Non-additive axiologies in large worlds

Christian Tarsney and Teruji Thomas

Global Priorities Institute | September 2020

GPI Working Paper No. 9-2020

Non-Additive Axiologies in Large Worlds

Christian Tarsney* and Teruji Thomas*

September 2020

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: additive axiologies support 'arguments from astronomical scale' which suggest (among other things) that it is overwhelmingly important for humanity to avoid premature extinction and ensure the existence of a large future population, while non-additive axiologies need not. 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 astronomical scale, and other arguments in practical ethics that seem to presuppose additive separability, may be truth-preserving in practice whether or not we accept additive separability as a basic axiological principle.

1 Introduction

The world we live in is both large and populous. Our planet, for instance, is 4.5 billion years old and has has borne life for roughly 4 billion of those years. At any time in its recent history, it has played host to billions of mammals, trillions of vertebrates, and quintillions of insects and other animals, along

*Global Priorities Institute, Faculty of Philosophy, University of Oxford

with countless other organisms. Our galaxy contains hundreds of billions of stars, many or most of which have planets of their own. The observable universe is billions of light-years across, containing hundreds of billions of galaxies—and is probably just a small fraction of the universe as a whole. It may be, therefore, that our biosphere is just one of many (perhaps infinitely many). Finally, the *future* is potentially vast: our descendants could survive for a very long time, and might someday settle a large part of the accessible universe, gaining access to a vast pool of resources that would enable the existence of astronomical numbers of beings with diverse lives and experiences.

These facts have ethical implications. Most straightforwardly, the potential future scale of our civilization suggests that it is extremely important to shape the far future for the better. This view has come to be called *longtermism*, and its recent proponents include Bostrom (2003, 2013), Beckstead (2013, 2019), Cowen (2018), Greaves and MacAskill (2019), and Ord (2020). There are many ways in which we might try to positively influence the far future—e.g., building better and more stable institutions, shaping cultural norms and moral values, or accelerating economic growth. But one particularly obvious concern is ensuring the *long-term survival* of our civilization, by avoiding civilization- or species-ending 'existential catastrophes' from sources like nuclear weapons, climate change, biotechnology, and artificial intelligence.[1] Longtermism in general, and the emphasis on existential catastrophes in particular, have major revisionary practical implications if correct, e.g., suggesting the need for major reallocations of resources and collective attention (Ord, 2020, pp. 57ff).

All these recent defenses of longtermism appeal, in one way or another, to the *astronomical scale* of the far future. For instance, Beckstead's central argument starts from the premises that 'Humanity may survive for millions, billions, or trillions of years' and 'If humanity may survive may survive for millions, billions, or trillions of years, then the expected value of the future is astronomically great' (Beckstead, 2013, pp. 1–2). Importantly for our purposes, the astronomical scale of the far future most plausibly results from the *astronomical number of individuals* who might exist in the far future: while the far future population might consist, say, of just a single galaxy-spanning individual, the futures that typically strike longtermists as most worth pursuing involve a very large number of individuals with lives worth living (and conversely, the futures most worth avoiding involve a very large number of individuals with lives worth not living).

[1]The importance of avoiding existential catastrophe is especially emphasized by Bostrom (2003, 2013) and Ord (2020).

Under this assumption, we can understand arguments like Beckstead's as instantiating the following schema.

> **Arguments from Astronomical Scale**
> Because far more welfare subjects or value-bearing entities are affected by A than by B, we can make a much greater difference to the overall value of the world by focusing on A rather than B.

Beckstead and other longtermists take this schema and substitute, for instance, 'the long-run trajectory of human-originating civilization' for A and 'the (non-trajectory-shaping) events of the next 100 years' for B. To illustrate the scales involved, Bostrom (2013) estimates that if we manage to settle the stars, our civilization could ultimately support at least 10^{32} century-long human lives, or 10^{52} subjectively similar lives in the form of simulations. Since only a tiny fraction of those lives will exist in the next century or millennium, it seems *prima facie* plausible that even comparatively minuscule effects on the far future (e.g., small changes to the average welfare of the far-future population, or to its size, or to the probability that it comes to exist in the first place) would be vastly more important than any effects we can have on the more immediate future.[2]

Should we find arguments from astronomical scale persuasive? That is, does the fact that A affects vastly more individuals than B give us strong reason to believe, in general, that A is vastly more important than B? Although there are many possible complications, the sheer numbers make these arguments quite strong if we accept an axiology (a theory of the value of possible worlds or states of affairs) according to which the overall value of the world is simply a *sum* of values contributed by each individual in that world—e.g., the sum of individual welfare levels. In this case, the effect that some intervention has on the overall value of the world scales linearly with the number of individuals affected (all else being equal), and so astronomical scale implies astronomical importance.

But can the overall value of the world be expressed as such a sum? This question represents a crucial 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. For

[2]Thus, for instance, in reference to the 10^{52} estimate, Bostrom claims that 'if we give this allegedly lower bound...a mere 1 per cent chance of being correct, we find that the expected value of reducing existential risk by a mere *one billionth of one billionth of one percentage point* is worth a hundred billion times as much as a billion human lives' (Bostrom, 2013, p. 19).

example, *total utilitarianism* 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.

When we consider non-additive axiologies, the force of arguments from astronomical scale becomes much less clear, especially in variable-population contexts (i.e. when comparing possible populations of different sizes). They therefore represent a challenge to the case for longtermism and, more particularly, to the case for the overwhelming importance of avoiding existential catastrophe. 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 scale-sensitivity of their additive counterparts. This makes arguments from astronomical scale less reliant on the controversial assumption of additive separability. It thereby increases the robustness of the practical case for the overwhelming importance of the far future and 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.[3] 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.[4]

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, at a minimum, substantially larger than the present and near-future human population. 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. Thus, our arguments suggest, it is not merely the potential scale of the *future* that has important ethical implications, but also the scale of the world as a whole—in particular, the scale of the background population.

