Calibrating variable-value population ethics*

Dean Spears† H. Orri Stefánsson‡

February 24, 2021

Abstract

Variable-Value axiologies propose solutions to the challenges of population ethics. 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 while assuming: 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 Variable-Value views imply conclusions that should seem repugnant to anyone who opposes Total Utilitarianism 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, these calibrations teach us something about the effort to avoid the Repugnant Conclusion. Our results join a recent literature arguing that prior efforts to avoid the Repugnant Conclusion hinge on inessential features of the formalization of repugnance. Some of this effort may therefore be misplaced.

*We are grateful for very useful written comments from Jake Nebel and Christian Tarsney, and for helpful suggestions and questions from the audience at Climate Ethics Workshop, Institute for Futures Studies in Stockholm, April 2020.

†**Email: dean@riceinstitute.org.** Economics Department and Population Research Center, University of Texas at Austin; Economics and Planning Unit, Indian Statistical Institute - Delhi Centre; IZA; Institute for Future Studies, Stockholm; r.i.c.e. (www.riceinstitute.org). Although this paper received no specific funding, Spears' research is supported by NICHD grants K01HD098313 and P2CHD042849. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

‡**Email: orri.stefansson@philosophy.su.se.** Department of Philosophy, Stockholm University; Swedish Collegium for Advanced Study, Uppsala; Institute for Future Studies, Stockholm. Stefánsson's research is supported by Riksbankens Jubileumsfond through a *Pro Futura Scientia* fellowship. Funding from Riksbankens Jubileumsfond to the Climate Ethics program is also gratefully acknowledged.

1 Introduction

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

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.

[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.

Somebody who abandons Mere Addition argues that merely 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 professor living in a developed country in 2020. Surely by adding a person like that to any population, without thereby making anyone else worse off, we have made the population better? 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 makes the resulting population worse, according to Average Utilitarianism. This 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 (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*.[2] These views are intended to avoid the Repugnant Conclusion while capturing the intuition that adding a well-off person to a small population 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.[3] These formalizations and the ensuing analysis has focused on *qualitative* properties of Variable-

[2]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.

[3]Examples include Ng (1989), Asheim and Zuber (2014), and Pivato (2020)

Value population ethics: with which *axioms* do these proposals comply? However, there has not been a similar focus on the *quantitative* implications of these 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 weakenings of the Mere Addition principle these views can accommodate without implying instances of the Repugnant Conclusion.[4]

Our aim in this paper is to fill the above gap in the population ethics literature. In particular, we shall assume some very weak instances of Mere Addition and calibrate what Variable-Value axiologies, 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 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 is higher than the average welfare of the world population up to 2020.[5]

Our main observation is that, when combined with the above two assumptions,

[4]Our aim is not to examine *all* Variable-Value views. In particular, 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]If the reader finds this empirical assumptions implausible, then she can of course read our argument and conclusion as being *merely conditional* on these assumptions.

Variable-Value axiologies 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, they do not imply the *qualitative* Repugnant Conclusion stated above — which holds for *all* populations of very well-off people. But these implications, we argue, should nevertheless seem every bit as repugnant to those who oppose to the Repugnant Conclusion.[6]

It might be worth providing some additional remarks to motivate our methodology.[7] First, we assume that even 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. 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 very well-off.

Second, we think that valuable lessons can be learnt from exploring what population axiologies imply given 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 we accept 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 haver counterintuitive implications in empirically plausible scenarios.

[6] 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.

[7] Many thanks to Christian Tarsney for making us see the need to address the motivation.

We proceed as follows. In Section 2 we lay out the formal framework of the paper, which allows us to state more formally the views and conditions we informally describe above, and introduce the reader to calibration methods in decision theory. Then, in Section 3, we use such methods to examine what two prominent Variable-Value views imply when they have been calibrated to the current world population and what we take to be reasonable assumptions about the future population. In Section 4 we use the same methods to present a more general result, that is, a result for all Variable-Value views that do not satisfy the strong version of Mere Addition. Finally, in Section 5 we ask what the upshot of our arguments is for population ethics and in particular for the focus in the population ethics literature on avoiding the Repugnant Conclusion.

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. 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 $\bar{x}$. So, $\bar{x}$ should be interpreted as the average lifetime welfare of people given by $\mathbf{x}$.

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 $\mathbf{X}$. So, for instance, let $\mathbf{x}'$ be the vector that results when the lifetime well-being of i and j in $\mathbf{x}$ are switched. Then the better-than relations that we shall consider are all indifferent between $\mathbf{x}$ and $\mathbf{x}'$,

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 $\mathbf{x}$. In other words, in $\mathbf{x}_{[]}$ the elements of $\mathbf{x}$ 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.

