Non-Additive Axiologies in Large Worlds

Christian Tarsney* and Teruji Thomas*

Version 3.0, April 2022
(Latest version here.)

Abstract

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 is practically important: among other things, additive axiologies generally assign great importance to large changes in population size, and therefore tend to support strongly prioritizing the long-term survival of humanity over the interests of the present generation. Non-additive axiologies, on the other hand, need not support this kind of reasoning. 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 some additive axiology—for instance, average utilitarianism converges to critical-level utilitarianism and various egalitarian theories converge to prioritiarianism. 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

*Global Priorities Institute, Faculty of Philosophy, University of Oxford. Comments welcome: christian.tarsney@philosophy.ox.ac.uk, teru.thomas@oxon.org.

a central dividing line in axiology, between axiologies that are *additively separable* (hereafter 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.

The abstract question of additive separability has 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 utilitarianism 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, the abstract question of 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.

The aim of this paper, however, is to show that under certain circumstances, a wide range of non-additive axiologies converge in their implications with some counterpart additive axiology. This convergence 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.

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]

[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, Budolfson and Spears (ms) consider the implications of background populations for issues related to the 'Repugnant Conclusion' (see §10.1 below). And, as we discovered while revising this paper,

With this in mind, our argument has two steps. First, we prove several results to the effect that, in the large-background-population limit (i.e., as the size of the background population Z tends to infinity), non-additive axiologies of various types converge with counterpart additive axiologies. Thus, these axiologies are effectively additive in the presence of sufficiently large background populations. 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 (and, more tentatively, variable value views) assign to avoiding existential catastrophe.

The paper proceeds as follows: section 2 introduces some formal concepts and notation, while section 3 formally defines additive separability and describes some important classes of additive axiologies. In sections 4–5, we survey several important classes of non-additive axiologies and show that they become additive in the large-background-population limit. In section 6, we argue that real-world background populations are large, and also briefly consider what their welfare distributions might look like. In sections 7–8, we answer two objections: that we should simply ignore background populations for decision-making purposes, and that we should apply 'axiological weights' to non-human welfare subjects that reduce their contribution to the size of the background population. Section 9 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 10 briefly describes three more potential implications of our results: they make it harder to avoid (a generalization of) the Repugnant Conclusion, help us to extend non-additive axiologies to infinite-population contexts, and suggest that agents who accept non-additive axiologies may be vulnerable to a novel form of manipulation. Section 11 is the conclusion.

---

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 (Shulman, 2014).

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 (except in §10.2, where we consider the significance of our results for infinite ethics). 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 non-zero, finitely supported function from the set $\mathscr{W}$ of all possible welfare levels to the set $\mathbb{Z}_+$ of all non-negative integers, specifying the number of welfare subjects at each level. Despite this formalism, we'll say that a welfare level w *occurs* in a population X to mean that $X(w) \neq 0$. An *axiology* $\succ_{\mathscr{A}}$ is a strict partial order $\succ_{\mathscr{A}}$ on the set $\mathscr{P}$ of all populations, with '$X \succ_{\mathscr{A}} Y$' meaning that population X is better than population Y according to $\succ_{\mathscr{A}}$.[3]

Almost all the axiologies we will consider in this paper can be represented by a *value function* $V_{\mathscr{A}} : \mathscr{P} \to \mathbb{R}$, meaning that $X \succ_{\mathscr{A}} Y$ if and only if $V_{\mathscr{A}}(X) > V_{\mathscr{A}}(Y)$.[4]

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

$$|X| \coloneqq \sum_{w \in \mathscr{W}} X(w).$$

Similarly, the total welfare is

$$\mathrm{Tot}(X) \coloneqq \sum_{w \in \mathscr{W}} X(w)w.$$

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

[3] 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 $(X \succ Z \leftrightarrow Y \succ Z)$.

[4] 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. 3 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* $\mathscr{V}_{\mathscr{A}}$ of value functions—in the sense that $X \succ_{\mathscr{A}} Y$ if and only if $V(X) > V(Y)$ for all $V \in \mathscr{V}_{\mathscr{A}}$—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.

assume that $\mathscr{W}$ 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

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

is also well-defined.

3 Additivity

We can now give a precise definition of additive separability.

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 $w \in \mathscr{W}$, $(X+Y)(w) = X(w) + Y(w)$. An axiology is *separable* if, for any populations X, Y, and Z,

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

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

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

with $f : \mathscr{W} \to \mathbb{R}$. 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 we will consider below have this feature. So for our purposes, the additive/non-additive and separable/non-separable distinctions are more or less extensionally equivalent.[5]

[5] For a detailed discussion of separability principles in population ethics, see Thomas (forthcoming). The main difference between separability and additivity is that the latter, but not the former, entails completeness (see fn. 4) 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.

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.[6]

Total Utilitarianism (TU)

$$V_{\mathrm{TU}}(X) = \mathrm{Tot}(X) = \sum_{w \in \mathscr{W}} X(w)w = \overline{X}|X|.$$

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.[7]

Critical-Level Utilitarianism (CL)

$$V_{\mathrm{CL}}(X) = \sum_{w \in \mathscr{W}} X(w)(w-c) = \mathrm{Tot}(X) - c|X| = (\overline{X}-c)|X|$$

for some constant $c \in \mathscr{W}$ (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 critical level $c = 0$. Note that, 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.[8]

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

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

[7] Critical-level views have been defended by Blackorby et al. (1997, 2005), among others.

[8] But a positive critical level also brings its own, arguably greater drawbacks—e.g., the Strong Sadistic Conclusion (Arrhenius, 2000).

same welfare improvement for someone who is already very well off. This intuition is captured by *prioritarian* theories.[9]

Prioritarianism (PR)

$$V_{\mathrm{PR}}(X) = \sum_{w \in \mathscr{W}} X(w) f(w)$$

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

CL is a special case of PR where f is linear, and TU is a special case where f is linear and passes through the origin. 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 two categories of non-additive axiologies and show that, in the presence of large enough background populations, they converge with some additive axiology. 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.

First, though, what do we mean by converging to an additive (or any other) axiology? The claim makes sense relative to a specified *type* of background population, e.g., all those having a certain average level of welfare.

Convergence
Axiology $\mathscr{A}$ converges to $\mathscr{A}'$ 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

$$X + Z \succ_{\mathscr{A}'} Y + Z \implies X + Z \succ_{\mathscr{A}} Y + Z.$$

[9] 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.

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

$$X \succ_{\mathscr{A}'} Y \implies X+Z \succ_{\mathscr{A}} Y+Z.$$

We can, in other words, compare $X+Z$ and $Y+Z$ with respect to $\mathscr{A}$ by comparing X and Y with respect to $\mathscr{A}'$—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 implication need not be a biconditional; thus, when $\mathscr{A}'$ does not have a strict preference between $X+Z$ and $Y+Z$ (e.g., when it is indifferent between them), convergence to $\mathscr{A}'$ does not imply anything about how $\mathscr{A}$ ranks those two populations. Because of this, every axiology converges to 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 to more discriminating axiologies. Specifically, we will only ever consider axiologies that satisfy the Pareto principle (which we discuss in §5.1).

4.1 Average utilitarianism

Average utilitarianism identifies the value of a population with the average welfare level of that population.[10]

Average Utilitarianism (AU)

$$V_{\mathrm{AU}}(X) = \overline{X} = \sum_{w \in \mathbb{W}} \frac{X(w)}{|X|}\, w.$$

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

[10] 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 (forthcoming)), 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,b) 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 §7.

Theorem 1. *Average utilitarianism converges 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 $V_{\mathrm{CL}_c}(X) > V_{\mathrm{CL}_c}(Y) \implies V_{\mathrm{AU}}(X+Z) > V_{\mathrm{AU}}(Y+Z)$.

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 (§§6–11).

4.2 '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.[11] It is useful to distinguish two types of this view, the second more general than the first.

Variable Value I (VV1)
$V_{\mathrm{VV1}}(X) = \overline{X}\, g(|X|)$, where $g : \mathbb{Z}_+ \to \mathbb{R}_+$ is increasing, concave, non-zero, 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 importance to the average level of welfare:

Variable Value II (VV2)
$V_{\mathrm{VV2}}(X) = f(\overline{X})g(|X|)$, where $f : \mathbb{R} \to \mathbb{R}$ is differentiable and strictly increasing, and $g : \mathbb{Z}_+ \to \mathbb{R}_+$ is increasing, concave, non-zero, 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 behavior as AU, converging to a critical-level view with the critical level given by the average welfare of the background population.

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

Theorem 2. *Variable value views converge 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* function that is strictly increasing, strictly 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.[12]

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) terms, one of which is a measure of inequality, and (ii) 'rank-discounting' views, which give less weight to the welfare of individuals who

[12] 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 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.