The paper proceeds as follows: section 2 introduces some formal concepts and notation, while section 3 formally defines additive separability

[3]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.)

[4]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, 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).

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 consider the size and other characteristics of real-world background populations and, in particular, argue that they are at least substantially larger than the present human population. 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 considers how real-world background populations affect the importance of avoiding existential catastrophe according to average utilitarianism and variable-value views. 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.

2 Formal setup

All of the population axiologies we will consider evaluate worlds based only on the number of welfare subjects at each welfare level. 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 theories of population axiology, like average utilitarianism, do not evaluate such empty worlds.

Thus for our 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* $\mathscr{A}$ 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 $\mathscr{A}$. 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 $X \succ_{\mathscr{A}} Y \iff V_{\mathscr{A}}(X) > V_{\mathscr{A}}(Y)$.

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

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

Similarly, the total welfare is

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

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 assume that $\mathscr{W}$ is given to us as a set of real numbers. With that in mind, the average welfare

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

is also well-defined.

3 Additivity

We can now give a precise definition of additive separability.

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]

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

[5] For a detailed discussion of separability principles in population ethics, see Thomas (forthcoming).

[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 (forthcoming), 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).

a given improvement in someone's welfare, regardless of how well off they were to begin with. We might intuit, however, that a one-unit improvement in the welfare of a very badly off individual has greater moral value than the same welfare improvement for someone who is already very well off. This intuition is captured by *prioritarian* theories.[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 strictly increasing and weakly concave.

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 can be replaced by

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

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 of 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, as the name suggests, identifies the value of a population with the average welfare level of that population.[10]

Average Utilitarianism (AU)

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

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 this and other results is deferred to §10.

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 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.

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.

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

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

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 the appendix), 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. Whether adding some individual to a population, or increasing the welfare of an existing individual, will increase or decrease equality 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 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).

[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.

5.1 Two-factor egalitarianism

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

> **Totalist Two-Factor Egalitarianism**
> $V(X) = \mathrm{Tot}(X) - I(X)|X|$, where I is some measure of inequality in X.
>
> **Averagist Two-Factor Egalitarianism**
> $V(X) = \overline{X} - I(X)$, where I is some measure of inequality in X.[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 $X/|X|\colon \mathscr{W} \to \mathbb{R}$ giving the proportion of the population in X having welfare w. 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, and then give some examples.

Theorem 3. *Suppose V is a value function of the form $V(X) = \mathrm{Tot}(X) - I(X)|X|$, or else $V(X) = \overline{X} - I(X)$, where I is a differentiable function of the distribution of X. Then the axiology $\mathscr{A}$ represented by V converges to an additive axiology relative to background populations with any given distribution D, with weighting function*[14]

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

[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.

[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 $D + t1_w$.

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)

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

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

[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.

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)

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

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

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, $\mathrm{QAM}(X) = \overline{X}$. More generally, QAA is ordinally equivalent to an averagist version of prioritarianism.

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 $\mathrm{MD}(w, D) := \sum_{x\in\mathscr{W}} D(x)|x - w|$ *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 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

[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.

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:

Geometric Rank-Discounted Utilitarianism (GRD)

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

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

[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]Using the standard notation in this paper, one can alternatively write

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

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).

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), and this is the case we will consider first, before returning to GRD.

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 $L > 0$.

In stating the result, 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.*

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.

In fact, GRD does *not* converge to an additive, Paretian axiology on any interesting range of populations. Roughly speaking, this is because,

[20] That is, there is some k such that f is convex on the interval (k, ∞). The assumption of eventual convexity is simply a technical assumption to be used in Theorem 6 below.

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. Nonetheless, it turns out that GRD *does* converge to a *separable*, Paretian axiology. We'll explain this carefully, but perhaps the most important take-away of this discussion will be that, given a large background population, GRD leads to some very strange and counterintuitive results. 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.[21]

In a sense, CLL_c is simply a limiting case of prioritarianism, where the priority given to the less-well-off is infinite. In particular, although it 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.[22]

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 population 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.*

Critical level leximin has a number of extreme and implausible features; as the theorem suggests, these will often be displayed by GRD when there is a large background population. For example, tiny benefits to worse-off individuals will often be preferred over astronomical benefits to even slightly

[21] 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.

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

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.[23] In fact, according to CLL_c, it makes things worse to add one person slightly below the critical level along with any number of people above the critical level; because of this, GRD implies what we might call the 'Snobbish Conclusion':

> **Snobbish Conclusion**
> Suppose X consists of one person with an arbitrarily good life, at level w, and any number of people with even better lives. Then there is some possible background population Z, in which the average welfare is far worse than w, and in which the very best lives are only slightly better than w, such that $Z + X$ is worse than Z.

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 Z such that $Z + X$ would be better than Z. 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 the limiting case, it makes things so much worse that it cannot be compensated by adding any number of even better lives.

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'

[23]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.

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 taken the separation between 'background' and 'foreground' populations as given, but it will now be helpful to make these notions more precise. 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 some *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 (perhaps simply equal to the number of individuals at each welfare level outside Earth's present future light cone).[24]

[24]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 research.