Let $(z)_n \in \mathbb{R}^n$ denote 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 $\mathbf{x}$, we denote this by $(\mathbf{x}, y)$.

Finally, $\precsim$ on $\mathbf{X}$ denotes 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,[8] which means that the relation generates a better-than *order*. The strict relation, $\prec$, and indifference, $\sim$, are respectively the asymmetric and symmetric counterparts of $\precsim$.

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:

[8] 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 by Parfit, 2016 and response by Arrhenius, 2016.) But, to keep things relatively manageable, we shall nevertheless in this paper assume completeness.

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:[9]

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 implies the Repugnant Conclusion.[10]

The Variable-Value views that we later discuss will be contrasted with both Average Utilitarianism (to be formally defined) and Critical-Level Generalized Utilitarianism (CLGU).[11] CLGU is a family of generalized total utilitarian views, which include for instance Critical-Level Utilitarianism, Prioritarianism, and Total Utilitarianism as special cases. The general view can be stated thus:

Critical-Level Generalized Utilitarianism (CLGU). *For any* $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{X}$*:*

$$\boldsymbol{x} \precsim \boldsymbol{y} \Leftrightarrow \sum_i \left[g\left(x_i\right) - g(c)\right] \leq \sum_i \left[g\left(y_i\right) - g(c)\right]$$

where g *is non-decreasing and non-convex, and* $c \geq 0$ *is the critical level at wich adding a new life becomes a social improvement.*

Some views in this CLGU family avoid the standard formalization of the Repugnant

[9] 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.

[10] Nebel (forthcoming) 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 Conclusion — we will not discuss Nebel's or other totalist views that avoid repugnance.

[11] CLGU was introduced by Blackorby and Donaldson (1984) and subsequently explored in depth by Blackorby et al. (2005) and Bossert (2017).

Conclusion, at the cost of entailing another counterintuitive result, such as Arrhenius's (2000) family of sadistic conclusions (Franz and Spears, 2020). CLGU is a significant family for our purposes because it is fully additively separable. That is, a CLGU better-than relation satisfies same-number, different-number, and existence independence axioms, always with a constant critical level (Blackorby et al., 2005). So, because our calibration results depend on additive separability being violated, no CLGU view is subject to the calibration arguments of our paper.

Average Utilitarianism can now simply be stated as:

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 also 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* $z \in \mathbb{R}$ *such that* $z > 0$, $\boldsymbol{x} \precsim (\boldsymbol{x}, z)$.

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

Anti-Natalist Conclusion. *There exists a* $z \in \mathbb{R}$*, where* $z > 0$*, and an* $\mathbf{x} \in \boldsymbol{X}$ *such that* $(\mathbf{x}, z) \prec \mathbf{x}$.

In the remainder of this paper, we examine a novel weakening of Mere Addition that we argue is highly plausible. 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, let's set that level at 97.5th percentile of lifetime well-being in

our current global population. To make things even more concrete, we shall occasionally assume that a typical professor in a developed country is at that level. And let $\mathbf{X}_R$ be the set of vectors that (we think) could realistically represent the lifetime well-being of the intertemporal world population. 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* $\mathbf{z} \in \boldsymbol{X}_R$ *such that* $(\mathbf{z}, y) \prec \mathbf{z}$*.*

While we ourselves are sceptical of the Anti-Natalist Conclusion, we think that there is even stronger reason to reject the Strong Anti-Natalist Conclusion.

As we show in the next two sections, however, unless Variable-Value views imply the Strong Anti-Natalist Conclusion, when these views have been calibrated to plausible empirical assumptions about the size and welfare of the future world population, they must imply many instances of the Repugnant Conclusion. This follows from our novel 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 maximiser is either surprisingly risk averse when stakes are large or surprisingly risk neutral when stakes are small.[12]

Our calibration result is that Variable-Value views are surprisingly anti-natalist when few extra lives are at stake (more specifically, they entail the Strong Anti-Natalist Condition) unless they are surprisingly totalist when more lives are at stake. (By "totalist" we mean making choices that should be found objectionable by those who reject Total Utilitarianism because of the Repugnant Conclusion.In other words, our result is that Variable-Value views can only accommodate an extremely weakened Mere Addition principle if they also imply countless instances of the Repugnant Conclusion.

