Calibrating variable-value population ethics: Implications for the effort to avoid the repugnant conclusion

March 10, 2022

Abstract

Population ethics has focused on the effort to avoid Parfit's *Repugnant Conclusion*. But a growing literature suggests that this effort has been misunderstood, and may therefore have been misplaced. Here, we add to this literature by exploring certain leading Variable-Value axiologies. These views avoid Parfit's Repugnant Conclusion, while satisfying some weak instances of the *Mere Addition* principle (for example, at small population sizes). We apply calibration methods to Variable-Value views conditional upon: first, some very weak instances of Mere Addition, and, second, some plausible empirical assumptions about the size and welfare of the intertemporal world population. We find that such facts calibrate Variable-Value views to be nearly totalist, and therefore imply conclusions that should seem repugnant to anyone who opposes Total Utilitarianism only due to the Repugnant Conclusion. So, any wish to avoid repugnant conclusions is not a good reason to choose a Variable-Value view. More broadly, our results join a recent literature arguing that prior efforts to avoid the Repugnant Conclusion hinge on inessential features of the formalization of repugnance — and therefore may be less normatively significant than is traditionally assumed.

1 Introduction

Much research in population ethics is motivated by the quest to avoid what Parfit (1984) called the *Repugnant Conclusion*, one version of which states that:[1]

[1] Parfit's own formulation of the Repugnant Conclusion states that: "For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better, even though its members have lives that are barely worth living." (Parfit, 1984, p. 388) Our formulation is closer to Arrhenius's (forthcoming). Spears and Budolfson (2021) have argued that formalizations of the Repugnant Conclusion should be broader — including, for example, additions to unaffected, intersecting populations — but for this paper we ignore that proposal and focus on what they call a "restricted" formalization.

The Repugnant Conclusion (Informal version)**.** *For any perfectly equal population of very well-off people, there is a better population consisting entirely of lives that are barely worth living.*

Total Utilitarianism, according to which a population is better the greater sum of welfare it contains, is widely recognized to entail the Repugnant Conclusion. No matter how well-off people are in some population A, and independently of A's size, there is some (potentially much bigger) imaginable population Z that contains a greater sum of welfare than A does — even though people in Z have lives that are each barely worth living (understood as having barely positive welfare).

Most paths to avoiding the Repugnant Conclusion begin by abandoning what Parfit called the *Mere Addition principle*, which can be stated thus:

Mere Addition (Informal version)**.** *By adding any life worth living to any population, without making anyone else worse off, we do not make the population worse.*

Total Utilitarianism implies the Mere Addition principle. But this principle is violated by *Average* Utilitarianism, according to which a population is better the greater average welfare it contains. And Average Utilitarianism avoids the Repugnant Conclusion: Population Z, whose members all have lives that are barely worth living, contains lower average welfare than A. So, A is better than Z, according to Average Utilitarianism.

Somebody who abandons Mere Addition thinks that adding a life worth living, without making anyone worse off, can make a population worse. But what about adding a life *well* worth living? Consider merely adding a person who lives a very good life by modern standards: say, a happy professor living in a developed country in 2022. Surely by adding a person like that to any population, without thereby making anyone else worse off, we have not made the population worse? Not according Average Utilitarianism. To see this in an absurd example: adding our professor to a single-person "population" whose member is only a tiny bit better-off than the professor makes the resulting population *worse*, according to Average Utilitarianism. In fact, if the future of humanity is as long and wonderful as some hope (Ord, 2020), then adding a person likes this to the *actual* intertemporal world population[2] makes the resulting population worse, according to Average Utilitarianism. This anti-natalist implication of Average Utilitarianism violates what we shall call *Weak Mere Addition* (which we state formally in section 2).

In light of the above counterintuitive implications of on the one hand Total Utilitarianism (implying the Repugnant Conclusion) and on the other hand of Average Utilitarianism (violating Weak Mere Addition), some theorists have been attracted to a family of views that are often called *Variable-Value views*.[3] Some views within this family avoid the Repugnant Conclusion altogether while capturing the intuition that adding a well-off person to a *small* population always makes the resulting population better. More specifically, these views hold that the quantity that added persons (with a fixed level of welfare) contribute towards the overall value of a population decreases as the size of the population increases, *cumulatively* contributing only a bounded amount, which is how such views escape the Repugnant Conclusion.

Various versions of Variable-Value views have been rigorously formalized. These formalizations and the ensuing analysis have focused on *qualitative* properties of Variable-Value views: with which *axioms* do they comply? However, there has not been a similar focus on the *quantitative* implications of Variable-Value views. In particular, one might wonder *how fast* the quantity that an added person contributes towards the overall value of a population diminishes as e.g. the size and average welfare of the population increases, and what implications that will have for various trade-offs between increasing the size and the average welfare of a population. Similarly, one might wonder precisely which weak-

[2] By "intertemporal world population" we mean the totality of humanity (see fn. 5) throughout history.