6.1 Population size

Past welfare subjects on Earth constitute the most obvious component of real-world background populations. 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 are also 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.[25] This is despite the significant decline in wild animal populations in recent centuries and millennia as a result of human encroachment.[26] 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 ($\sim$ 66 million years ago), with a population birth/death rate of 0.1 per individual per year. This gives us a background population of $\sim 6.6 \times 10^{17}$ individuals. Being a bit less conservative (though perhaps still objectionably 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

[25] 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).

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

(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.[27]

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).

[27] In the name of conservatism, we are setting aside several hypotheses that might generate much larger background populations. First, of course, even the restriction to vertebrates excludes potential welfare subjects like crustaceans and insects. Second, we Earthlings may not be the only welfare subjects. The observable universe contains roughly 2 trillion galaxies (Conselice et al., 2016), and the universe as a whole is likely to be many times larger (Vardanyan et al., 2011). The universe could therefore contain many other biospheres like Earth's. It might also contain advanced, spacefaring civilizations, which could support enormous populations on the order of 10^{30} individuals or more (Bostrom, 2003, 2011). So the extraterrestrial background population could be many—indeed, indefinitely many—orders of magnitude larger than the populations of past mammals or vertebrates on Earth. Finally, 'micropsychist' versions of panpsychism (advocated or sympathetically entertained by Chalmers (1996), Rosenberg (2004), Strawson (2006), and Roelofs and Buchanan (2019), among others) suggest that the number of welfare subjects is much larger than we intuitively suppose: perhaps as large as the number of elementary particles in the universe (or more, if these 'micro-subjects' can combine and recombine to form further, numerically distinct 'macro-subjects'). Since the number of *quarks* in the *observable* universe alone seems to be at least on the order of 10^{80} (Penrose, 1989, p. 340; Padilla, 2017), this hypothesis implies an *extremely* large background population.

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 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 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.

Here are 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

[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., forthcoming). However, though the temporal value asymmetry is a clear and robust *psychological* phenomenon, it has proven notoriously difficult to come 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.

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 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 affect 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

[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.

[31] The 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'.

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 way one might try 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 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, lifespan, or by a combination of 'spatial' and 'temporal' factors (e.g., lifespan times neuron count). 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.

Here are three replies to this suggestion: First, of course, one might lodge straightforward ethical objections to axiological weights. 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 (and for this reason, may be especially hard to reconcile with egalitarian views in axiology). It's also hard to imagine a plausible principle that assigns reduced axiological weight to non-human animals without also assigning reduced axiological weight to some humans, which many will find ethically unacceptable.

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

[33] Thanks to Tomi Francis and Toby Ord, who each separately suggested this objection.

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 $\sim 6.6 \times 10^{18}$ down to $\sim 2.3 \times 10^{15}$. 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, 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.

Third and 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 population in most present-day choice situations. To begin with, 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 value of avoiding 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 the background populations in real-world

[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.

choice situations are very large—at least, multiple orders of magnitude larger than the affectable portion of the present and near-future population. And in §§7–8, we resisted two strategies for deflating the size of real-world background populations.

If we are right about the size of real-world background populations, this provides a weak *prima facie* reason to believe that our limit results are practically significant—i.e., that what is true in the limit will be true in practice, for the most plausible versions of the various families of non-additive axiologies we have considered. That is, the absolute and relative size of real-world background populations weakly suggests that we should expect plausible non-additive axiologies to agree closely with their additive counterparts in real-world choice situations. 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 are likely to be truth-preserving in practice.

9.1 Present welfare vs. future population size

But we will focus on a particular issue in practical ethics where we can say something a bit more concrete and definite. As we suggested in §1, perhaps the most important practical implication of our results concerns the importance of existential catastrophes—more specifically, the extent to which the potentially astronomical scale of the far future makes it astronomically important to avoid existential catastrophe. An 'existential catastrophe', for our purposes, is any near-future event that would drastically reduce the future population size of human-originating civilization (e.g., human extinction).[35] To keep the discussion manageable, we will focus on AU and, secondarily, VV1/VV2. This lets us isolate the central relevant feature of insensitivity to scale (or asymptotic insensitivity to scale) in the absence of background populations, without the essentially orthogonal feature of inequality aversion.[36] We will also focus on the case where the future generations that will exist if we avoid existential catastrophe

[35]This is a broader category of events than '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).

[36]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.

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 much 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.)

The importance of avoiding existential catastrophe can be measured by comparing the value of avoiding existential catastrophe with the value of improving the welfare of the affectable pre-catastrophe population (which, for simplicity, we will hereafter call 'the current generation'). We would like to know how the answer to this question depends on the welfare and (especially) the size of the background population.

To formalize the question, let C represent the current generation as it will be if we prioritize its welfare at the expense of allowing an existential catastrophe. Let C' denote the current generation as it will be if we instead prioritize avoiding an existential catastrophe. Thus $\overline{C} > \overline{C'}$, but we assume that $|C| = |C'|$. (This is mostly harmless: it just means that we designate as the members of C' the first $|C|$ individuals in the affectable population in the world where we avoid existential catastrophe.) 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.

In short, we consider a choice between $Z + C$ and $Z + C' + F$. In terms of this choice, the importance of avoiding existential catastrophe can be made precise in several different ways. We will consider the following three:

> *Maximum incurred cost.* Holding fixed the average welfare $\overline{C}$ of the current generation in the world where existential catastrophe occurs, what is the greatest reduction in welfare for the current generation that is worth accepting to avoid existential catastrophe?
>
> *Maximum opportunity cost.* Holding fixed the average welfare $\overline{C'}$ of the current generation in the world where existential catastrophe *does not* occur, what is the greatest improvement in the welfare of the current generation that is worth forgoing to to avoid existential catastrophe?
>
> *Value difference ratio.* Holding fixed both $\overline{C}$ and $\overline{C'}$, and thinking of $Z + C'$ as the status quo, what is the ratio between the changes in value that would result from (i) avoiding existential catastrophe by adding F, versus (ii) raising the welfare of the current generation from $\overline{C'}$ to $\overline{C}$?