3 Two prominent examples

This section turns to two prominent examples of Variable-Value population axiologies. Both of these are well-known in the literature to avoid traditional formalizations of the Repugnant Conclusion. We ask what these views recommend in repugnance-type tradeoffs between large and small populations, once calibrated to satisfy Weak Mere Addition and facts or predictions about the world population. Because they are algebraically tractable and well-studied, it is instructive to see why these particular versions of Variable-Value views imply instances of the repugnant conclusions, before considering (in the next section) a more general argument that applies to all Variable-Value views that violate the strong Mere Addition principle that Total Utilitarianism entails.

[12]Nebel and Stefánsson (2020) apply a similar logic to inequality averse views about how to order populations of a fixed size, in particular, to Prioritarianism and Rank-Discounted Utilitarianism, and find that such views can only be moderately inequality averse when small differences in welfare are at stake — e.g. preferring that everyone is equally well off at level w to half the population being at level $w - 0.9$ while the other half is at level $w + 1$ — if they are extremely inequality averse when larger welfare differences are at stake.

3.1 Number-dampened generalized utilitarianism

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* $\alpha \in (0, 1)$ *such that for any* $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{X}$*:*

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

NDGU reduces the value of additions to the population as population size grows. One way to see this, and connect it to our formulation of the Repugnant Conclusion, is that $(y)_k \prec (z)_n$ if $\frac{y}{z} > \left(\frac{k}{n}\right)^{\alpha}$. From this we can see that as α approaches 0, the ratio between y and z needed for NDGU to imply that $(z)_n \precsim (y)_k$ becomes smaller; which is just another way to say that the closer α comes to 0, the closer NDGU gets to Average Utilitarianism, and the fewer instances of the Repugnant Conclusion it implies. However, because $\alpha > 0$, NDGU will imply some instances of the Repugnant Conclusion, in the sense that for *some intuitively large* difference between y and z, and for any k, there will be an n such that $(y)_k \prec (z)_n$. Moreover, this illustrates that while NDGU violates the strong version of Mere Addition that Total Utilitarianism implies (since $\alpha < 1$), it does imply some weak instances of Mere Addition. Our aim now is to explore what happens when we assume that it satisfies some particular (very plausible) instances of Weak Mere Addition.

This algebraic formulation applies a particular, one-parameter family of functional forms to the concave-transformation proposal of Hurka (1983) and Ng (1989). Here we expand our historical horizons: The relevant fact is that adding a well-off (by today's standards) professor lowers average lifetime well-being because there will be many future people who will be even better off. In making this assumption, we follow recent literature on the possible long-term human future, such as by Greaves and MacAskill (2019) and Ord (2020). If future lives are to be so many and so good, then α must be high if the addition of a well-off professor is not to be a worsening.

If there will be 2×10^{12} people overall, for example, and if the average lifetime well-being will be three times as high as that of the modern professor's, then that implies an α of at least 0.67. If the future is even better than that and the average person will be five times as well-off as our professor, then we have an α of 0.80 — greater because today's extra happy professor pulls the intertemporal average down by more.

These, too, imply repugnant-like quantitative consequences. For instance, $\alpha = 0.67$ implies that a population of 31,695,627 people with lifetime well-being of 0.1 is better than a population of 1,000 people with lifetime well-being of 100. And $\alpha = 0.80$ implies that a population of 5,639,614 people with lifetime well-being of 0.1 is better than a population of 1,000 people with lifetime well-being of 100. [13] Both of these better-than judgements (i.e., those entailed by $\alpha = 0.67$ and $\alpha = 0.80$) should, we contend, be found repugnant by those who oppose Total Utilitarianism due to the Repugnant Conclusion.

3.2 Rank-discounted generalized utilitarianism

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

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.*

A version of this view is defended by, for instance, 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 a perfectly-equal population remains finite, no matter how large it becomes. Therefore, if k, in our formal statement of the Repugnant

[13] The reason why the number of people with 0.1 lifetime well-being that is needed to outweigh 1,000 people with lifetime well-being of 100 is smaller in the second case, is that as α becomes larger, the resulting axiology becomes closer to a totalist one.

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 qualitative (intuitively repugnant) features of the Repugnant Conclusion.

RDGU does not satisfy the strong Mere Addition principle that Total Utilitarianism entails. This is because adding a life lowers the weights of any otherwise-existing higher-utility lives, which may reduce social welfare by more than the additional life increases 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.