[3] Hurka (1983) coined the term, and was probably the first to suggest such a view in response to Parfit's Repugnant Conclusion, but views in this family have since been proposed or investigated by Ng (1989), Sider (1991), Asheim and Zuber (2014), and Pivato (2020), although not all of these authors endorsed the Variable-Value axiology that they identified or explored.

enings of the Mere Addition principle these views can accommodate without implying seemingly repugnant instances of the Repugnant Conclusion.[4]

Our aim in this paper is to fill this quantitative gap in the population ethics literature. In particular, we shall conditionally assume some very weak and, we propose, intuitively compelling instances of Mere Addition and calibrate what Variable-Value views, that satisfy such weak instances of Mere Addition (but violate the stronger Mere Addition principle that Total Utilitarianism entails), imply under what we take to be plausible empirical assumptions about the future. Informally, the weak Mere Addition that we assume ensures that merely adding people who are very well-off by modern standards, such as happy professors in the developed world, does not make the population worse. The empirical assumption we make is that the future of humanity is long and prosperous, such that, in particular, the average welfare of the total intertemporal world population[5] is much higher than the average welfare of the world population up to 2022.[6]

To be clear upfront: from the point of view of the universe there is nothing special about relatively well-off professors in 2022 when evaluating the full intertemporal population. We write of this case merely because we expect that both you, the reader, and we, the authors, have some intuition about it. Nor need it be true for our argument to succeed that

[4] Our aim is not to examine *all* Variable-Value views. In particular, because we are principally investigating the usefulness of dropping Mere Addition in response to the Repugnant Conclusion, we shall not be concerned with those variable-value views that satisfy the strong version of Mere Addition (i.e., the version entailed by Total Utilitarianism), such as the theory examined in Sider's (1991). Instead, the aim is to examine those views that (unlike Average Utilitarianism) imply some weak instance of Mere Addition, without implying the strong version of Mere Addition.

We note also that a normative reason for excluding from our examination the view in Sider (1991) is that it implies what Arrhenius's (forthcoming) calls "The Very Anti Egalitarian Conclusion: For any perfectly equal population of at least two persons with positive welfare, there is a population which has the same number of people, lower average (and thus lower total) welfare and inequality, which is better." In fact, Sider himself rejects the view due to implications like this (Sider, 1991, 270).

[5] In what follows, we focus on *human* populations. This is not because we think that the welfare of animals is unimportant. But how precisely to integrate animal welfare into population ethics is far from evident. Therefore, for the sake of simplicity, we focus on human welfare and populations of people.

[6] If the reader finds our empirical assumptions implausible, then she can of course read our argument and conclusion as being *merely conditional* on these assumptions.

the future *must* be prosperous and populous; we only need the conditional claim that *if* the future is prosperous and populous, you would nevertheless support a mere addition of a happy present-day professor.

Our main observation is that, when combined with the above two assumptions, these Variable-Value views calibrate to be nearly totalist. So they imply countless instances of the Repugnant Conclusion. (By an "instance" of the Repugnant Conclusion, we mean the judgement that some particular population consisting only of lives that are barely worth living is better than some particular perfectly equal population of very well-off people.) Of course, these Variable-Value views do not imply the *qualitative* Repugnant Conclusion stated above — which holds for *all* populations of very well-off people. But the aforementioned implications, we argue, should nevertheless seem every bit as repugnant to those who oppose to the Repugnant Conclusion.[7] So, on the face of it, these results would seem problematic for any arguments that the Repugnant Conclusion requires us to reject Total Utilitarianism in favor of a Variable-Value view.

It might be worth providing some additional remarks to motivate our approach.[8] First, we assume that some (but perhaps not all) of those who are happy with giving up the traditional Mere Addition principle will nevertheless find it hard to reject some very weak instances of the principle. After all, we seem to have stronger reasons to think that a mere addition of a very well off person does not make the world worse than we have to think that a mere addition of a life barely worth living does not make the world worse.[9] Therefore, there is, we think, something to be gained from exploring what happens when we replace Mere Addition with a weaker principle that only applies to people who are

[7] In fact, according to the principle of "unrestricted instantiation" (Tännsjö, 2020), these implications *must* be seen as repugnant if the Repugnant Conclusion is to be an argument against Total Utilitarianism.

[8] Many thanks to [blinded] for making us see the need to address the motivation.

[9] If one is, say, a convinced total utilitarian, then one might nevertheless have equally strong (subjective) reason to think that a mere addition of a very well off person *could* make the world worse as one has to think that a mere addition of a life barely worth living *could* make the world worse. (Cf. Lenman, 2000,Greaves, 2016.)

very well-off.

Second, we think that valuable lessons can be learnt from exploring what population axiologies imply when calibrated to reasonable empirical assumptions, as opposed to merely exploring what these axiologies imply in theory. In particular, our finding that Variable-Value views have counterintuitive implications, given empirical assumptions that are plausible for our actual world population, provides a valuable lesson that is not learnt from simply learning that these axiologies have counterintuitive implications given assumptions about the world population that we take to be false. For that shows that Variable-Value views do not only have counterintuitive implications in hypothetical scenarios; they also have counterintuitive implications in empirically plausible scenarios.