Broadly, we want to know how the presence of Z affects these measures of importance. We know they depend, for one thing, on the size of F; we want particularly to know how this dependence is mediated by the size of Z. In the extreme case, as $|Z| \to \infty$, we know from our results in §4 that AU, VV1, and VV2 all converge to $\mathrm{CL}_{\overline{Z}}$. And according to $\mathrm{CL}_{\overline{Z}}$, the value of adding F to the population scales with $|F|$, so that when $|F|$ is astronomically large, the importance of avoiding existential catastrophe, by any of these measures, will be astronomically great. We should therefore expect, a bit roughly, that AU will give great importance to avoiding existential catastrophe when both $|Z|$ and $|F|$ are large, and more precisely that its measures of importance will agree with those of $\mathrm{CL}_{\overline{Z}}$. The task is to say more about how this works at a qualitative level, and then (in §9.5) to give some indicative numerical results.

9.2 Measure 1: Maximum incurred cost

First, we hold fixed the welfare of the the current generation in the catastrophe world (where F does not exist), and consider the greatest welfare cost we are willing to impose on the current generation to avoid catastrophe and thereby add F to the population.

According to the $\mathrm{CL}_{\overline{Z}}$, the axiology to which AU, VV1, and VV2 converge in the limit, this is simply the critical-level sum of welfare in F, given by $|F|(\overline{F}-\overline{Z})$. That is, when $\mathrm{Tot}(C)-\mathrm{Tot}(C') = |F|(\overline{F}-\overline{Z})$, CL is indifferent between $Z+C'+F$ and $Z+C$. According to AU, analogously, the maximum cost we are willing to impose on the current generation is the cost at which $\overline{Z+C'+F} = \overline{Z+C}$. We solve for it, therefore, by rearranging this equation into an equation for $\mathrm{Tot}(C)-\mathrm{Tot}(C')$ in terms of Z, C, and F.[37] This rearranged equation turns out to be:

$$\mathrm{Tot}(C)-\mathrm{Tot}(C') = \frac{|Z||F|(\overline{F}-\overline{Z})+|C||F|(\overline{F}-\overline{C})}{|Z+C|}. \tag{3}$$

The key thing to notice about this equation is its surprising implication that the importance of avoiding existential catastrophe in the 'maximum incurred cost' sense scales linearly with $|F|$, with or without a background population. As we will see, this is not the case for the other two measures of the importance of avoiding existential catastrophe we consider. The right way to interpret this fact is as follows: if $\overline{F} > \overline{C}$ and $|F| \gg |C|$, then AU is willing to impose enormous costs on the current generation to enable

[37] If we instead wanted to focus on the *average* (per capita) welfare cost imposed on the current generation, we could just divide both sides of the following equation by $|C|$.

the existence of F, since if F exists, C' will be only a very small part of the resulting population and must have extremely low average welfare to reduce $\overline{C'+F}$ below $\overline{C}$. And on the other hand, if $\overline{C} > \overline{F}$ and $|F| \gg |C|$, then AU will require an enormous increase in the welfare of the current generation (i.e., that $\overline{C'} \gg \overline{C}$) to compensate for the reduction in average welfare created by F.

Nevertheless, even by this measure, the size of the background population makes a difference because it determines the 'effective critical level' to which $\overline{F}$ is compared—the average welfare level above which adding F to the population has positive value, and below which it has negative value. When $|C| \gg |Z|$, the right-hand side of (3) is approximately $|F|(\overline{F}-\overline{C})$;[38] thus AU agrees closely with $\mathrm{CL}_{\overline{C}}$ rather than $\mathrm{CL}_{\overline{Z}}$ and is only willing to impose any positive cost at all on the current generation to avoid existential catastrophe when (with some approximation) $\overline{F} > \overline{C}$. But when $|Z| \gg |C|$, the right-hand side of (3) is approximately $|F|(\overline{F}-\overline{Z})$—i.e., the value given by $\mathrm{CL}_{\overline{Z}}$. This shift could either increase or decrease the value of avoiding existential catastrophe (depending on whether $\overline{Z}$ is greater than or less than $\overline{C}$), and could reverse the sign of the value of avoiding existential catastrophe if $\overline{F}$ is between $\overline{Z}$ and $\overline{C}$. Most notably for our purposes, the effective critical level will be closer to $\overline{Z}$ than to $\overline{C}$ if $|Z| > |C|$, and will be very close to $\overline{Z}$ if $|Z| \gg |C|$ (since $|Z||F|(\overline{F}-\overline{Z})$ rather than $|F||C|(\overline{F}-\overline{C})$ will dominate the numerator in (3)). So by this measure, AU closely agrees with its corresponding additive limit theory as long as the background population is substantially larger than the current generation, i.e., $|Z| \gg |C|$.

9.3 Measure 2: Maximum opportunity cost

Now let's ask the converse question: holding fixed the welfare of the current generation in the world *without* existential catastrophe (i.e. holding fixed $\overline{C'}$), how large a welfare *gain* for the current generation should we be willing to *forgo* to avoid existential catastrophe?