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, ms).

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
$V(X) = \text{Tot}(X) - I(X)|X|$, 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.[13]

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 $X/|X|: \mathscr{W} \to \mathbb{R}$ 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 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, i.e. $D \coloneqq Z/|Z|$. 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 $V(X) = \text{Tot}(X) - I(X)|X|$, or else $V(X) = \overline{X} - I(X)$, where I is a differentiable function of the distribution of X. Then the axiology $\mathscr{A}$ represented by V converges to*

[13] 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.

an additive axiology relative to background populations with any given distribution D, with weighting function[14]

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

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

A few points in the theorem require further explanation. We will explain the relevant notion of *differentiability* when it comes to the proof (see Remark 1 in the appendix); as usual, functions that are easy to write down tend to be differentiable, but it isn't automatic. 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 assumed to be 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 this result, 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:

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

$\mathrm{MD}(X)$ represents the average welfare inequality between any two individuals in X. $\mathrm{MD}(X)|X|$ can therefore be understood as measuring total pairwise inequality in X. Consider, then, the following totalist two-factor view:

Mean Absolute Difference Total Egalitarianism (MDT)

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

[14] Here $1_w \in \mathscr{P}$ 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 $\mathscr{W} \to \mathbb{R}_+$, such as, in particular, D and $D + t1_w$.

where $\alpha \in (0, 1/2)$ is a constant that determines the relative importance of inequality.[15]

Second, consider the following averagist two-factor view, which identifies overall value with a quasi-arithmetic mean of welfare:[16]

Quasi-Arithmetic Average Egalitarianism (QAA)

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

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

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

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

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

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

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

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

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

[15] 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.

[16] 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.

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.[17]

Versions of rank-discounted utilitarianism have been discussed and advocated under various names in both philosophy and economics, e.g. 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 decreasing function.[18]

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. We will consider the most obvious approach, simply taking equation (2) as a definition regardless of the size of X.[19] 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:

[17] 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.

[18] To connect this to the standard notation in this paper, one can alternatively write

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

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

[19] An alternative approach would be to extend to variable-populations the 'veil of ignorance' description of rank-discounting described by Buchak (see also McCarthy et al. (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).

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 non-increasing 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 non-increasing, positive function $f : \mathbb{R} \to \mathbb{R}$ that is eventually convex[20] 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 to 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. Uniquely among the axiologies we consider, GRD does *not* converge to 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 to it by any particular additive, Paretian axiology.

[20] 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.

Nonetheless, it turns out that GRD *does* converge to 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.[21] 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 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.

[21] 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 consists 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, $V_{\text{GRD}}(X) - V_{\text{GRD}}(Z) = 0.1\beta^3 - \beta^{N+3}$, again positive for N large enough; then $X \succ_{\text{GRD}} Z$, illustrating the second claim.

To sum up, many forms of egalitarianism, including many forms of rank-discounted utilitarianism, converge to interesting additive axiologies. Geometric Rank-Discounted Utilitarianism provides one counterexample, although it does converge to 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 investigate the implications of the preceding results, and especially their practical implications for morally significant real-world choices. As we have seen, how closely a given non-additive axiology agrees with its additive counterpart in some real-world choice situation depends on the size of the population that can be treated as 'background' in that choice situation. And *what* that additive counterpart will be (i.e., which version of CL or PR) depends on the average welfare of the background population, and perhaps on its entire welfare distribution. In this section, therefore, we consider the size and (to a lesser extent) the welfare of real-world background populations.

We note that nothing in this section (or the next two) shows conclusively that the background population is large enough for our limit results to be effective, but we do establish a *prima facie* case for their relevance. In §9, we will seek firmer conclusions in a stylized case.

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. Importantly, the background population is not defined in terms of individuals whose existence and welfare levels are unaffected by the choice at hand. Rather, given a choice between populations $X_1, X_2, ...X_n$, the population Z that can be treated as background with respect to that choice is defined by $Z(w) = \min_i X_i(w)$. That is, the background population consists of the minimum number of welfare subjects at each welfare level who are guaranteed to exist regardless of the agent's choice. For this Z and for each X_i, there is then a population X_i^* such that $X_i = X_i^* + Z$. The choice between $X_1, X_2, ...X_n$ 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.

Clearly, this means that different real-world choices will involve different background populations. In particular, more consequential choices (that have far-reaching effects on the overall population) allow less of the population to be treated as background, whereas choices whose effects are tightly

localized (or otherwise limited) may allow nearly the entire population to be treated as background. But we can also define a 'shared' background population for a *set* of choice situations, by considering all the overall populations that might be brought about by any *profile* of choices in those situations. Thus we can speak, for instance, of the population that is 'background' with respect to all the choices faced by present-day human agents, consisting of the minimum number of individuals at each welfare level that the overall population will contain whatever we all collectively do. This might simply be the number of individuals at each welfare level outside Earth's present future light cone.[22] Importantly, however, the size of the background population can exceed the number of 'unaffectable' individuals (e.g., those outside our future light cone). It might be, for instance, that our present choices entirely determine which particular future individuals will exist, but that whatever we do, the future population will include some minimum number of individuals at various welfare levels, in which case those minima will contribute to the background population.

6.1 Population size

We will make two claims about the size of real-world background populations, 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. This suggests that our limits results are relevant when comparing options that only affect present humans (and more generally, any choices where none of the potential foreground populations are larger than the present human population). As we will see in §9, it can also be enough to substantially increase the relative importance of avoiding existential catastrophe and ensuring the existence of a large future population, as compared to benefiting present humans, even though in this case one of the potential foreground populations (the large future population that will exist if we avoid existential catastrophe) may be much larger than the background population. Second, with much lower confidence, we will argue that real-world backgound populations *may well be* much larger (again, multiple orders of magnitude) than the entire population in our future light cone, even on the supposition that Earth-originating civilization will eventually settle much of the accessible universe and support astronomically large populations. The ground for this second claim is that,

[22] Here and below, we assume a causal decision theory, which guarantees that causally inaccessible populations can be treated as 'background'. How we can identify background populations, and how their practical significance changes, in the context of non-causal decision theories are interesting questions for future work.

if there is life elsewhere in the universe, the great majority of it is likely to be outside our future light cone. This suggests that our limit results may be relevant to *all* our real-world choices, including those that have far-reaching effects on the long-term future within our future light cone.

Let's start by establishing the first claim. The most obvious component of real-world background populations is past welfare subjects on Earth. Estimates of the number of human beings who have ever lived are on the order of 10^{11} (Kaneda and Haub, 2018), of whom only $\sim 7 \times 10^9$ are alive today. But of course *Homo sapiens* are not the only welfare subjects. At any given time in the recent past, for instance, there have also been many billions of mammals, birds, and fish being raised by humans for meat and other agricultural products. And given their very high birth/death rates, past members of these populations greatly outnumber present members.

But since human agriculture is a relatively recent phenomenon, farmed animals make only a relatively small contribution to the total background population. Wild animals make a far greater contribution. There are today, conservatively, 10^{11} mammals living in the wild, along with similar or greater numbers of birds, reptiles, and amphibians, and a significantly larger number of fish—conservatively 10^{13}, and possibly far more.[23] This is despite the significant decline in wild animal populations in recent centuries and millennia as a result of human encroachment.[24] Inferring the total number of past mammals, vertebrates, etc., from the number alive at a given time requires us to make assumptions about population birth/death rates. Unfortunately, we have not been able to find data that allow us to estimate overall birth/death rates for the wild mammal or wild vertebrate populations as a whole with any confidence. So we will simply adopt what strikes us as a very safely conservative assumption of 0.1 births/deaths per individual per year in wild animal populations (roughly corresponding to an average individual lifespan of 10 years). The actual rates are almost certainly much higher (especially for vertebrates), implying larger total past populations.

Being extremely conservative, then, we might suppose that all and only mammals are welfare subjects and that 10^{11} mammals have been alive on Earth at any given time since the K-Pg boundary event (the extinction event that killed the dinosaurs, $\sim$66 million years ago), with a population birth/death rate of 0.1 per individual per year. This gives us a background

[23] For useful surveys of evidence on present animal population sizes, see Tomasik (2019) and Bar-On et al. (2018) (especially pp. 61-4 and Table S1 in the supplementary appendix).

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

population of $\sim 6.6 \times 10^{17}$ individuals. Being a bit less conservative, we might suppose that all and only vertebrates are welfare subjects and that 10^{13} vertebrates have been alive on Earth at any time in the last 500 million years (since shortly after the Cambrian explosion), with the same population birth/death rate of 0.1 per individual per year. This gives us a background population of $\sim 5 \times 10^{20}$ individuals. It is worth noting that even the restriction to vertebrates may be objectionably conservative, excluding potential welfare subjects like crustaceans and insects.