2 Formal framework for population ethics

Our framework, terminology, and notation follow closely that of Asheim and Zuber (2014). Let $\mathbb{N}$ denote the set of natural numbers and $\mathbb{R}$ the set of real numbers, while $\mathbb{R}^+ \subset \mathbb{R}$ denotes the set of strictly positive real numbers. Let $\mathbf{X} = \bigcup_{n \in \mathbb{N}} \mathbb{R}^n$ denote the set of possible finite distributions of lifetime well-being. More formally, $\mathbf{X} = \bigcup_{n \in \mathbb{N}} \mathbb{R}^n$ is a set of of vectors of real numbers, where each number represents the lifetime well-being of some person. A generic such vector for a population of m people is denoted $\mathbf{x} = (x_1, ..., x_m)$, where x_i denotes the lifetime well-being of individual i. The size of the population given by $\mathbf{x}$ is denoted by $\mathcal{N}(\mathbf{x})$ (and will, as mentioned, always be finite). For any vector $\mathbf{x}$, we write the average lifetime well-being of its members as $\bar{x}$.

Let $\precsim$ on $\mathbf{X}$ denote a (weak) better-than relation on $\mathbf{X}$, such that for any $\mathbf{x}, \mathbf{y} \in \mathbf{X}$, $\mathbf{x} \precsim \mathbf{y}$ means that $\mathbf{y}$ is at least as good as $\mathbf{x}$. Throughout the discussion we shall assume that the better-than relation is transitive, reflexive, and complete,[10] which means that the relation

[10] Although the assumption of completeness is standard in the population economics literature, some population ethicists have made attempts to avoid the Repugnant Conclusion by relaxing it. (See e.g. attempt

generates a better-than *order*. The strict relation, $\prec$, and indifference, $\sim$, are respectively the asymmetric and symmetric counterparts of $\precsim$.

Built into our framework is an *anonymity* axiom, which holds that the "better-than relation" we study is invariant under permutations of the vectors in **X**. For instance, let **x**′ be the vector that results when the lifetime well-being of i and j in **x** are switched. Then the better-than relations that we shall consider are all indifferent between **x** and **x**′, that is, they deem these two distributions to be equally good. Intuitively, this means that it does not matter *who* receives what welfare; all that matters is how many people have each welfare level. This assumption rules out some person-affecting views.

For any $\mathbf{x} \in \mathbf{X}$ with m members, let $\mathbf{x}_{[]} = (x_{[1]}, ..., x_{[r]}, ..., x_{[m]})$ be the nondecreasing reordering of **x**. In other words, in $\mathbf{x}_{[]}$ the elements of **x** have been put in a nondecreasing order, such that for each rank $r \in \{1, ..., m\}$, $x_{[r]} \leq x_{[r+1]}$, meaning that individual with rank $r+1$ is at least as well off as individual with rank r. The anonymity assumption ensures that when two or more individuals are equally well-off, how they are ranked relative to each other does not affect the ranking of populations.

Finally, $(z)_n \in \mathbb{R}^n$ denotes the perfectly-equal distribution where all n individuals have lifetime well-being z. And let $(\mathbf{x}, (z)_n)$ denote distribution $\mathbf{x} \in \mathbf{X}$ with n added individuals that all have lifetime well-being z. When only one individual with well-being level y is added to **x**, we denote this by $(\mathbf{x}, y)$.

With this formalization, different axiological views, such as those discussed above, can be seen as different views about the structure of $\precsim$. This allows for convenient formal statements of the views and conditions we informally discussed in the last section. For instance, Total Utilitarianism can be formulated thus:

by Parfit, 2016; see also discussion in Arrhenius, 2016.) However, since the aim of this paper is to explore Variable-Value views which imply a complete better-than relation, the assumption of completeness is harmless here.

Total Utilitarianism (TU). *For any $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{X}$:*

$$\boldsymbol{x} \precsim \boldsymbol{y} \Leftrightarrow \sum_i x_i \leq \sum_i y_i$$

We can now also state the Repugnant Conclusion more formally:[11]

The Repugnant Conclusion (Formal version). *For all $y, z \in \mathbb{R}$, where $y > z > 0$, and for any $k \in \mathbb{N}$, there is a $n \in \mathbb{N}$ such that $(y)_k \prec (z)_n$.*

It is easy to verify that Total Utilitarianism, as formulated above, implies the Repugnant Conclusion.[12]

The Variable-Value views that we later discuss will be contrasted with Average Utilitarianism:

Average Utilitarianism (AU). *For any $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{X}$:*

$$\boldsymbol{x} \precsim \boldsymbol{y} \Leftrightarrow \bar{x} \leq \bar{y}$$

It can be easily verified that Average Utilitarianism does not imply the Repugnant Conclusion. However, Average Utilitarianism is well-known to violate the Mere Addition principle, which we can now formally state as:

Mere Addition (Formal version). *For any $\boldsymbol{x} \in \boldsymbol{X}$, and for any $z \in \mathbb{R}^+$, $\boldsymbol{x} \precsim (\boldsymbol{x}, z)$.*

Denying Mere Addition, for a complete ordering, is equivalent to entailing what we call the Anti-Natalist Conclusion:

[11] This formalization is slightly different from that of Blackorby et al. (2005), who require that $n > k$. See Spears and Budolfson (2021) for a discussion of heterogeneity in formalizations of the Repugnant Conclusion in the prior literature.