Here again, CL gives the answer $|F|(\overline{F}-\overline{Z})$. To find AU's answer, we rearrange $\overline{Z+C'+F} = \overline{Z+C}$ into an equation for $\mathrm{Tot}(C)-\mathrm{Tot}(C')$, this time in terms of Z, F, and C'. This gives us:

$$\mathrm{Tot}(C)-\mathrm{Tot}(C') = \frac{|Z||F|(\overline{F}-\overline{Z})+|C'||F|(\overline{F}-\overline{C'})}{|Z+C'+F|}. \tag{4}$$

Now the size of the background population takes on greater significance. Consider three cases:

[38] Formally, 'if $a \gg b$ then x is approximately y' means that $\lim_{a/b\to\infty} x/y = 1$. In this case, the limit converges uniformly in $|F|$.

Case 1: $|\mathbf{F}| \gg |\mathbf{C}'| \gg |\mathbf{Z}|$. In this case, the right-hand side of (4) is approximately $|C'|(\overline{F} - \overline{C'})$, and the value of avoiding existential catastrophe as measured by maximum opportunity cost is therefore approximately independent of $|F|$.[39]

Case 2: $|\mathbf{F}| \gg |\mathbf{Z}| \gg |\mathbf{C}'|$. In this case, the right-hand side of (4) is approximately $|Z|(\overline{F} - \overline{Z})$. Thus the value of avoiding existential catastrophe as measured by maximum opportunity cost is approximately proportional to $|Z|$, which may be astronomically large but is also (we are supposing) much less than $|F|$. Note also that the effective critical level is now close to $\overline{Z}$ rather than $\overline{C'}$ as in Case 1.

Case 3: $|\mathbf{Z}| \gg |\mathbf{F}| \gg |\mathbf{C}'|$. In this case, the right-hand side of (4) is approximately $|F|(\overline{F} - \overline{Z})$, in agreement with $\mathrm{CL}_{\overline{Z}}$. Thus the value of avoiding existential catastrophe as measured by maximum opportunity cost is approximately proportional to $|F|$, and will be astronomically large if $|F|$ is astronomically large and $(\overline{F} - \overline{Z})$ is non-trivial.

While there are a number of points of interest in this analysis, the quick takeaway is that the maximum opportunity cost increases without bound as we increase *both* $|F|$ and $|Z|$ (while holding all else equal)—a situation reflected in Cases 2 and 3 but not Case 1. So, qualitatively, arguments from astronomical scale can go through if we attend to the potentially astronomical scale of both the future population *and* the background population.

9.4 Measure 3: Value difference ratio

Finally, we treat $Z + C'$ as a baseline, and ask whether it is better to avoid existential catastrophe by adding F or to improve C' to C . More precisely, we consider the ratio of the value of these improvements:

$$R = \frac{V(Z + C' + F) - V(Z + C')}{V(Z + C) - V(Z + C')}.$$

According to $\mathrm{CL}_{\overline{Z}}$, R is equal to $\frac{|F|(\overline{F} - \overline{Z})}{|C|(\overline{C} - \overline{C'})}$. According to AU, of course, R is equal to $\frac{\overline{Z+C'+F} - \overline{Z+C'}}{\overline{Z+C} - \overline{Z+C'}}$. But again, we need to do some rearranging to make clear how this ratio is affected by the sizes of Z, C, and F. Specifically, in the case of AU, the formula for R rearranges to

$$\frac{1}{\overline{C} - \overline{C'}}\left(\overline{F}\frac{|F||Z + C|}{|C||Z + C + F|} - \overline{C'}\frac{|F|}{|Z + C + F|} - \overline{Z}\frac{|Z||F|}{|C||Z + C + F|}\right). \quad (5)$$

[39] Formally, a claim to the effect of 'if $a \gg b \gg c$ then x is approximately y' means that $x/y \to 1$ as a/b and $b/c \to \infty$; more precisely, for any $\epsilon > 0$, there exists $n > 0$ such that if both a/b and b/c are bigger than n, then $x/y \in (1 - \epsilon, 1 + \epsilon)$.

This expression is unattractive, but informative. Again, let's consider three cases:

Case 1: $|\mathbf{F}| \gg |\mathbf{C}| \gg |\mathbf{Z}|$. In this case, (5) is approximately $\frac{\overline{F}-\overline{C'}}{\overline{C}-\overline{C'}}$, and the importance of avoiding existential catastrophe by the value difference ratio measure is therefore approximately independent of $|F|$.

Case 2: $|\mathbf{F}| \gg |\mathbf{Z}| \gg |\mathbf{C}|$. In this case, (5) is approximately $\frac{|Z|}{|C|} \times \frac{\overline{F}-\overline{Z}}{\overline{C}-\overline{C'}}$. Thus the importance of avoiding existential catastrophe by the value difference ratio measure is approximately proportional to $\frac{|Z|}{|C|}$. And again, note that when $|Z| \gg |C|$, the effective critical level is close to $\sim \overline{Z}$ rather than $\overline{C'}$.

Case 3: $|\mathbf{Z}| \gg |\mathbf{F}| \gg |\mathbf{C}|$. In this case, (5) is approximately $\frac{|F|}{|C|} \times \frac{\overline{F}-\overline{Z}}{\overline{C}-\overline{C'}}$, in agreement with $\mathrm{CL}_{\overline{Z}}$. Thus the importance of avoiding existential catastrophe by the value difference ratio measure is now approximately proportional to $\frac{|F|}{|C|}$, and will be astronomically large if $\frac{|F|}{|C|}$ is astronomically large and $\frac{\overline{F}-\overline{Z}}{\overline{C}-\overline{C'}}$ is non-trivial.