So far, this only considers past life on Earth. Once we look beyond Earth, we get support for the stronger but more speculative claim. 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 galaxies. However, the *accessible* universe (the portion of the universe that it is possible in principle for us to reach, travelling no faster than the speed of light) is significantly smaller, containing only about 20 billion galaxies (at least according to our present best understanding of the expansion of the universe, which places limits on the accessible universe) (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 et al., 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,[25] and some models suggest that it is *vastly* larger than the observable universe.[26] But at a minimum, we should expect that more than 99% of any life elsewhere in our universe is outside our future light cone, and therefore guaranteed to belong to background rather than foreground populations. The same goes for intelligent life and advanced civilizations elsewhere in the universe. Thus, for any hypothesis about the future size of our civilization, no matter how extravagant, if we assume that there are a significant number of advanced civilizations of similar size elsewhere in the universe, then we should conclude that the large majority of those civilizations—and therefore the large majority of the universal population—are inaccessible, and background. But this is, of course, entirely speculative, and we take no stance on the existence of life or civilization elsewhere in the universe.[27]

[25] On the possibility of an infinite universe, see §10.2 below.

[26] For instance, Greene (2004) 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' (p. 285). From one such inflationary model, Page (2007) extrapolates (though without fully endorsing) a lower bound of roughly $10^{10^{10^{122}}}$ Hubble volumes.

[27] Given our current state of ignorance on this point, what effect the possibility of ex-

6.2 Welfare

Anything we say about the distribution of welfare levels in the background population will of course be enormously speculative. So although the question has important implications, we will limit ourselves to a few brief remarks.

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, so the average welfare level of the background population is very close to zero. If the capacity for positive/negative welfare scales with brain size (or related features like cortical neuron count), this would reinforce the same conclusion. It seems likely that average welfare in these populations will be negative, at least on a hedonic view of welfare (Ng, 1995; Horta, 2010). These assumptions together would imply, for instance, that AU, VV1 and VV2 converge to a version of CL with a slightly negative critical level (perhaps very similar in practice to TU).

> **Hypothesis 2** The background population mainly consists of the members of advanced alien civilizations. If, for instance, the average biosphere produces 10^{23} wild animals over its lifetime, but one in a million biospheres gives rise to an interstellar civilization that produces 10^{35} individuals on average over *its* lifetime, then the denizens of these interstellar civilizations would greatly outnumber wild animals in the universe as a whole. Under this hypothesis, given the limits of our present knowledge, all bets are off: average welfare of the background population could be very high (Ord, 2020, pp. 235–9), very low (Sotala and 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 non-trivial degree of welfare inequality in the background population—compare, for instance, the lives of a well-cared-for pet dog and a factory-farmed layer hen. Self-reported welfare levels in the contemporary human population indicate substantial inequality (see for

traterrestrial life has on the practical implications of non-additive axiologies will depend very much on how those axiologies handle uncertainty (a topic we are mostly avoiding in this paper, except in §9), and what this means for situations where we're uncertain about the size and other characteristics of the background population (a topic we are *entirely* avoiding).

instance Helliwell et al. (2019), Ch. 2), and while *contemporary* humans need not belong to the background population with respect to present-day choice situations, it seems safe to infer that there has been substantial welfare inequality in human populations in at least the recent past. For non-human animals, of course, we do not even have self-reports to rely on, and so any claims about the distribution of welfare are still more tentative. But there is, for instance, some literature on farm animal welfare that suggests significant inter-species welfare inequalities (e.g. Norwood and Lusk (2011, pp. 224–9), Browning (2020)).

That said, it could still turn out that the background population is dominated by welfare subjects who lead fairly uniform lives—e.g., by small animals who almost always experience lifetime welfare slightly below 0, or by members of alien civilizations that converge reliably on some set of values, social organization, etc., that produce enormous numbers of individuals with near-equal welfare.

7 Objection 1: Causal domain restriction

We have shown that various non-additive axiologies converge to additive axiologies in the large-background-population limit. But proponents of non-additive views might wish to avoid drawing practical conclusions from these results. After all, much of the point of being, say, an average utilitarian rather than a critical-level utilitarian is to reach the right practical conclusions in cases where AU seems more plausible than CL. That point is defeated if, in practice, AU is nearly indistinguishable from CL.

The simplest way to avoid the implications of our limit results is to claim that, for decision-making purposes, agents should simply ignore most or all of the background population. This idea can be spelled out in various ways, but it seems to us that the most principled and plausible precisification is a *causal domain restriction* (Bostrom, 2011), according to which an agent should evaluate the potential outcomes of her actions by applying the correct axiology only to those populations that might exist *in her causal future* (presumably, her future light cone).[28] Since background

[28]A causal domain restriction might be motivated by the *temporal value asymmetry*, our tendency to attach greater affective and evaluative weight to future events than to otherwise equivalent past events (Prior, 1959; Parfit, 1984, Ch. 8). It is sometimes claimed that this asymmetry characterizes only our self-regarding (and not our other-regarding) preferences (see e.g. Parfit, 1984, p. 181; Brink, 2011, pp. 378–9; Greene and Sullivan, 2015, p. 968; Dougherty, 2015, p. 3), but recent empirical studies appear to contradict this claim (Caruso et al., 2008; Greene et al., 2021). However, though the temporal value asymmetry is a clear and robust *psychological* phenomenon, it has proven notoriously difficult to come

populations of the sort described in the last section will mostly lie outside an agent's future light cone, a causal domain restriction may drastically reduce the size of the population that can be treated as background, and hence the practical significance of our limit results.

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.[29]

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 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.[30] 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

up with any normative *justification* for asymmetric evaluation of past and future events (see for instance Moller (2002), Hare (2013)).

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

[30] One general lesson of this example is that, when a group of timelike-related agents or agent-stages accept the same causal-domain-restricted non-separable axiology, an earlier agent in the group will have an incentive (i.e., will pay some welfare cost) to push axiologically significant events forward in time, into the future light cones of later agents, so that their evaluations of their options will more closely agree with hers.

feature of that normative theory, though we leave it to the reader to decide just how unwelcome.[31]

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 be *identity-affecting* with respect to many or most 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—see, e.g., Parfit (1984, Ch. 16)), but will have little if any effect on the *number* of individuals at each welfare level in that population. And this alone supplies quite a large background population—perhaps 10^{13} mammals and 10^{16} 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 to human civilization, even the 'short-termist' (bednet) intervention 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.[32]

8 Objection 2: Counting some for less than one

Another possible way to avoid the limit behaviors described in §§4–5 is to claim that not all welfare subjects make the same contribution to the 'size' of a population, as it should be measured for axiological purposes. Roughly speaking: although we should not deny *tout court* that fish are welfare subjects, perhaps, when evaluating outcomes, a typical fish should effectively count as only (say) one tenth of a welfare subject, given its cognitive and physiological simplicity. If, in a typical choice situation, the background population is predominantly made up of such simple creatures, then it

[31] This argument is essentially due to Rabinowicz (1989); see also the cases of intertemporal conflict for future-biased average utilitarianism in Hurka (1982b, pp. 118–9).

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

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

might be dramatically smaller (in the relevant sense) than it would first appear.[33]

A bit more formally, we can understand this strategy as assigning a real-valued *axiological weight* to each individual in a population, and turning populations from integer-valued to real-valued functions, where $X(w)$ now represents not the *number* of welfare subjects in X with welfare w, but the *sum of the axiological weights* of all the welfare subjects in X with welfare w. Axiological weights might be determined by factors like brain size, neuron count, or lifespan. 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.

We have three replies to this suggestion as well. First, of course, one might lodge straightforward ethical objections to axiological weights, since they seem to contradict the ideals of impartiality and equal consideration that are often seen as central to ethics in general and axiology in particular.

Second, the most natural measures by which we could assign axiological weights generate population size adjustments that, though large, still leave us with background populations significantly larger than the present human population. For instance, suppose we stick with our conservative assumption that only mammals are welfare subjects, but also weight by cortical neuron count. And, very conservatively, let's take mice as representative of non-human mammals in general. Humans have roughly 2875 times as many cortical neurons as mice (Roth and Dicke, 2005, p. 251). Normalizing our axiological weights so that present-day humans have an average weight of 1, this would mean that non-human mammals have an average weight of 3.48×10^{-4}, which would cut our estimate of the size of the mammalian background population from $\sim 6.6 \times 10^{17}$ down to $\sim 2.3 \times 10^{14}$. If we *also* weight by lifespan, and generously assume that present-day humans have an average lifespan of 100 years, then the effective mammalian background population is reduced to $\sim 2.3 \times 10^{13}$.[34] Thus, even after making a host of conservative assumptions (only counting mammals as welfare subjects, taking a conservative estimate of the number of mammals alive at a time,

[33] Thanks to Tomi Francis and Toby Ord, who each separately suggested this objection. A view along these lines is proposed by Kagan (2019); see especially his §4.5. On p. 109, Kagan suggests (by way of illustration) that mice might have a weight of 0.02, far more favorable to our case than the values we consider below.