[12] Nebel (2022) has recently formulated a version of totalism that avoids the Repugnant Conclusion, by including a lexical threshold in the conception of individual welfare. As our aim here is not to discuss the extent to which Total Utilitarianism implies the Repugnant Conclusion — but rather the extent to which Variable-Value views imply the repugnant conclusions — we will not discuss Nebel's or other totalist views that avoid repugnance.

Anti-Natalist Conclusion. *There exists a $z \in \mathbb{R}^+$ and an $\boldsymbol{x} \in \boldsymbol{X}$ such that $(\boldsymbol{x}, z) \prec \boldsymbol{x}$.*

In the remainder of this paper, we examine a novel weakening of Mere Addition that we argue is highly plausible (and that we believe will seem plausible to even some of those who are willing to reject the strong Mere Addition principle). To state the principle, let us stipulate that there is well-being level beyond which lives at that level are excellent by the standards of 21st-century developed countries; and let $\mathbb{E} \subset \mathbb{R}^+$ be the set of well-being levels that are excellent by this same standard. For concreteness, we shall occasionally assume that a typical happy professor in a present-day developed country is at that level — we assume that most readers of this paper will have empirical familiarity with such a life.[13] And let $\mathbf{X}_R$ be the set of vectors that could realistically represent the lifetime well-being of the entire intertemporal (human) world population.[14] We can now finally state:

Weak Mere Addition. *For any $\boldsymbol{x} \in \boldsymbol{X}_R$, and for any $y \in \mathbb{E}$, $\boldsymbol{x} \precsim (\boldsymbol{x}, y)$.*

While denying Mere Addition, for a complete ordering, "only" implies accepting the Anti-Natalist Conclusion, denying Weak Mere Addition in addition implies accepting a Strong Anti-Natalist Conclusion:

Strong Anti-Natalist Conclusion. *There exists a well-being level $y \in \mathbb{E}$ and a population $\boldsymbol{z} \in \boldsymbol{X}_R$ such that $(\boldsymbol{z}, y) \prec \boldsymbol{z}$.*

We ourselves are sceptical of the Anti-Natalist Conclusion. But we think that there is even stronger reason to reject the Strong Anti-Natalist Conclusion — and we expect that this intutition will be widely shared. The latter is of course logically stronger than the

[13] Note that since $\mathbb{E}$ is a subset of $\mathbb{R}^+$, the concreteness assumption implies that a typical professor in a developed country, and more generally people in the top part of the current global well-being distribution, have lives worth living. While some may question this assumption (e.g., Benatar, 1997), we hope that it will strike most readers as innocent.

[14] The exact content of this set is not important for our argument. We merely introduce the set to emphasize that our argument does not require the use of any arbitrary world population, but only (what we take to be) empirically plausible populations.

former (since $\mathbb{E}$ is a strict subset of $\mathbb{R}^+$). Moreover, we think that, intuitively, lives that are excellent by the standards of 21st-century developed countries are *much* better than barely worth living. Some may question this intuition (e.g., Benatar, 1997), but we shall simply take it for granted. But then the Strong Anti-Natalist Conclusion will be correspondingly "much" stronger than the Anti-Natalist Conclusion.

3 What do we mean by Variable-Value axiologies?

Before discussing in detail the preiviously mentioned implications of Variable-Value axiologies, let us explain what types of axiologies we have in mind.

An infinity of population axiologies could have value vary, for instance, by whether population size is odd or even. Here we follow the population ethics literature which has understood the term "Variable-Value axiologies" to refer to a particular structure of well-behaved families of social welfare functions that are designed to respond to the tension between on the one hand satisfying (some version of) Mere Addition and on the other hand avoiding the Repugnant Conclusion. In his first paragraph on Variable-Value approaches, Arrhenius (forthcoming) summarizes: "These principles are sometimes called 'compromise theories' since a Variable-Value Principle can be said to be a compromise between Total and Average Utilitarianism. With small populations enjoying high welfare, a Variable Value Principle behaves like Total Utilitarianism and assigns most of the value to the total sum of welfare. For large populations with low welfare, the principle mimics Average Utilitarianism and assigns most of the value to average welfare" (p. 88). In the context of Ng's (1989) trilemma among Mere Addition, Non-Antiegalitarianism,[15] and avoiding the Repugnant Conclusion, we interpret the core of the Variable-Value idea to be

[15] Non-Antiegalitarianism says that a perfectly equal population with is better than a population with the same number of people, inequality, and lower average welfare. (See, e.g., Arrhenius 2000.)

a principled approach to rejecting Mere Addition in favor of the other two.[16] So, we are interested in complete, transitive, and anonymous (recall from the last section) families of social welfare functions that:

- are well-behaved in the sense of satisfying (*ex post*) Pareto, Extended Egalitarian Dominance, Non-Antiegalitarianism, and other non-controversial principles in the literature;[17]
- avoid the Repugnant Conclusion by rejecting Mere Addition;
- compromise between TU for small populations and a populations-size-insensitive alternative for large populations; and
- can be calibrated by a parameter that quantifies the distance between TU and the size-insensitive alternative as two convergent endpoints.