As with the maximum opportunity cost, the most basic qualitative point is that the value difference ratio R increases without bound 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 the value difference ratio will be greater than 1 (thus favouring extinction-avoidance) for a robust range of the other parameters.

9.5 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 catastrophes, for some reasonable and even conservative estimates of the background population and other parameters; 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 $|C| = |C'| = 10^{10}$ (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 population, if we avoid existential catastrophe, arrived at 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

Axiology	**|Z|**	**MIC**	**MOC**	**VDR**
AU	$\|Z\| = 0$	5×10^{16}	$\sim 10^{10}$	~ 2
AU	$\|Z\| = 10^{13}$	$\sim 1.9985 \times 10^{17}$	$\sim 2.0008 \times 10^{13}$	$\sim 4.0016 \times 10^{3}$
AU	$\|Z\| = 10^{20}$	$\sim 2 \times 10^{17}$	$\sim 1.998 \times 10^{17}$	$\sim 3.996 \times 10^{7}$
CL	—	2×10^{17}	2×10^{17}	4×10^{7}

TABLE 1: The importance of avoiding existential catastrophe, as measured by maximum incurred cost (MIC), maximum opportunity cost (MOC), and value difference ratio (VDR), according to AU for different background population sizes and $\text{CL}_{\overline{Z}}$, with $\overline{F} = 2$, $|F| = 10^{17}$, $\overline{C} = 1.5$, $\overline{C'} = 1$, $|C| = |C'| = 10^{10}$, $\overline{Z} = 0$, and $|Z|$ as specified in each row.

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).

In terms of average welfare, we have much less to go on. For simplicity let's assume that $\overline{F} = 2$ (corresponding 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 $\overline{C'} = 1$ (except when considering maximum incurred cost, where $\overline{C'}$ is a dependent variable) and $\overline{C} = 1.5$ (except when considering maximum opportunity cost, where $\overline{C}$ is a dependent variable).

Table 1 gives the importance of avoiding existential catastrophe according to AU and $\text{CL}_{\overline{Z}}$, under these assumptions, for all three measures of importance and all three background population sizes. In general, we see that with three- or four-order-of-magnitude differences in the population sizes of C, F, and Z, the approximations arrived at above are accurate to at least the third or fourth significant figure. And more specifically, in the case where $|Z| \gg |F| \gg |C|$, AU agrees with $\text{CL}_{\overline{Z}}$ on all three measures to at least the fourth significant figure.

9.6 Conclusions

In summary: when the background population is small or non-existent, the importance of avoiding existential catastrophe according to AU is approximately proportional to $\overline{F} - \overline{C'}$ or $\overline{F} - \overline{C}$ (depending on which measure we consider), and approximately independent of population size, and is therefore unlikely to be astronomically large. 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 $\mathrm{CL}_{\overline{Z}}$. 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), AU agrees closely with $\mathrm{CL}_{\overline{Z}}$ about the importance of avoiding existential catastrophe, treating it as approximately linear in $|F|$, by all three of the measures we considered. The exception to this pattern is the 'maximum incurred cost' measure, by which the importance of avoiding existential catastrophe scales with $|F|$ regardless of the size of the background population.

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': AU closely approximates $\mathrm{CL}_{\overline{Z}}$ in every respect we have considered only when $|Z| \gg |F|$ (or at any rate, only when $|Z| > |F|$). But it behaves in important ways like $\mathrm{CL}_{\overline{Z}}$ as long as $|Z| \gg |C|$—both in that it is disposed to assign astronomical 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 AU evaluates existential catastrophes in practice.

Our conclusions about AU also partially generalize to VV1 and VV2. In the case of VV1: for any two populations X and Y, if $\overline{X} > \overline{Y}$, $|X| \geq |Y|$, and $\overline{X} \geq 0$, then clearly any VV1 axiology will prefer X to Y. For our purposes, this means that any VV1 axiology, so long as it assigns non-negative value to the non-catastrophe population $Z + C' + F$ (i.e., so long as $\overline{Z + C' + F} \geq 0$), will prefer that population to the catastrophe population $C + Z$ whenever AU does. Analogously, in the case of VV2 (which, recall, applies an increasing transformation f to the average welfare of a population): for any two populations X and Y, if $\overline{X} > \overline{Y}$, $|X| \geq |Y|$, and $f(\overline{X}) \geq 0$, then clearly any VV2 axiology will prefer X to Y. For our purposes, this means that any VV2 axiology, so long as it assigns non-negative value to the non-catastrophe population $Z + C' + F$ (i.e., so long as $f(\overline{Z + C' + F}) \geq 0$), will prefer that population to the catastrophe population $Z + C$ whenever AU does.

Putting these observations together, any VV axiology, as long as it assigns positive value to the non-catastrophe population, will prefer it to the catastrophe population whenever AU does. This means, among other things, that under this condition, the importance of avoiding existential catastrophe as measured by maximum incurred cost or maximum opportunity cost, will be at least as great according to VV as according to AU.[40] With respect to value difference ratio, things are a bit more complicated: when $\overline{Z+C'+F} \geq 0$ *and* $\overline{Z+C'+F} \geq \overline{Z+C'}$, VV1 is guaranteed to assign more importance than AU to avoiding existential catastrophe by this measure. But we cannot say anything analogous about VV2 in this case, since the transformation f it applies to average welfare can be arbitrarily concave or convex.[41]