[34] When we weight by lifespan, we can derive population size simply from the number of individuals alive at a time multiplied by time, without needing to make any assumptions about birth or death rates.

ignoring times before the K-Pg boundary event, weighting by cortical neuron count and lifespan, and taking mice as a stand-in for all non-human mammals), we are still left with a background population more than three orders of magnitude larger than the present human population.

Finally, as we have already argued, even if we entirely ignore non-humans we may still find that background populations are large relative to foreground populations in most present-day choice situations. Past humans outnumber present humans by more than an order of magnitude (as we saw in §6). And it seems plausible that the large majority even of the present and near-future human population is approximately background in most choice situations (as we argued at the end of §7). Thus, even if we *both* severely deprioritize or ignore non-humans *and* adopt a causal domain restriction, we might *still* find that background populations are usually large relative to foreground populations.

9 The importance of existential catastrophe

Taking stock: in §§4–5, we showed that various non-additive axiologies converge to additive axiologies in the presence of large enough background populations. In §6, we argued that real-world background populations are quite large—at the very least, multiple orders of magnitude larger than the affectable portion of the present and near-future population in most choice situations. And in §§7–8, we resisted two strategies for deflating the size of real-world background populations.

If real-world background populations are indeed large relative to foreground populations, this provides some *prima facie* reason to believe that our limit results may be practically significant. That is, it at least intuitively suggests that for many plausible non-additive views in the families we have considered, what is true in the limit will be true in practice, namely, that they agree closely with their additive counterparts. More generally, it suggests that even if we don't accept (additive) separability as a fundamental axiological principle, it may nevertheless be a useful heuristic for real-world decision-making purposes—i.e., that arguments in practical ethics that rely on separability assumptions may still succeed in practice.

But in this section we will give a more particular, concrete illustration of the practical import of the preceding arguments. 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. On additive views, the amount of present welfare we should be

willing to sacrifice to ensure the existence of a future population X (assuming X has positive value) scales linearly with $|X|$. And since the potential future population of human-originating civilization is astronomically larger than the present human population, as long as the average welfare of that future population would be significantly above the critical level, we should be willing to accept very large sacrifices for the present generation to ensure its existence. But non-additive views need not endorse this sort of reasoning—in particular, AU and other similar views do not.

We will therefore consider 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.[35] An 'existential catastrophe', for our purposes, is any near-future event that would drastically reduce the future population size of human-originating civilization.[36] For expository purposes, we focus on the case where the future generations that will exist if we avoid existential catastrophe have higher average welfare than the background population, so that AU assigns positive value to avoiding existential catastrophe, at least in the large-background-population limit. But most of what we say about the value of avoiding existential catastrophe on this assumption also applies, *mutatis mutandis*, to the *dis*value of avoiding existential catastrophe on the opposite assumption that the potential future population has lower average welfare than the background population.

9.1 Setup

We want to know how AU balances the competing objectives of avoiding existential catastrophe and improving the welfare of the affectable pre-catastrophe population (which, for simplicity, we will hereafter call 'the current generation'), in the presence of a background population. In particular, we want to know how the relationship between the size of the potential future population and the importance of avoiding existential catastrophe is mediated by the size of the background population, and whether AU

[35] For example, while totalist two-factor egalitarianism in 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.

[36] This includes, but is not limited to, premature human extinction. For instance, an event that prevented humanity from ever settling the stars, while allowing us to survive for a very long time on Earth, could be an existential catastrophe in our sense. It is also importantly distinct from the usual concept of 'existential catastrophe' in the philosophical literature, which is roughly 'any event that would permanently curtail humanity's long-term potential for value' (see for instance Bostrom, 2013, p. 15; Ord, 2020, p. 37).

takes on the scale-sensitivity of its limiting theory CL for realistically-sized background populations, or only in some unrealistically far-off limit. To formalize these questions, let C^+ represent the current generation as it will be if we prioritize its welfare over avoiding existential catastrophe. Let C^- denote the current generation as it will be if we instead prioritize avoiding existential catastrophe. Thus $\overline{C^+} > \overline{C^-}$, but we assume that $|C^+| = |C^-|$ (and will therefore designate this quantity simply $|C|$).[37] Let F denote the future population that will exist only if we avoid existential catastrophe. And suppose there is a background population Z, which includes past terrestrial welfare subjects, perhaps distant aliens, and perhaps unaffectable present/future welfare subjects like wild animals.

Now, we have so far focused entirely on axiologies understood as rankings of populations, and ignored the question of how these axiologies should be extended to rank *probability distributions* over populations. But in thinking about the practical importance of existential catastrophe, the question of risk is unavoidable. In practice, we do not face choices that lead to certainty of existential catastrophe on the one hand and certainty of survival (i.e., non-catastrophe) on the other. Rather, we face choices where different options carry slightly different probabilities of existential catastrophe. Thus, rather than asking, for instance, 'How much sacrifice should we be willing to impose on the current generation to exchange certainty of catastrophe for certainty of survival?', a better question is 'What reduction in the *probability* of existential catastrophe would compensate a given reduction in the welfare of the current generation?' To address this question, we must adopt some rule for the ranking of risky prospects, and we will therefore assume that an average utilitarian should respond to risk by *maximizing expected average welfare*. Call this extended theory $\mathbb{E}$AU. We have, unfortunately, no compelling argument that $\mathbb{E}$AU is the best extension of AU to cases of risk. But we have to assume *something*, maximizing expected average welfare seems like a natural default, and there no alternative decision rule for AU (or for other non-additive axiologies) that has achieved anything like widespread acceptance.[38]

[37] Since any existential catastrophe would occur some way into the future, we can have some influence on pre-catastrophe population sizes. But even in cases where this influence is substantial, the assumption that $|C^+| = |C^-|$ is mostly harmless, since we can interpret $|C|$ as the size of the non-background population in the outcome where existential catastrophe occurs (in which outcome, note, the non-background population is smallest), and, in any other outcome, we designate the first $|C|$ members of the non-background population as the members of C^+ or C^-.

[38] (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. Although this view can behave quite differently from $\mathbb{E}$AU in general,

This rule in hand, we consider a choice between two options, one of which improves the welfare of the current generation from $\overline{C^-}$ to $\overline{C^+}$ and the other of which increases the probability of survival (i.e., avoiding existential catastrophe) by an increment of Δp from a baseline of p. That is, the options to be compared are:

$\boldsymbol{O_1}$ $F + C^+ + Z$ with probability p, $C^+ + Z$ otherwise

$\boldsymbol{O_2}$ $F + C^- + Z$ with probability $p + \Delta p$, $C^+ + Z$ otherwise

9.2 Qualitative analysis

In the rest of the section, we consider the question: How large a reduction in the probability of existential catastrophe is needed to offset a reduction in the average welfare of the current generation from $\overline{C^+}$ to $\overline{C^-}$? That is, we treat all other features of the choice situation as fixed, and take as our dependent variable of interest Δp^*, the value of Δp at which we are indifferent between O_1 and O_2. We can treat the reciprocal of Δp^* (i.e., $(\Delta p^*)^{-1}$) as one measure of 'the importance of existential catastrophe'—the larger $(\Delta p^*)^{-1}$ becomes, the more we should be willing to trade even large gains in the welfare of the current generation for even small reductions in the probability of existential catastrophe.

Let's first consider the value of Δp^* according to the axiology $\mathrm{CL}_{\overline{Z}}$, to which AU converges in the large-background-population limit. More specifically, consider $\mathbb{E}\mathrm{CL}_{\overline{Z}}$, which ranks risky prospects by their expected critical-level sums. It is straightforward to show that, according to $\mathbb{E}\mathrm{CL}_{\overline{Z}}$,

$$\Delta p^* = \frac{|C|}{|F|} \frac{\overline{C^+} - \overline{C^-}}{\overline{F} - \overline{Z}}. \tag{3}$$

For our purposes, the most important feature of this equation is that Δp^* is inversely related to $|F|$, meaning that the importance of existential catastrophe (as measured by $(\Delta p^*)^{-1}$) scales linearly with $|F|$.