For clarity, we restrict our attention to the very large set of axiologies that, for perfectly equal populations of size n, reduce to $g(\bar{u})f(n)$ for some increasing g and some concave, bounded f. We ask quantitative questions about f. In particular, we use a calibration method (informally described in the next section), which determines how quickly f approaches its bound as population size increases. A totalist f would be linear. We ask quantitatively: how totalist does a Variable-Value view have to be to be plausible? The answer, we argue, turns out to be rather totalist.

Two well-known examples of Variable-Value axiologies that satisfy all of the above discussed properties are:

[16] For this reason, we disregard Sider's (1991) example of Geometrism, which rejects Non-Antiegalitarianism, and any other candidate axiology that does so; we are unaware of any author in the population ethics literature (including Sider in fact) who defends any Antiegalitarian Variable-Value proposal as plausibly the true population axiology.

[17] Since we will not make direct use of these principles, it suffices to define them informally. Pareto says that if every person is at least as well-off in population A as in population B, then A is at least as good as B. Extended Egalitarian Dominance says that if there is perfect equality in population A which is of greater size than population B, and every person in A has higher positive welfare than every person in B, then A is better than B. (See, e.g., Arrhenius forthcoming.) Non-Antiegalitarianism is defined in footnote 15.

- Number-Dampened Generalized Utilitarianism (with an appropriate choice of functional form): a calibrated mix of TU and AU; and
- Rank-Discounted Generalized Utilitarianism: a calibrated mix of TU and leximin.

We will investigate these views in detail in sections 5.1 and 5.2 after first informally presenting our general argument in the next section.

4 Our general argument, informally

To clarify: it is not our own view, the authors', that the Repugnant Conclusion is indeed repugnant or must be avoided by the correct theory of population ethics (Zuber et al., 2021). But many population ethicists have held that view, and we respond to them here. The population ethics literature is unclear on what exactly it is that is supposed to make the Repugnant Conclusion repugnant. We do not intend to take a stance on this question. In the rest of this paper, we will informally describe a judgement as being "repugnant" when it involves preferring a much-larger, much-worse-off perfectly-equal population over a large but much-smaller, much-better-off perfectly equal population. We mean "repugnant in the sense of the repugnant conclusion." Perhaps no such judgement is actually repugnant at all, but if repugnance is to be found there, we propose that it needs to be found *consistently* in any similarly "repugnant" judgements (and hereafter without the quotation marks).

But let's say that you do find the Repugnant Conclusion repugnant; that you find Non-Antiegalitarianism utterly unrejectable; and that you therefore abandon Mere Addition in favor of a Variable-Value axiology. Have you then escaped repugnance? We suggest that it depends upon the calibration of the partially-totalist mix.

Here is why: We conjecture that you, the reader, have a strong conviction that the mere deletion of your life, or mine, or that of any other very-well-off person by present-day standards, would not make the intertemporal human population better (even if it could be accomplished by magic, the "mere" deletion having no effect on the welfare of anyone else) than the actual intertemporal human population in which this very-well-off person by present-day standards in fact exists. A great life judged by the standards of our times does not in and of itself make the world worse.

A natural question is why we are emphasizing our times, your life, or mine. We are evaluating the actual intertemporal population, after all, from the point of view of timeless population ethics. From the point of view of the universe, why is today special? The answer — as briefly mentioned in the introduction — is that of course it is not. But we are writing to you, the reader. And you, the reader, live in the present and come to our argument with beliefs and intuitions and, in the case of your life, hedonic experiences. We think lives like yours and ours are easy, readily available, and informative test cases for the present purposes. In fact, test cases like these seem to illustrate quite clearly that there is at least one population and one welfare level such that we do not believe adding that welfare level to that population makes matters worse: namely, adding a good life, like a professor's in our times, to the actual intertemporal population.

So why is agreement on this case important? Because it disciplines the calibration of the mix of any Variable-Value axiology, the mix between totalism, on the one hand, and, on the other hand, the number-insensitive counterpart (such as AU or leximin). We propose that two facts are probably true of the actual intertemporal population, and even if they are not true, we propose that if we assume them as hypothetically true, you would still find that the mere addition of a very-well-off person by present-day standards would not make the world worse. The two facts are:

- The future is vast: the actual intertemporal population is enormous.

- The future is splendid: the actual intertemporal population is full of lives much better than ours, i.e., much better than that of even a happy present-day professor.

These facts, plus our judgement about the mere addition of very-well-off person by present-day standards, bound our calibration of quantitatively how non-totalist a Variable-Value axiology can be. In steps:

1. Because the population is enormous, we will be making decisions like an averagist (or otherwise like a non-totalist, depending on the details of the particular Variable-Value mix), unless the tuning parameter is calibrated to move away from totalism only very slowly.

2. Because someone like us is relatively badly-off compared to the splendid full distribution, adding such a person pulls down the average, disadvantages the lexical ladder, or otherwise looks undesirable to the non-totalist part of the axiological mix.