A crucial limitation of our discussion, however, is that we have only considered the *objective importance of existential catastrophes*, and not the prospective or decision-theoretic significance of existential *risks* (i.e., risks of existential catastrophe). If we assume a straightforward expectational decision theory according to which average utilitarians, for instance, should simply maximize expected average welfare, then the astronomical decision-theoretic significance of existential risk would follow straightforwardly from the astronomical axiological significance of existential catastrophe in the 'value difference ratio' sense (assuming, of course, that we can have non-negligible effects on the probability of existential catastrophe). We have sidestepped the question of risk, however, because there are good reasons to think that non-additive axiologists should be in the market for something other than this simple expectational theory of decision-making under risk[42], and there is not yet any unproblematic or widely accepted alternative in

[40]Consider VV2, of which VV1 is a special case (where $f(\overline{X}) = \overline{X}$). If $f(\overline{Z+C'+F+C'}) = f(\overline{Z+C})$, and is positive, then $g(|Z+C'+F|)f(\overline{Z+C'+F}) > g(|Z+C|)f(\overline{Z+C})$, since g is increasing. Thus, all else being equal, VV2 axiologies will require either a larger value of $\overline{C}$ or a smaller value of $\overline{C'}$ to equalize the value of the populations, meaning that the maximum incurred cost/maximum opportunity cost that it will accept to avoid existential catastrophe is greater.

This does not necessarily mean that VV will converge with $\mathrm{CL}_{\overline{Z}}$ faster than AU, with respect to these measures, as the size of the background population increases. After all, g may be arbitrarily close to linear up to arbitrarily large population sizes, allowing VV to remain in close agreement with TU rather than $\mathrm{CL}_{\overline{Z}}$ for arbitrarily large populations. But it does mean that VV will converge with $\mathrm{CL}_{\overline{Z}}$ faster than AU if it is converging from below.

[41]If $\overline{Z+C'+F}$ and $\overline{Z+C'+F} - \overline{Z+C'}$ are both non-negative, then $\frac{g(|Z+C+F|)\overline{Z+C'+F} - g(|Z+C|)\overline{Z+C'}}{g(|Z+C|)\overline{Z+C} - g(|Z+C|)\overline{Z+C'}} \geq \frac{\overline{Z+C'+F} - \overline{Z+C'}}{\overline{Z+C} - \overline{Z+C'}}$. (Again, this means that VV1 will converge with $\mathrm{CL}_{\overline{Z}}$ faster than AU, with respect to the value difference ratio measure, if it is converging from below.) However, since VV2's f need only be increasing, $\frac{f(\overline{Z+C'+F}) - f(\overline{Z+C'})}{f(\overline{Z+C}) - f(\overline{Z+C'})}$ can differ to an arbitrarily extreme degree from $\frac{\overline{Z+C'+F} - \overline{Z+C'}}{\overline{Z+C} - \overline{Z+C'}}$ (except when $\overline{Z+C'+F} = \overline{Z+C}$).

[42]See for instance Thomas (2016, ch. 3), McCarthy et al. (2020, Prop. 4.8), Nebel (forthcoming), Tarsney (unpublished).

the literature. We therefore leave the question of how AU, VV1, VV2, and other non-additive axiologies evaluate existential risk in the presence of large background populations for future research.

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 preferences 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.[43] Most of the extant proposals for ranking infinite worlds, in both philosophy and economics, aim to extend total utilitarianism.[44] 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.[45] 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, at least when these populations 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.

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

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

[44]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), Wilkinson (forthcoming), and Clark (ms), among many others.

[45]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).

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.[46] 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

[46] 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.

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. 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. The most notable implication of these arguments is that 'arguments from astronomical scale'—in particular, for the overwhelming importance of existential catastrophes—need not depend on an assumption of axiological 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 the real-world background population; (2) the significance of risk/uncertainty, particularly with respect to these characteristics 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 Results

Recall that $\mathcal{W}$ is the set of welfare levels, and $\mathcal{P}$ consists of all non-zero, finitely supported functions $\mathcal{W} \to \mathbb{Z}_+$. By a *type* of populations we mean a set $T \subset \mathcal{W}$ 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 $V : \mathcal{P} \to \mathbb{R}$ and a positive function $s : \mathbb{N} \to \mathbb{R}$. Define*

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

as Z ranges over populations of some type T. If 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

$$V_{\mathrm{AU}}(X+Z) - V_{\mathrm{AU}}(Z) = \frac{(\overline{X} - \overline{Z})|X|}{|X| + |Z|} = \frac{V_{\mathrm{CL}_c}(X)}{|X| + |Z|}. \tag{6}$$

Setting $s(n) = n$ we find $V_{\mathrm{AU}}^s(X) = V_{\mathrm{CL}_c}(X)$, 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 (6), 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,

$$|Z|(V_{\mathrm{CL}_c}(X) - V_{\mathrm{CL}_c}(Y)) > |X| V_{\mathrm{CL}_c}(Y) - |Y| V_{\mathrm{CL}_c}(X). \tag{7}$$

Given that $V_{\mathrm{CL}_c}(X) - V_{\mathrm{CL}_c}(Y) > 0$, the desired conclusion (7) 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

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

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[47]

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

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. $\square$

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 (6),

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

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. $\square$

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 $\overline{X} \ge 0 \ge \overline{Y}$, 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 $|X| \ge |Y|$, 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

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

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

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. $\square$

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*[48]

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

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.

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 $\tilde{I}\colon \mathscr{D} \to \mathbb{R}$ 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:[49] for all $P, Q \in \mathscr{P}_{\mathbb{R}}$, the limit

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

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

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

[48] 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 $D + t1_w$.