To that end, we assume that a typical professor in the developed world is at the 97.5th percentile of lifetime well-being in our current world population,[14] so only 2.5 percent of people are better-off. That is around 182 million people. Under RDGU, adding such a professor reduces the weight on those 182 million very well-off people. The implication is that, if adding such a professor is not a worsening, β must be very close to 1. In particular, β must be greater than 0.99999995 if the 182 million people are all *no more than 1.0001 times as well-off* as our well-off professor, and even closer to one if the better-off people are even better-off than that (which is of course more realistic). β reaches 0.999999999 if the better-off people are a little more than six times as well-off as the professor, for example.[15]

Now consider what these high βs imply for repugnant-like tradeoffs. How many

[14]For simplicity we here ignore past and future generations, but the large number of future people who would, we assume, be better off than today's happy professors only increases the force of this argument.

[15]Note that our argument can accommodate a positive critical level and a prioritarian transformation: simply subtract the critical level and/or do the prioritarian transformation before the rank-based weighting.

people each with a life at 0.01 would be needed to be better than a population with 100 people at 100? The answer ranges from 1,025,864 for $\beta = 0.99999995$ down to just above one million as β becomes closer to 1.[16,17] But those who believes that the Repugnant Conclusion must be avoided at all theoretical cost will presumably not be happy to have 1.026 million people at 0.01 instead of 100 people at 100. And yet, such is the consequence of choosing RDGU and maintaining that creating the happy professor is not a worsening given our actual world population.

4 A more general argument

The quantitative results of Section 3 depended upon two specific functional forms. Here we present a more general argument which applies to any variable-value goodness function of the form for any $\mathbf{u} \in \mathbf{X}$: $W(\mathbf{u}) = \bar{u} \times f(\mathcal{N}(\mathbf{u}))$, where f is positive, increasing, and concave.[18] Our arguments would extend readily to cases where $\bar{u}$ were replaced by a more general "equally-distributed equivalent" (Atkinson, 1970), because the populations in the standard Repugnant Conclusion are perfectly equal. We however use the arithmetic mean for simplicity. NDGU is the special case where $f(\mathcal{N}) = \mathcal{N}^{\alpha}$; RDGU is the special case where $f(\mathcal{N}) = \frac{1-\beta^{\mathcal{N}}}{1-\beta}$, for perfectly-equal populations. Notice that the algebra of $\bar{u} \times f(\mathcal{N}(\mathbf{u}))$ resembles the algebra of (risk averse) expected utility — an affine probability times a concave von Neumann-Morgenstern transformation — which is why our argument follows from Rabin's (2000).

Figure 1 illustrates the argument. First, fix a lifetime well-being level for a weak mere

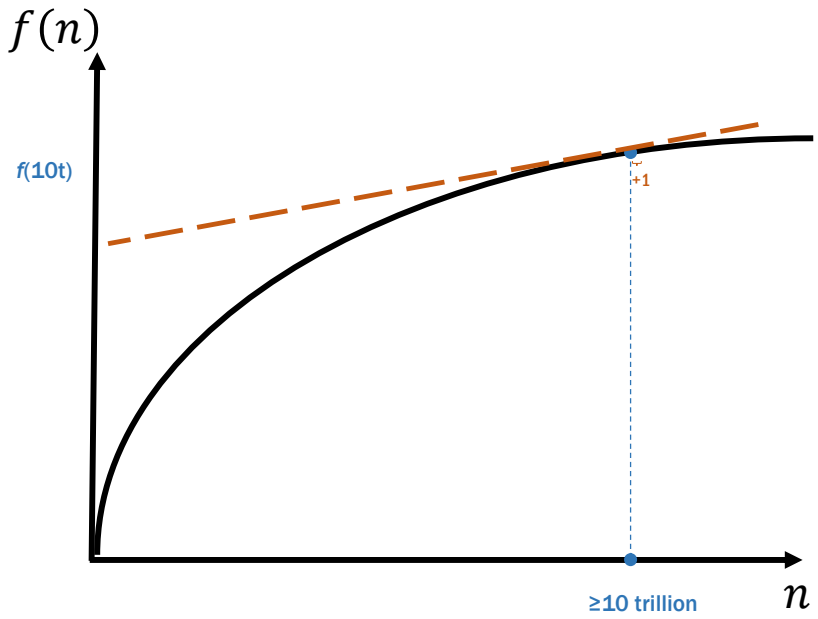
[16]The sum of a geometric series is $\frac{1-\beta^n}{1-\beta}$. Because well-being is constant in both populations, this requires solving for n such that $0.1 \times \frac{1-\beta^n}{1-\beta} > 100 \times \frac{1-\beta^{100}}{1-\beta}$.