---

the main qualitative conclusions described below still hold: rough independence from population size in Case 1, dependence on $|C|/|Z|$ in Case 2, and dependence on $|C|/|F|$ in Case 3. If we wished to avoid questions of risk entirely, we could measure the importance of avoiding existential catastrophe according to AU in a risk-free context and ask how this changes with the size of the background population. In particular, we could (i) hold $\overline{C^-}$ fixed and find the value of $\overline{C^+}$ at which $F + C^- + Z \sim_{\mathrm{AU}} C^+ + Z$, or (ii) hold $\overline{C^+}$ fixed and find the value of $\overline{C^-}$ at which $F + C^- + Z \sim_{\mathrm{AU}} C^+ + Z$. With some minor complications, these approaches yield qualitatively similar conclusions to those we reach in the rest of this section, with respect to how the importance of avoiding existential catastrophe depends on $|Z|$. But we omit these analyses for the sake of concision.

By contrast, $\mathbb{E}AU$ is indifferent between O_1 and O_2 when

$$\Delta p(\overline{F+C^-+Z}-\overline{C^-+Z}) = p(\overline{F+C^++Z}-\overline{F+C^-+Z}) + (1-p)(\overline{C^++Z}-\overline{C^-+Z}).$$

Solving for Δp, one finds

$$\Delta p^* = \frac{(\overline{C^+}-\overline{C^-})(p(|C|^2+|C||Z|)+(1-p)(|F||C|+|C|^2+|C||Z|))}{|F||C|(\overline{F}-\overline{C^-})+|F||Z|(\overline{F}-\overline{Z})} \quad (4)$$

This expression is unattractive, but informative—it lets us see, in broad strokes, how the importance of avoiding existential catastrophe according to $\mathbb{E}AU$ changes with $|Z|$. Consider the following three cases, where in each case we assume that the baseline probability of survival p is intermediate, i.e., neither much larger nor much smaller than $(1-p)$:

Case 1: $|\mathbf{F}| \gg |\mathbf{C}| \gg |\mathbf{Z}|$. In this case, where the background population is comparatively small (or non-existent), (4) becomes approximately $(1-p)\frac{\overline{C^+}-\overline{C^-}}{\overline{F}-\overline{C^-}}$. [39] There are three things to observe about this formula. First, because $|F| \gg |C|$, improving C^- to C^+ has a much greater impact on average welfare in the state where existential catastrophe occurs and F does not exist. This is why the p term in (4) drops out. Second, the term $\overline{F}-\overline{C^-}$ in the denominator indicates that $\overline{C^-}$ acts like an 'effective critical level', in the sense that we will only ever prefer avoiding existential catastrophe to benefiting the current generation if $\overline{F} > \overline{C^-}$. Third, and most importantly for our purposes, the importance of avoiding existential catastrophe as measured by Δp^* is approximately independent of $|F|$, so that the astronomical size of the future population does not make it astronomically important that we ensure its existence.

Case 2: $|\mathbf{F}| \gg |\mathbf{Z}| \gg |\mathbf{C}|$. In this case, where Z is intermediate in size between F and C, (4) becomes approximately $(1-p)\frac{|C|}{|Z|}\frac{\overline{C^+}-\overline{C^-}}{\overline{F}-\overline{Z}}$. Because F is still much larger than Z or C, improving C^- to C^+ still has a much greater impact on average welfare when F does not exist, so the p term still drops out. But now, because Z is the largest part of the pre-existing population, the 'effective critical level' against which F is evaluated becomes $\overline{Z}$ rather than $\overline{C^-}$. Finally, and most importantly for our purposes, Δp^* is now proportionate to $\frac{|C|}{|Z|}$. Since $|Z| \gg |C|$, this means that Δp^* will be very small (assuming $\frac{\overline{C^+}-\overline{C^-}}{\overline{F}-\overline{Z}}$ is not very large), but it does

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

not yet scale inversely with $|F|$—rather, the importance of existential catastrophe is limited by the size of the background population.

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, (4) becomes approximately $\frac{|C|}{|F|}\frac{\overline{C^+}-\overline{C^-}}{\overline{F}-\overline{Z}}$—exactly the formula for Δp^* given by $\mathbb{E}\mathrm{CL}_{\overline{Z}}$. That is, in this case, we should expect $\mathbb{E}$AU and $\mathbb{E}\mathrm{CL}_{\overline{Z}}$ to agree, to a good approximation, about the importance of existential catastrophe.

The most basic qualitative point to take away from this analysis is that Δp^* decreases toward zero as we increase *both* $|F|$ and $|Z|$. The fact that possible future and actual background populations are both likely to be extremely large suggests that Δp^* is likely to be small (thus favouring extinction-avoidance) for a robust range of the other parameters.

9.3 A numerical illustration

So far, our analysis has remained qualitative; we'll now put in some numbers, with the purpose of illustrating two things: first, the practical point that even AU will give great weight to avoiding existential catastrophe, for some reasonable and even conservative estimates of the background population and other parameters; and second, the more theoretical point that AU converges to CL with high precision, given these same estimates.

For the sizes of the foreground populations, let's suppose that $|C| = 10^{10}$ (a realistic estimate of the size of the present and near-future human population) and $|F| = 10^{17}$ (a fairly conservative estimate of the potential size of the future human-originating population, if we avoid existential catastrophe, obtained by assuming 10^{10} individuals per century for the next billion years). For $|Z|$, we will consider three values: $|Z| = 0$ (i.e., the absence of any background population), $|Z| = 10^{13}$ (a rounding-down of our most conservative estimate of the number of past mammals, weighted by lifespan and cortical neuron count, from §8), and $|Z| = 10^3 \times |F| = 10^{20}$ (arrived at by assuming that the universe contains 1000 other advanced civilizations, of the same scale that our civilization will achieve if we avoid existential catastrophe). $|Z| = 10^{13}$ represents a conservative version of our high-confidence claim from §6 that real-world background populations are at least orders of magnitude larger than the present human population. $|Z| = 10^{20}$ corresponds to our more speculative claim that real-world background populations may well be orders of magnitude larger than the whole population of our future light cone.

Axiology	$\lvert\mathbf{Z}\rvert$	Approximation	$\Delta\mathbf{p}^*$
$\mathbb{E}$AU	0	$(1-p)\frac{\overline{C^+}-\overline{C^-}}{\overline{F}-\overline{C^-}} = 0.25$	0.25000005
$\mathbb{E}$AU	10^{13}	$(1-p)\frac{\lvert C\rvert}{\lvert Z\rvert}\frac{\overline{C^+}-\overline{C^-}}{\overline{F}-\overline{Z}} = 1.25\times10^{-4}$	$\sim 1.2496\times10^{-4}$
$\mathbb{E}$AU	10^{20}	$\frac{\lvert C\rvert}{\lvert F\rvert}\frac{\overline{C^+}-\overline{C^-}}{\overline{F}-\overline{Z}} = 2.5\times10^{-8}$	$\sim 2.50125\times10^{-8}$
$\mathbb{E}\mathrm{CL}_{\overline{Z}}$	—	—	2.5×10^{-8}

TABLE 1: The importance of avoiding existential catastrophe, as measured by Δp^*, according to $\mathbb{E}$AU for different background population sizes and $\mathbb{E}\mathrm{CL}_{\overline{Z}}$, with $\overline{F}=2$, $|F|=10^{17}$, $\overline{C}=1.5$, $\overline{C'}=1$, $|C|=|C'|=10^{10}$, $\overline{Z}=0$, $p=0.5$, and $|Z|$ as specified in each row. The third column gives the approximations used in the previous subsection for comparison.

In terms of average welfare, we have much less to go on. For simplicity let's assume that $\overline{F}=2$ (we can choose the units of welfare so that this corresponds to very good but generally normal human lives) and $\overline{Z}=0$ (plausible for the case where Z consists mainly of wild animals, somewhat less plausible for the case where it consists mainly of the member of other advanced civilizations). And let's assume that $\overline{C^-}=1$ and $\overline{C^+}=1.5$.

Finally, we assume that $p=0.5$.

Table 1 gives the values of Δp^* according to $\mathbb{E}$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 or fourth significant figure. In particular, in the third case, where $|Z|\gg|F|\gg|C|$, $\mathbb{E}$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.

In summary, the preceding analysis suggests the following conclusions: (1) When the background population is small or non-existent, the importance of avoiding existential catastrophe according to $\mathbb{E}$AU is approximately independent of population size, depending only on the average welfares of the potential foreground populations, and is therefore unlikely to be astronomically large. (2) When the background population 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 astronomically large, while still falling well short of its importance according to $\mathbb{E}\mathrm{CL}_{\overline{Z}}$. (3) Finally, if the background population is much larger even than the potential future

population (as it would be, for instance, if it includes many advanced civilizations elsewhere in the universe), $\mathbb{E}$AU agrees closely with $\mathbb{E}\text{CL}_{\overline{Z}}$ about the importance of avoiding existential catastrophe, treating it as approximately linear in $|F|$.