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

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:

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

For totalist egalitarianism, we find that

$$V^s(X) = \mathrm{Tot}(X) - \partial_X I(D) - I(D)|X|.$$

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

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

Similarly, for averagist egalitarianism,

$$\begin{aligned} V^s(X) &= (\overline{X} - \overline{D})|X| - \partial_X I(D) \\ &= \sum_{w\in\mathscr{W}} X(w)f(w) \quad \text{with} \quad f(w) = w - \partial_{1_w} I(D) - \overline{D}. \end{aligned}$$

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 $V(X^+ + Z) \geq V(X+Z)$ 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 $\mathrm{MD}(w,D) := \sum_{x\in\mathscr{W}} D(x)|x-w|$ *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. Following the proof of Theorem 3, we know 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).

□

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 $X_{\le w} = \sum_{x \le w} X(w)$ 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

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

The assumption that X is moderate means that, in those cases where $X(w) \ge 1$, so that the first inner sum is non-trivial, we also have $Z_{\le w} \to \infty$. We see therefore that each summand in the first double-sum tends to Lw. The first double sum then converges to $\sum_{w \in \mathscr{W}} X(w) L w = L\,\mathrm{Tot}(X)$. 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 $Z(w) \ge 1$ (making the inner sum 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 $X_{<w} \ge 1$. We can also assume that $Z_{<w}$ is large enough that f is convex in the relevant range; then

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

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 population 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 $\tilde{X}$ and $\tilde{Y}$ be populations of equal size, obtained from X and Y by adding people at the critical level c. The assumption that $X \succ_{\mathrm{CLL}} Y$ means that, for the first m such that $\tilde{X}_m \ne \tilde{Y}_m$, we have $\tilde{X}_m > \tilde{Y}_m$. This shows that $\tilde{Y}_m < c$, so that in fact $\tilde{Y}_m = Y_m$. For brevity define $w := Y_m$.

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 for 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 in (w, v), and Z_+ only involves those in $[v, \infty)$. Note that $X_- = Y_-$ and $X_0 = Y_0 = 0$ but (because D covers W and is only supported up to c) $Z_0 \neq 0$. We have

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

A similar expression holds for Y in place of X. Therefore

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

where the remainder R is such that $\lim_{n\to\infty} R = 0$. Now we use the standard fact that $\sum_{i=1}^{m} \beta^i = \beta \frac{1-\beta^m}{1-\beta}$. It follows that $V(X_w) - V(Y_w) = \beta \frac{\beta^{|Y_w|} - \beta^{|X_w|}}{1-\beta} w$. Therefore

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

Note that $\beta^{|X_w|} - \beta^{|Y_w|} > 0$. To conclude that $V(X+Z) > V(Y+Z)$ for all n large enough, it suffices to show that

$$\lim_{n\to\infty} V(Z_0) > \frac{\beta w}{1-\beta}.$$

In fact, if v' is the lowest welfare level greater than w occurring in D, then $v' \in (w, v)$ and $\lim_{n\to\infty} V(Z_0) = \frac{\beta v'}{1-\beta}$. □

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.

Beckstead, N. (2013). *On the Overwhelming Importance of Shaping the Far Future.* Ph. D. thesis, Rutgers University Graduate School - New Brunswick.

Beckstead, N. (2019). A brief argument for the overwhelming importance of shaping the far future. In H. Greaves and T. Pummer (Eds.), *Effective Altruism: Philosophical Issues*, pp. 80–98. Oxford: Oxford University Press.

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.

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.

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.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Clark, M. (2019). Infinite ethics, intrinsic value, and the Pareto principle. Unpublished manuscript.

Conselice, C. J., A. Wilkinson, K. Duncan, and A. Mortlock (2016). The evolution of galaxy number density at $z < 8$ and its implications. *The Astrophysical Journal 830*(2), 1–17.

Cowen, T. (2018). *Stubborn Attachments: A Vision for a Society of Free, Prosperous, and Responsible Individuals*. San Francisco: Stripe Press.

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.

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.

Greaves, H. and W. MacAskill (2019). The case for strong longtermism. Global Priorities Institute Working Paper No. 7-2019.

Greene, P., A. J. Latham, K. Miller, and J. Norton (forthcoming). Hedonic and non-hedonic bias towards the future. *Australasian Journal of Philosophy*.

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

Gustafsson, J. E. (forthcominga). 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.

Gustafsson, J. E. (forthcomingb). Population axiology and the possibility of a fourth category of absolute value. *Economics and Philosophy*.

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.

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.

Nebel, J. M. (forthcoming). Rank-weighted utilitarianism and the veil of ignorance. *Ethics.*

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.

Padilla, T. (2017). How many particles in the universe? Numberphile. URL: https://www.youtube.com/watch?v=lpj0E0a0mlU.

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

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

Penrose, R. (1991 (1989)). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics.* New York: Penguin Books.

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.

Roelofs, L. and J. Buchanan (2019). Panpsychism, intuitions, and the great chain of being. *Philosophical Studies 176*(11), 2991–3017.

Rosenberg, G. H. (2004). *A Place for Consciousness: Probing the Deep Structure of the Natural World.* Oxford: Oxford University Press.

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.

Strawson, G. (2006). Realistic monism: Why physicalism entails panpsychism. *Journal of Consciousness Studies 13*(10-11), 3–31.

Tarsney, C. J. Average utilitarianism implies solipsistic egoism. Unpublished manuscript, March 2020.

Thomas, T. (2016). *Topics in Population Ethics.* Ph. D. thesis, Oxford University.

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.

Wilkinson, H. (forthcoming). Infinite aggregation: Expanded addition. *Philosophical Studies.*