3. Because we judge that adding the person is nevertheless not a worsening, it must be the case that the tuning parameter is calibrated to move away from totalism only very slowly, so that the totalist benefits of the addition outweigh the non-totalist costs of adding a relatively-badly-off person.

And this brings us to the implications for the Repugnant Conclusion. If the tuning parameter is such that the Variable-Value axiology is, in the end, calibrated to be quite close to totalism, then it will often agree with totalism about how to rank populations. And that means that it will make many repugnant judgements where it prefers larger, worse-off populations to smaller, better-off ones, agreeing with totalism even in many quantitatively extreme cases. The universally-quantified Repugnant Conclusion is escaped, but *repugnance* is not, whatever that is. So the spectre of repugnance seems hardly a reason to choose a Variable-Value approach. The next two sections make this general argument quantitatively precise.

Some readers will have recognized that our argument presents an application of a familiar logic in decision theory: calibration of variable-value objective functions to reveal tensions between intuitions for large-quantity decisions and intuitions for small-quantity decisions. The leading result in this literature is Rabin's (2000) celebrated argument about expected utility theory. Formally, we merely extend Rabin's argument about choice under risk to analogous functional forms in population ethics.

Rabin established that an expected utility maximizer can only be moderately risk averse when relatively small sums of money are involved — e.g. always turning down 50-50 gambles between losing \$100 and winning \$105 — if she is extremely risk averse when larger sums of money are involved — e.g. turning down 50-50 gambles between losing \$2,000 and winning any (including infinite) amount of money. So, the lesson of Rabin's argument is that an expected utility maximizer is either surprisingly risk averse when stakes are large or surprisingly risk neutral when stakes are small. Similarly, the lesson of our calibration exercise is that Variable-Value views are either surprisingly totalist or surprisingly strongly anti-natalist, when applied to a vast and splendid future.[18]

5 Formal arguments using two Variable-Value axiologies

5.1 Number-Dampened Generalized Utilitarianism: A bounded case

The first view in the Variable-Value family that we shall consider can be stated as follows:

Number-Dampened Generalized Utilitarianism (NDGU). *There is a concave (and increas-*

[18] Nebel and Stefánsson (forthcoming) apply a similar logic to inequality averse views about how to order populations of a fixed size, in particular, to Prioritarianism and Generalized-Gini Egalitarianism, and find that such views can only be moderately inequality averse when small differences in welfare are at stake if they are extremely inequality averse when larger welfare differences are at stake.

ing) real-valued function f such that for any $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{X}$:

$$\boldsymbol{x} \precsim \boldsymbol{y} \Leftrightarrow \bar{x}\, f(\mathcal{N}(\boldsymbol{x})) \leq \bar{y}\, f(\mathcal{N}(\boldsymbol{y}))$$

To our knowledge, this family was introduced to the literature implicitly in the diagrams of Hurka (1983). The concavity of f means that NDGU reduces the value of additions to the population as population size grows. Moreover, if (but only if) f is bounded, then NDGU avoids the universally-quantified Repugnant Conclusion.

For illustrative purposes, we assume in our example that

$$f(\mathcal{N}(\mathbf{x})) = 1 - e^{-\frac{\mathcal{N}(\mathbf{x})}{\alpha}}$$

for some $\alpha > 0$. The parameter α is the crucial parameter of calibration that tunes how quickly, as population size increases, f transitions from a TU-like gradient for small populations to an AU-like nearly constant 1 for large populations. Larger α is more totalist;[19] α nearer to zero is more averagist.

Our judgement that it is not worse to add today's happy professor puts a lower bound on α. Assume that you would agree that a mere addition of today's happy professor would make the world better even if you knew that optimists like Ord (2020) were correct,[20] such that, say, the human population would have at least 10^{12} people in it ($1,000$ generations as large as ours) and that the average wellbeing would be at least 5 times that of today's happy professor. Because today's happy professor pulls down the average, α must be large (that is, towards totalism) if adding this person is an improvement. How large? For these hypothetical numbers: at least 2.345×10^{12}. For the argument that follows, it is

[19] To see this, consider multiplying a population size by λ. How does $\frac{f(\lambda n)}{f(n)}$ behave as α gets large? Using L'Hôpital's rule, the limit of the ratio is described by $\lambda e^{\frac{n(1-\lambda)}{\alpha}}$, which goes to λ for any n as α gets large.

[20] Actually, it might be more accurate to call Ord a "conditional optimist": he is optimistic that the future will be long and prosperous *if* we manage to avoid the looming existential catastrophes.

not essential that you believe that the actual size and wellbeing of the full intertemporal population is exactly as we just assumed. Instead, it suffices, for our purposes, that you accept that these assumptions *could* be true, and that you would support a "mere addition" of the happy professor even if these assumptions were true.