In this very specific context, therefore, we can now say how large the background population needs to be for large-background-population limiting behavior to 'kick in': $\mathbb{E}$AU closely approximates $\mathbb{E}\text{CL}_{\overline{Z}}$ only when $|Z| \gg |F|$. But it behaves in important ways like $\mathbb{E}\text{CL}_{\overline{Z}}$ as long as $|Z| \gg |C|$, both in that it may assign great importance to avoiding existential catastrophes, and in that the effective critical level that determines whether that importance is positive or negative is approximately $\overline{Z}$. This lends significance to our conclusion in §6 that real-world background populations are much larger than the current generation (i.e., the affectable present and near-future population), whether or not they are large relative to the potential future population as a whole. The former fact alone is enough to have a significant effect on how $\mathbb{E}$AU evaluates existential catastrophes in practice.

10 Other implications

We conclude by briefly surveying three other interesting implications of our limit results and, more generally, of the influence of background populations on the verdicts of non-separable axiologies.

10.1 Repugnant Addition

The Repugnant Conclusion, recall, is the conclusion (implied by TU among other axiologies) that for any positive welfare levels $l_1 < l_2$ and any number n, there is a population where everyone has welfare l_1 that is better than a population of n individuals all with welfare l_2. One of the motivations for population axiologies with an 'averagist' flavor (like AU, VV1, VV2, and QAA) is to avoid the Repugnant Conclusion. But the results in §§4–5 imply that, although they avoid the Repugnant Conclusion as stated above, these views cannot avoid the closely related phenomenon of 'Repugnant Addition': for any positive welfare levels $l_1 < l_2$ and any number n, if Y consists of n individuals all with welfare l_2, there is some population X in which everyone has welfare l_1 and some population Z such that $X + Z$ is better than $Y + Z$. As per the results in §4, AU/VV1/VV2 support Repugnant Addition with respect to a large enough background population Z with $\overline{Z} \leq 0$ (and indeed, when $\overline{Z} < 0$, they support the much more repugnant conclusion that, for

any population Y in which everyone has positive welfare, there is a larger population X in which everyone has negative welfare such that $X+Z$ is better than $Y+Z$).

The difficulty of avoiding Repugnant Addition has been noticed independently by Budolfson and Spears (ms), who provide a thorough exploration of the phenomenon covering a broader range of axiologies than we have considered here. So rather than saying any more about this implication, we direct the reader to their results.

10.2 Infinite ethics

A long-standing and unresolved challenge for axiology is how to extend axiologies from finite to infinite contexts.[40] Most of the extant proposals for ranking infinite worlds, in both philosophy and economics, aim to extend total utilitarianism.[41] However, these proposals can easily be adapted to extend other additive axiologies. For instance, a simple extension of total utilitarianism (suggested in Lauwers and Vallentyne (2004)) simply compares any two populations by summing the differences in welfare between the two populations for each individual, treating an individual who doesn't exist in a population as having welfare 0.[42] This axiological criterion can easily be adapted to a critical-level or prioritarian theory by simply replacing welfare with some increasing function of welfare.

It is much less clear, however, how to extend non-additive theories to the infinite context, and there has so far been little if any discussion of this question. Our limit results, however, suggest a partial answer: when comparing two infinite populations that differ only finitely, we are quite literally in (and not merely approaching) the large-background-population limit. So it is natural to think that a non-additive axiology $\mathscr{A}$ that has an additive counterpart $\mathscr{A}'$ should agree exactly with that additive counterpart in the infinite context. For instance, if we are average utilitarians and we live in an infinite world, but we can only affect a finite part of that infinite world, then we should simply compare the possible outcomes of our choices by the appropriate infinite generalization of critical-level utilitarianism, where the critical level is the average welfare level in the background population.

[40] For surveys of the difficulties of infinite axiology, see for instance Asheim (2010), Bostrom (2011), and Ch. 1 of Askell (2018).

[41] See, for instance, Atsumi (1965), Diamond (1965), Von Weizsäcker (1965), Vallentyne and Kagan (1997), Lauwers and Vallentyne (2004), Bostrom (2011), Arntzenius (2014), Jonsson and Voorneveld (2018), and **?**, among many others.

[42] Formally, $X \succcurlyeq Y$ if and only if $\sum_{p_i \in X \cup Y} w_{\mathbf{x}}(p_i) - w_{\mathbf{y}}(p_i)$ converges unconditionally to a value ≥ 0, where for any $p_i \notin X$, $w_{\mathbf{x}}(p_i) = 0$ (and likewise for Y).

This suggestion is well-defined only if we have a well-defined notion of *relative frequency* for infinite worlds—specifically, the relative frequency of different welfare levels in an infinite population, which lets us make sense of further notions like a *welfare distribution* and *average welfare*. A natural suggestion here (advocated, for instance, by Knobe et al. (2006)) is to use the *limiting* relative frequency in uniformly expanding spatiotemporal regions, providing that this limit exists and is the same for all starting locations. There is plenty of debate to be had about this proposal, but this is not the place for that debate. At any rate, it seems plausible (though far from indisputable) that there should be *some* way of making sense of the relative frequencies of particular welfare levels in an infinite population.

10.3 Opportunities for manipulation

The results in §§4–5 have one other interesting implication: they suggest a way in which agents who accept non-separable axiologies can be *manipulated*. 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 enormous number of welfare subjects with welfare very close to zero—for instance, breeding quintillions of very small, short-lived animals with mostly bland experiences—and send us proof 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. If they could create for us a large enough background population with average welfare sufficiently close to zero, the Andromedans could move us arbitrarily close to *de facto* total utilitarianism.

It's not obvious whether such a strategy would be efficient, but it might be, if creating small, short-lived welfare subjects with bland experiences (and transmitting the necessary proof of their existence) is sufficiently cheap. Since the cost of creating a welfare subject with welfare x presumably increases with $|x|$ (and plausibly increases at a super-linear rate), it might well make sense for the Andromedans to devote some of their resources to this manipulation strategy rather than spending all their resources directly on creating welfare subjects with high welfare.

As the preceding results demonstrate, this kind of manipulability is not unique to average utilitarians, but applies also to agents who accept variable-

value or non-separable egalitarian views.[43] Moreover, the potential for manipulation is not symmetric: since the Andromedans accept a separable axiology, what they choose to do in their galaxy will not be affected by their beliefs about what we are doing in ours (except in the ordinary ways, involving potential causal interactions between our galaxies).

Diverting though these speculations might be, the real-world opportunities for this sort of axiological manipulation may be quite limited. Setting aside the likelihood of nearby planets or galaxies being monopolized by proponents of rival axiologies, if there is a large enough pre-existing background population in the universe as a whole (say, outside the region accessible either to us or to the Andromedans), then it may be very hard for the Andromedans to have any significant impact on the size or welfare distribution of the background population as a whole. This might be welcome news for them: if the average welfare of the background population is already close to zero, then they will get what they want from us averagists, without having to work for it. But if the average welfare in the background population is non-zero, then we may not behave quite as the Andromedans would most prefer.

This illustrates a general point: the preceding arguments are not necessarily good news for total utilitarians, or for proponents of any other separable axiology in particular. In the presence of large background populations, non-separable axiologies can converge with a wide range of separable counterparts, which disagree among themselves about how to rank populations and how to act for the best. So although large background populations generate *some* convergence among axiologies on particular practical conclusions, axiological disputes remain practically significant.

11 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

[43] But manipulating egalitarians may be more expensive, if it requires creating beings with a wide distribution of welfare levels. Likewise, agents who accept a critical-level view other than TU may find it more expensive to manipulate in this way, since they may need to create welfare subjects at or near what they regard as the critical level—unless, for instance, creating welfare subjects with welfare close to zero can reduce the average welfare of a pre-existing background population toward that critical level.

may have important practical implications for tradeoffs between avoiding existential catastrophe and benefiting 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 rely 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 §9.

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.

Convergence on *S*
Axiology $\mathscr{A}$ converges to $\mathscr{A}'$, 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

$$X+Z \succ_{\mathscr{A}'} Y+Z \implies X+Z \succ_{\mathscr{A}} Y+Z.$$

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

Theorem 6. BRD *converges 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)

1. 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 $X_k \succ Y_k$.
2. If X and Y differ only in that Y has additional individuals at welfare level c, then X and Y are equally good.[44]

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.[45]

To state the theorem, fix a set $W \subset \mathscr{W}$ of welfare levels. Say that a population X is *supported* on W 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 $W \subset \mathscr{W}$ be any set of welfare levels, and D a distribution that covers W.* GRD *converges to* 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.*

B Proofs

Recall that $\mathscr{W}$ is the set of welfare levels, and $\mathscr{P}$ consists of all non-zero, finitely supported functions $\mathscr{W} \to \mathbb{Z}_+$. By a *type* of populations we mean a set $T \subset \mathscr{W}$ 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.

