generations? Impossibility of Parfit's Theory X. *Economics and Philosophy*, *5* (2), 235–53. https://doi.org/10.1017/s0266267100002406

Ng, Yew-Kwang (1995). Towards Welfare Biology: Evolutionary Economics of Animal Consciousness and Suffering. *Biology and Philosophy*, *10* (3), 255–85. https://doi.org/10.1007/BF00852469

Norwood, F. Bailey and Jayson L. Lusk (2011). *Compassion, by the Pound: The Economics of Farm Animal Welfare*. Oxford University Press.

Ord, Toby (2020). *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury Publishing.

Ord, Toby (2021). The Edges of Our Universe. arXiv: https://arxiv.org/abs/2104.01191v2 [gr-qc]

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

Parfit, Derek (1997). Equality and Priority. *Ratio*, *10* (3), 202–21. https://doi.org/10.1111/1467-9329.00041

Pressman, Michael (2015). A Defence of Average Utilitarianism. *Utilitas*, *27* (4), 389–424. https://doi.org/10.1017/s0953820815000072

Rabinowicz, Wlodzimierz (1989). Act-Utilitarian Prisoner's Dilemmas. *Theoria*, *55* (1), 1–44. https://doi.org/10.1111/j.1755-2567.1989.tb00720.x

Roth, Gerhard and Ursula Dicke (2005). Evolution of the Brain and Intelligence. *Trends in Cognitive Sciences*, *9* (5), 250–57.

Shulman, Carl (2014). Population Ethics and Inaccessible Populations. Unpublished manuscript.

Sidgwick, Henry (1907). *The Methods of Ethics* (7th ed.). Macmillan and Company. (Original work published 1874)

Smil, Vaclav (2013). *Harvesting the Biosphere: What We Have Taken from Nature*. The MIT Press.

Sotala, Kaj and Lukas Gloor (2017). Superintelligence as a Cause or Cure for Risks of Astronomical Suffering. *Informatica*, *41* (4), 389–400.

Spears, Dean and Mark Budolfson (2021). Repugnant Conclusions. *Social Choice and Welfare*, *57* (3), 567–88.

Thomas, Teruji (2022). Separability and Population Ethics. In Gustaf Arrhenius, Krister Bykvist, Tim Campbell, and Elizabeth Finneron-Burns (Eds.), *The Oxford Handbook of Population Ethics* (271–95). Oxford University Press.

Tomasik, Brian. How Many Wild Animals Are There? Retrieved from https://reducing-suffering.org/how-many-wild-animals-are-there/

Vardanyan, Mihran, Roberto Trotta, and Joseph Silk (2009). How Flat Can You Get? A Model Comparison Perspective on the Curvature of the Universe. *Monthly Notices of the Royal Astronomical Society*, *397* (1), 431–44.

Vardanyan, Mihran, Roberto Trotta, and Joseph Silk (2011). Applications of Bayesian Model Averaging to the Curvature and Size of the Universe. *Monthly Notices of the Royal Astronomical Society: Letters*, *413* (1), L91–L95.

Weirich, Paul (1983). Utility Tempered with Equality. *Noûs*, *17* (3), 423–39. https://doi.org/10.2307/2215258