Now that α has been bounded, we can generate repugnant judgements. For example, for any high quality of life:

- a population with 10 billion people[21] in which everyone enjoyed such a high-quality life

would be worse than

- a population in which 4.5 trillion people live much worse lives, each living a life with only half of a percent of the wellbeing of those in the smaller population.[22]

Formally, that is:

$$1 - e^{-\frac{10^{10}}{2.345\times 10^{12}}} < 0.005 \times \left(1 - e^{-\frac{4.5\times 10^{12}}{2.345\times 10^{12}}}\right).$$

Now, population ethicists' intuitive reactions to the above example will presumably vary. Our claim is simply that whatever repugnance is supposed to be found in the Repugnant Conclusion is surely found in the above judgment too. So NDGU of this form does not escape repugnance, for better or worse, if it is calibrated such that it supports a mere addition of today's happy professor.

5.2 Rank-Discounted Generalized Utilitarianism

The second Variable-Value view that we shall consider can be stated as follows:[23]

[21] Is this enough for an instance of repugnance? Ten billion is exactly the population size that Parfit used for the small population in introducing the Repugnant Conclusion. It is 10 generations each as populous as the entire world at a point in the 19th century.

[22] Here we assume that the measure of wellbeing is zero-normalised around the point at which life becomes worth living (as is traditionally done in population ethics and population economics), which is why it is meaningful to talk of percentages of wellbeing.

[23] Recall that $\mathbf{x}_{[]} = (x_{[1]}, ..., x_{[r]}, ..., x_{[m]})$ is the nondecreasing reordering of $\mathbf{x}$.

Rank-Discounted Generalized Utilitarianism (RDGU). *There is a* $\beta \in (0,1)$ *such that for any* $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{X}$*:*

$$\boldsymbol{x} \precsim \boldsymbol{y} \Leftrightarrow \sum_r \beta^r g\left(x_{[r]}\right) \leq \sum_r \beta^r g\left(y_{[r]}\right)$$

where g *is increasing and weakly concave.*

This view was introduced and characterized by Asheim and Zuber (2014). It avoids the (universally quantified) Repugnant Conclusion because $\beta^1 + \beta^2 + \beta^3$... is a convergent series, which ensures that the aggregated value of a perfectly-equal population is bounded and remains finite, no matter how large it becomes. Therefore, if k, in our formal statement of the Repugnant Conclusion, is sufficiently large, and if y is sufficiently larger than z, then there is *no* n such that, by RDGU, $(y)_k \prec (z)_n$. In other words, the (universally quantified) Repugnant Conclusion does not follow from RDGU. But, if β is sufficiently close to 1, then even large y could be part of an instance of the Repugnant Conclusion with a z that is small enough to capture the purportedly repugnant features of the Repugnant Conclusion.

Like α tuned NDGU, β tunes RDGU. Asheim and Zuber (2014) prove that as β approaches 1, RDGU approaches totalism,[24] and as β approaches 0, RDGU approaches a variable-population version of leximin.

RDGU does not satisfy the strong Mere Addition principle that Total Utilitarianism entails. This is because adding a life to a population lowers the weights of any otherwise-existing better-off lives, which may worsen the population by more than the additional life improves it. However, RDGU must satisfy *some* Weak Mere Addition principle, since $\beta > 0$, which means that *some* mere additions are valuable. And, in fact, the closer β is to 1, the closer RDGU comes to implying the strong Mere Addition principle, in the sense of implying stronger instances of Mere Addition. We want to examine what RDGU implies if we assume that it satisfies particular (very plausible) instances of Weak Mere Addition.

[24] In fact RDGU approaches Critical-Level Generalized Utilitarianism, but for simplicity we are ignoring non-zero critical levels here.

Let us calibrate again to a plausibly quantified version of our actual population, again following futurists like Ord (2020) about the plausible size and value of the long-term population. In evaluating whether to merely add a person, RDGU can ignore everyone worse off than that person. Assume that, all in all, there eventually will have been 10^{12} people better-off than today's happy professor by at least a factor of 5 (after g-transformation, if g is concave rather than linear). Then β is bounded by the following inequality:

$$g(\text{happy professor}) + \beta \sum_{r=0}^{10^{12}-1} \left(\beta^r \times 5g(\text{happy professor})\right) > \sum_{r=0}^{10^{12}-1} \left(\beta^r \times 5g(\text{happy professor})\right),$$

factoring out the $g(\text{happy professor})$ terms:

$$1 + 5\beta \sum_{r=0}^{10^{12}-1} \beta^r > 5 \sum_{r=0}^{10^{12}-1} \beta^r,$$

which implies $\beta > 0.999999999999778$. Recall that as β approaches 1, RDGU approaches totalism. As before, you need not believe that the future will take be as long and splendid as we just assumed; you need only agree that today's happy professor is worth adding even if it is.

As in the case of NDGU, a calibrated version of RDGU generates repugnant judgements. For example, for any high quality of life:

- a population with 10 billion people in which everyone enjoyed such a high-quality life

would be worse than

- a population in which 2.65 trillion people live much worse lives, each living a life with only half of a percent of the wellbeing (or g-transformed wellbeing, if g is not linear), as the lives in the smaller population.

So, it turns out that just like NDGU, RDGU needs to be calibrated to be rather totalist to avoid being implausibly anti-natalist about adding to the assumed population people like you, the reader, and us, the authors. Furthermore, such a calibrated version of RDGU will make judgements that, we suspect, those who criticise Total Utilitarianism for its purported repugnance will find hard to accept.