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

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

[44] 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.

[45] See McCarthy et al. (2020, Example 2.7) for details in the constant-population-size case.

as Z ranges over populations of some type T. If the axiology with value function V^s is separable, then the axiology with value function V converges to it, relative to background populations of type T.

Proof. Let Z be a background population of type T. Suppose that $V^s(X+Z) > V^s(Y+Z)$. Given that the corresponding axiology is separable, we must have $V^s(X) > V^s(Y)$. Then, if $|Z|$ is large enough,

$$\big(V(X+Z)-V(Z)\big)s(|Z|) > \big(V(Y+Z)-V(Z)\big)s(|Z|),$$

whence, rearranging, $V(X+Z) > V(Y+Z)$. □

Theorem 1. *Average utilitarianism converges 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 $V_{\mathrm{CL}_c}(X) > V_{\mathrm{CL}_c}(Y) \implies V_{\mathrm{AU}}(X+Z) > V_{\mathrm{AU}}(Y+Z)$.

Proof. In this case, a brief calculation shows

$$V_{\mathrm{AU}}(X+Z) - V_{\mathrm{AU}}(Z) = \frac{(\overline{X}-\overline{Z})|X|}{|X|+|Z|} = \frac{V_{\mathrm{CL}_c}(X)}{|X|+|Z|}. \tag{5}$$

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 (5), 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,

$$|Z|(V_{\mathrm{CL}_c}(X) - V_{\mathrm{CL}_c}(Y)) > |X|V_{\mathrm{CL}_c}(Y) - |Y|V_{\mathrm{CL}_c}(X). \tag{6}$$

Given that $V_{\mathrm{CL}_c}(X) - V_{\mathrm{CL}_c}(Y) > 0$, the desired conclusion (6) follows from (1). □

Theorem 2. *Variable value views converge 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

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

We now apply two lemmas, proved below.

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

Lemma 3. *We have* $\big(f(\overline{X+Z}) - f(\overline{Z})\big)|Z| \to f'(c)V_{\mathrm{CL}_c}(X)$ *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

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

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

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

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[46]

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

Cross-multiplying,

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

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. □

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

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 $\overline{X+Z}$ 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 (5),

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

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. First, note that $V_{\mathrm{VV1}}(X)$ has the same sign as $\overline{X}$. So if $\overline{X} \geq 0 \geq \overline{Y}$, then it is automatic that $V_{\mathrm{VV1}}(X) > V_{\mathrm{VV1}}(Y)$. (The condition that $X \succ_{\mathrm{TU}} Y$ and $X \succ_{\mathrm{AU}} Y$ excludes the case where $\overline{X} = 0 = \overline{Y}$.) Thus it remains to consider the case when $\overline{X}$ and $\overline{Y}$ are both positive or both negative.

First suppose they are positive. If $|X| \geq |Y|$, then, since g is increasing and $\overline{X} > \overline{Y}$, $V_{\mathrm{VV1}}(X) = \overline{X} g(|X|) > \overline{Y} g(|Y|) = V_{\mathrm{VV1}}(Y)$, as required. If, instead, $|Y| > |X|$, then we have

$$\frac{V_{\mathrm{VV1}}(X)}{V_{\mathrm{VV1}}(Y)} = \frac{\overline{X} g(|X|)}{\overline{Y} g(|Y|)} \geq \frac{\overline{X}|X|}{\overline{Y}|Y|} > 1$$

and therefore $V_{\mathrm{VV1}}(X) > V_{\mathrm{VV1}}(Y)$. Here, the first inequality uses the concavity of g, and the second the fact that $\mathrm{Tot}(X) > \mathrm{Tot}(Y) > 0$.

The case where $\overline{X}$ and $\overline{Y}$ are negative is similar, with careful attention to signs. □

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 $\mathscr{A}$ represented by V converges to an additive axiology relative to background populations with any given distribution D, with weighting function*[47]

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

[47] Here $1_w \in \mathscr{P}$ 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 $\mathscr{W} \to \mathbb{R}_+$, such as, in particular, D and $D + t1_w$.

If the Pareto principle holds with respect to $\mathscr{A}$, then f is weakly increasing, and if Pigou-Dalton transfers are weak improvements, then f is weakly 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 $\mathscr{P}_{\mathbb{R}}$ be the set of finitely-supported, non-zero functions $\mathscr{W} \to \mathbb{R}_+$. Let $\mathscr{D} \subset \mathscr{P}_{\mathbb{R}}$ be the subset of distributions, i.e. those functions that sum to 1. The first part of the requirement is that there is a function $\tilde{I} : \mathscr{D} \to \mathbb{R}$ 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 $\mathscr{P}_{\mathbb{R}}$ that is scale-invariant, i.e. $I(nX) = I(X)$ for all reals $n > 0$ and all $X \in \mathscr{P}_{\mathbb{R}}$. The second part of the requirement is that I, so extended, is differentiable, in the following sense:[48] for all $P, Q \in \mathscr{P}_{\mathbb{R}}$, the limit

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

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 $\mathscr{P}$ but on the larger set $\mathscr{P}_{\mathbb{R}}$ (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 additive in X and therefore represents an additive axiology $\mathscr{A}'$. More specifically, for

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

each welfare level w let 1_w be a population with one person at level w. We then have

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

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

$$f(w) = w - \partial_{1_w} I(D) - I(D). \tag{7}$$

Similarly, for averagist egalitarianism,

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

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 $\mathscr{A}$, 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 $\mathscr{A}'$, and it follows that f is weakly concave. □

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

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

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

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

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

In particular, MD is differentiable and Theorem 3 applies. We know from equation (7) in the proof of Theorem 3 that MDT converges to the additive axiology $\mathscr{A}'$ with weighting function

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

□

Theorem 5. QAA *converges to* PR*, relative to background populations with a given distribution* D*. Specifically,* QAA_g *converges to* 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 to prioritarianism with a priority weighting function

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

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

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

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 $Z_{\leq w} \to \infty$. We see therefore that each summand in the first double-sum tends to Lw. The first double sum then converges to $\sum_{w \in \mathscr{W}} X(w)Lw = L\,\mathrm{Tot}(X)$. 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 converges 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

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

Moreover, the number of terms, $Z(w)$, 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 $W \subset \mathscr{W}$ be any set of welfare levels, and D a distribution that covers W.* GRD *converges to* 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 $X \succ_{\mathrm{CLL}} Y$. Let Z be a population with distribution D, so $Z = nD$ for some $n > 0$. We have to show that $X + Z \succ_{\mathrm{GRD}} Y + Z$ 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 $X \succ_{\mathrm{CLL}} Y$ implies that $X' \succ_{\mathrm{CLL}} Y'$. According to the first condition chracterizing CLL, we have $X'_k > Y'_k$ for the first k such that $X'_k \neq Y'_k$. This shows that $Y'_k < c$, so that in fact $Y'_k = Y_k$. For brevity define $w \coloneqq Y_k$.

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

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

Apply to this the value function $V = V_{\mathrm{GRD}}$:

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

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

$$\frac{V(X+Z)-V(Y+Z)}{\beta^{|X_-+Z_-+Z_w|}} = V(X_w)-V(Y_w)+(\beta^{|X_w|}-\beta^{|Y_w|})V(Z_0)+R \qquad (8)$$

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 (8) 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 (8), we use the standard fact that $\sum_{i=1}^{m}\beta^i = (1-\beta^m)\frac{\beta}{1-\beta}$. Since $V(X_w) = \sum_{i=1}^{|X_w|}\beta^i w$, 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 (8) 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, if v' is the lowest welfare level greater than w occurring in D, then $v' \in (w,v)$ and $\lim_{n\to\infty} V(Z_0) = \frac{\beta v'}{1-\beta} > \frac{\beta w}{1-\beta}$.

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 $D_0 \neq 0$. There are two cases. First, if $w, v \in W$, then D involves a welfare level between them, because it covers W. Otherwise, $w \in W$ but $v = c+1$. Then D involves $c \in (w,v)$. Either way, D involves some welfare level in (w,v), so $D_0 \neq 0$. □

References

Adler, M. (2009). Future generations: A prioritarian view. *George Washington Law Review 77*, 1478–1520.

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

Arneson, R. J. (2000). Luck egalitarianism and prioritarianism. *Ethics 110*(2), 339–349.