6 Lesson and concluding remarks

Recall that the intuitive appeal of Variable-Values views was supposed to be that they could avoid the Repugnant Conclusion while satisfying at least some weak instance of the Mere Addition principle. We have now seen, however, that if these views satisfy what we take to be a very plausible, and certainly weak, instance of Mere Addition, and if in addition we make plausible empirical assumptions about the intertemporal world population, then these Variable-Value views have implications that, we suggest, those who oppose Total Utilitarianism due to the Repugnant Conclusion will find repugnant.[25]

What should we conclude from our results? Most narrowly, a lesson of our results is that when calibrated to the real world — that is, the actual world population and what we think are plausible empirical assumptions about the future population — leading Variable-Value views would substantially agree with Totalist views on how to rank most policies that affect a relatively small number of people. Moreover, if we assume that policy choices typically affect only a relatively small number of people — that is, small in relation to the total intertemporal world population — then the implication is that these Variable-Value views and Totalist views typically recommend the exact same courses of action (especially if the menu of possible options is coarse and bounded). The only escape would be for these Variable-Value views to be strikingly anti-natalist, such that they do not even satisfy

[25] In fact, given Tännsjö's principle of unrestricted instantiation (recall fn. 7), these implications *must* be deemed repugnant if the Repugnant Conclusion is to be used as an argument against Total Utilitarianism.

weak instances of Mere Addition that would involve the lives of well-off readers of this paper.

More broadly, these results teach us something about the effort to avoid the Repugnant Conclusion. Population ethicists have long understood that it is impossible to escape all implications that initially seem undesirable or unintuitive. But some unintuitive claims are true. The quantitative results of this paper are unintuitive, too. This paper adds to a growing recent literature — including Spears and Budolfson (2021) on additions to an unaffected population and Arrhenius and Stefánsson (2018) on risky choice between uncertain populations — that finds repugnant conclusions even under approaches to population ethics commonly understood to avoid repugnance. Collectively, these results suggest that the effort to avoid the Repugnant Conclusion has, in some ways, hinged on questionable features of the formalization of repugnance (such as the features that exclude the cases documented in this paper); that some of this effort may therefore be misplaced; and that perhaps avoidance of the Repugnant Conclusion should not be a core goal of population ethics research. We therefore suggest that our arguments here support the position in Zuber et al. (2021), a recent statement of agreement by many authors from diverse perspectives who suggest that avoiding the Repugnant Conclusion has been overemphasized by population ethics research.

References

Arrhenius, Gustaf. 2000. "An impossibility theorem for welfarist axiologies." *Economics & Philosophy*, 16(2): 247–266.

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

Arrhenius, Gustaf and H Orri Stefánsson. 2018. "Population ethics under risk." working paper, IFFS.

Arrhenius, Gustaf. forthcoming. *Population Ethics: The Challege of Future Generations*: Oxford University Press.

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

Benatar, David. 1997. "Why It Is Better Never to Come Into Existence." *American Philosophical Quarterly*, 34(3): 345–355.

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

Greaves, Hilary. 2016. "Cluelessness." *Proceedings of the Aristotelian Society*, 116(3): 311–339.

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

Lenman, James. 2000. "Consequentialism and Cluelessness." *Philosophy and Public Affairs*, 29(4): 342–370.

Nebel, Jacob M.. 2022. "Totalism Without Repugnance." in Jeff McMahan, Tim Campbell, James Goodrich, and Ketan Ramakrishnan eds. *Ethics and Existence: The Legacy of Derek Parfit*: Oxford: Oxford University Press: 200–231.

Nebel, Jacob M. and H. Orri Stefánsson. forthcoming. "Calibration Dilemmas in the Ethics of Distribution." *Economics and Philosophy*.

Ng, Yew-Kwang. 1989. "What Should We Do About Future Generations?: Impossibility of Parfit's Theory X." *Economics & Philosophy*, 5(2): 235–253.

Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity*: Hachette Books.

Parfit, Derek. 1984. *Reasons and Persons*: Oxford.

——— 2016. "Can we avoid the repugnant conclusion?" *Theoria*, 82(2): 110–127.

Pivato, Marcus. 2020. "Rank-additive population ethics." *Economic Theory*, 69(4): 861–918.

Rabin, Matthew. 2000. "Risk Aversion and Expected-utility Theory: A Calibration Theorem." *Econometrica*, 68(5): 1281–1292.

Sider, Theodore R. 1991. "Might theory X be a theory of diminishing marginal value?" *Analysis*, 51(4): 265–271.

Spears, Dean and Mark Budolfson. 2021. "Repugnant Conclusions." working paper; prior version is IZA Discussion Paper 12668.

Tännsjö, Torbjörn. 2020. "Why Derek Parfit had reasons to accept the Repugnant Conclusion." *Utilitas*: 1–11.

Zuber, Stéphane, Dean Spears, Johan Gustafsson, Mark Budolfson, and others. 2021. "What should we agree on about the repugnant conclusion?" *Utilitas*: 1–5.