Arntzenius, F. (2014). Utilitarianism, decision theory and eternity. *Philosophical Perspectives 28*(1), 31–58.

Arrhenius, G. (2000). An impossibility theorem for welfarist axiologies. *Economics and Philosophy 16*(2), 247–266.

Asheim, G. B. (2010). Intergenerational equity. *Annual Review of Economics 2*(1), 197–222.

Asheim, G. B. and S. Zuber (2014). Escaping the repugnant conclusion: Rank-discounted utilitarianism with variable population. *Theoretical Economics 9*(3), 629–650.

Askell, A. (2018). *Pareto Principles in Infinite Ethics*. Ph. D. thesis, New York University.

Atsumi, H. (1965). Neoclassical growth and the efficient program of capital accumulation. *The Review of Economic Studies 32*(2), 127–136.

Bar-On, Y. M., R. Phillips, and R. Milo (2018). The biomass distribution on Earth. *Proceedings of the National Academy of Sciences 115*(25), 6506–6511.

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

Blackorby, C., W. Bossert, and D. Donaldson (1997). Critical-level utilitarianism and the population-ethics dilemma. *Economics and Philosophy 13*(2), 197–230.

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

Bostrom, N. (2003). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas 15*(3), 308–314.

Bostrom, N. (2011). Infinite ethics. *Analysis and Metaphysics 10*, 9–59.

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy 4*(1), 15–31.

Brink, D. O. (2011). Prospects for temporal neutrality. In C. Callender (Ed.), *The Oxford Handbook of Philosophy of Time*. Oxford: Oxford University Press.

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

Broome, J. (1997). Is incommensurability vagueness? In R. Chang (Ed.), *Incommensurability, Incomparability and Practical Reason*. Harvard University Press.

Browning, H. (2020). *If I Could Talk to the Animals: Measuring Animal Welfare*. Ph. D. thesis, Australian National University.

Buchak, L. (2017). Taking risks behind the veil of ignorance. *Ethics 127*(3), 610–644.

Budolfson, M. and D. Spears (2018). Why the Repugnant Conclusion is inescapable. Unpublished manuscript, December 2018. URL: `https://scholar.princeton.edu/sites/default/files/cfi/files/budolfson_spears_2018_repugnant_cfi.pdf`.

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

Caruso, E., D. Gilbert, and T. Wilson (2008). A wrinkle in time: Asymmetric valuation of past and future events. *Psychological Science 19*(8), 796–801.

Crisp, R. (2003). Equality, priority, and compassion. *Ethics 113*(4), 745–763.

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

Diamond, P. A. (1965). The evaluation of infinite utility streams. *Econometrica: Journal of the Econometric Society 33*(1), 170–177.

Dougherty, T. (2015). Future-bias and practical reason. *Philosophers' Imprint 15*(30), 1–16.

Fishburn, P. C. (1970). Intransitive indifference in preference theory: A survey. *Operations Research 18*(2), 207–228.

Fleurbaey, M. (2010). Assessing risky social situations. *Journal of Political Economy 118*, 649–80.

Frankfurt, H. (1987). Equality as a moral ideal. *Ethics 98*(1), 21–43.

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

Greene, P., A. J. Latham, K. Miller, and J. Norton (2021). Hedonic and non-hedonic bias toward the future. *Australasian Journal of Philosophy 99*(1), 148–163.

Greene, P. and M. Sullivan (2015). Against time bias. *Ethics 125*(4), 947–970.

Gustafsson, J. E. (2020). Population axiology and the possibility of a fourth category of absolute value. *Economics and Philosophy 36*(1), 81–110.

Gustafsson, J. E. (forthcoming). Our intuitive grasp of the repugnant conclusion. In G. Arrhenius, K. Bykvist, and T. Campbell (Eds.), *The Oxford Handbook of Population Ethics*. Oxford: Oxford University Press. URL: `https://johanegustafsson.net/papers/our-intuitive-grasp-of-the-repugnant-conclusion.pdf`.

Hardin, G. (1968). The tragedy of the commons. *Science 162*(3859), 1243–1248.

Hare, C. (2013). Time – The Emotional Asymmetry. In A. Bardon and H. Dyke (Eds.), *A Companion to the Philosophy of Time*, pp. 507–520. Wiley-Blackwell.

Harsanyi, J. C. (1977). Morality and the theory of rational behavior. *Social Research 44*(4), 623–656.

Helliwell, J. F., R. Layard, and J. D. Sachs (2019). *World Happiness Report 2019*. New York: Sustainable Development Solutions Network.

Horta, O. (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, J. L. (1987). The diminishing marginal value of happy people. *Philosophical Studies 51*(1), 123–137.

Hurka, T. (1982a). Average utilitarianisms. *Analysis 42*(2), 65–69.

Hurka, T. (1982b). More average utilitarianisms. *Analysis 42*(3), 115–119.

Hurka, T. (1983). Value and population size. *Ethics 93*(3), 496–507.

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

Jonsson, A. and M. Voorneveld (2018). The limit of discounted utilitarianism. *Theoretical Economics 13*(1), 19–37.

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

Kaneda, T. and C. Haub (2018). How many people have ever lived on earth? Population Reference Bureau. First published in 1997, updated in 2002, 2011, and 2018. Accessed 22 November 2019. URL: https://www.prb.org/howmanypeoplehaveeverlivedonearth/.

Knobe, J., K. D. Olum, and A. Vilenkin (2006). Philosophical implications of inflationary cosmology. *The British Journal for the Philosophy of Science 57*(1), 47–67.

Kowalczyk, K. Equality and population size. Unpublished manuscript, November 2019.

Lauwers, L. and P. Vallentyne (2004). Infinite utilitarianism: More is always better. *Economics and Philosophy 20*(2), 307–330.

McCarthy, D. (2015). Distributive equality. *Mind 124*(496), 1045–1109.

McCarthy, D., K. Mikkola, and T. Thomas (2020). Utilitarianism with and without expected utility. *Journal of Mathematical Economics 87*, 77–113.

Mill, J. S. (1863). *Utilitarianism*. London: Parker, Son, and Bourne.

Moller, D. (2002). Parfit on pains, pleasures, and the time of their occurrence. *Canadian Journal of Philosophy 32*(1), 67–82.

Ng, Y. (1989). What should we do about future generations? Impossibility of Parfit's Theory X. *Economics and Philosophy 5*(2), 235–253.

Ng, Y. (1995). Towards welfare biology: Evolutionary economics of animal consciousness and suffering. *Biology and Philosophy 10*(3), 255–285.

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

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

Ord, T. (2021). The edges of our universe.

Page, D. N. (2007). Susskind's challenge to the Hartle-Hawking no-boundary proposal and possible resolutions. *Journal of Cosmology and Astroparticle Physics 01*(004), 1–20.

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

Parfit, D. (1997). Equality and priority. *Ratio 10*(3), 202–221.

Pressman, M. (2015). A defence of average utilitarianism. *Utilitas 27*(4), 389–424.

Prior, A. N. (1959). Thank goodness that's over. *Philosophy 34*(128), 12–17.

Rabinowicz, W. (1989). Act-utilitarian prisoner's dilemmas. *Theoria 55*(1), 1–44.

Roth, G. and U. Dicke (2005). Evolution of the brain and intelligence. *Trends in Cognitive Sciences 9*(5), 250–257.

Shulman, C. (2014). Population ethics and inaccessible populations. *Reflective Disequilibrium*. Accessed 25 August 2020. URL: `https://reflectivedisequilibrium.blogspot.com/2014/08/population-ethics-and-inaccessible.html`.

Sidgwick, H. (1907 (1874)). *The Methods of Ethics* (7th ed.). London: Macmillan and Company.

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

Sotala, K. and L. Gloor (2017). Superintelligence as a cause or cure for risks of astronomical suffering. *Informatica 41*(4), 389–400.

Thomas, T. (forthcoming). Separability. In G. Arrhenius, K. Bykvist, and T. Campbell (Eds.), *The Oxford Handbook of Population Ethics*. Oxford: Oxford University Press.

Tomasik, B. (2019). How many wild animals are there? First published 2009, updated 7 August 2019. Accessed 15 November 2019. URL: https://reducing-suffering.org/how-many-wild-animals-are-there/.

Vallentyne, P. and S. Kagan (1997). Infinite value and finitely additive value theory. *The Journal of Philosophy 94*(1), 5–26.

Vardanyan, M., R. Trotta, and J. 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.

Von Weizsäcker, C. C. (1965). Existence of optimal programs of accumulation for an infinite time horizon. *The Review of Economic Studies 32*(2), 85–104.

Weirich, P. (1983). Utility tempered with equality. *Noûs 17*(3), 423–439.