[17]Here we follow the convention of normalizing the g measure around the 0 welfare level, that is, we assume that $g(0) = 0$.

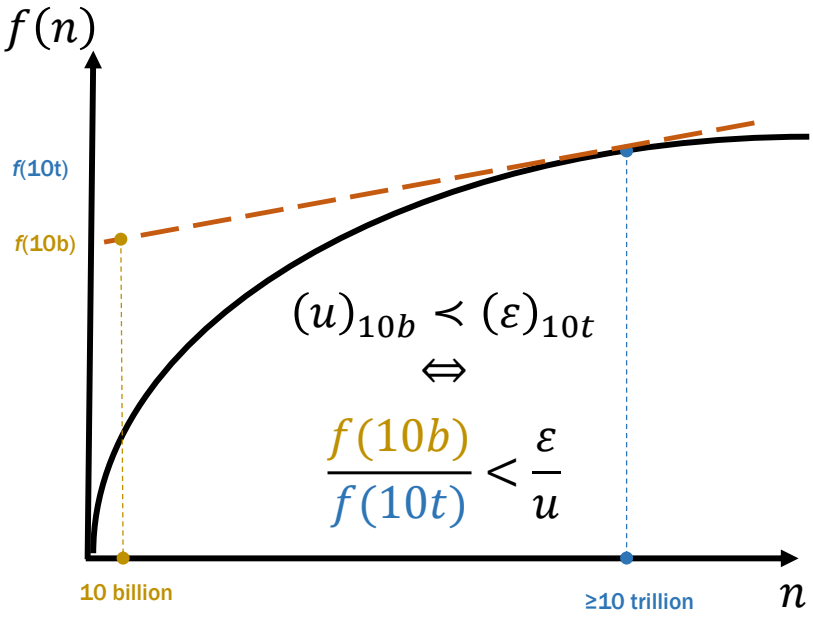
[18]Note that such a view violates the strong Mere Addition principle entailed by Total Utilitarianism. Hence, the view examined in Sider (1991) does not have this form. But, as mentioned in fn. 4, there are strong normative reasons for excluding from consideration the view in Sider (1991).

Figure 1: A graphical representation of concave variable-value calibration

Panel a. Addition of a good life bounds the slope of f at 10 trillion



Panel b. This slope bounds the ratio of f at 10 trillion and 10 billion



For an explanation of the choice of population sizes, see fn. 19.

addition assumption: the lowest lifetime well-being level that you are confident that, if added to our intertemporal population, would not decrease the goodness of the population. So far, we have been using the typical lifetime well-being of a developed-country professor, or more specifically the 97.5th percentile of the 2020 global socioeconomic distribution. But here, to get the strongest possible argument, we want the *lowest* lifetime well-being level for which mere addition reasoning can be safely applied. We expect that, for most readers, this is a level below that of most lives in 2020 developed countries. We denote this a well-being level of 1, where 0 is a neutral life (neither worth living nor worth not living).

Next fix a lifetime well-being level of what we expect the business-as-usual average lifetime well-being level to be for the overall intertemporal human population. By "business-as-usual" we mean the situation that will occur if the one life at well-being level 1 is *not* added. We label this average lifetime well-being level γ for "good" and assume that the long-run future of humanity is such that γ is greater than 1.

That it is not worse to add a life at 1, given the functional form of W, implies that:[19]

$$\frac{\gamma \times 10 \text{ trillion} + 1}{10 \text{ trillion} + 1} f(10 \text{ trillion } + 1) > \gamma f(10 \text{ trillion}).$$

This inequality bounds the slope of f at 10 trillion. By the concavity of f, the slope at 10 billion can be no less positive than the slope at 10 trillion. Extending this linearly towards zero provides an upper bound on f at 10 billion — but if concavity is steep the actual value may be well below this bound, resulting in even more repugnant-like implications.

Now we are in a situation to draw quantitative repugnant conclusions. The precise numbers depend on γ.[20] We expect that readers will take γ to be large. Recall that γ is the ratio of long-term business-as-usual average lifetime well-being to the lowest lifetime

[19] 10 trillion is an Ord (2020)-type estimate of the plausible size of the intertemporal human population. 10 billion is a commonly-used size of the high-welfare population (A) in the Repugnant Conclusion literature.

[20] This dependence is because of the arithmetic of averages.

well-being level such that we are confident that mere addition at that well-being level does not make the population worse. But our results are striking even if γ is not large.

- If $\gamma = 100$:
  - 10 trillion lives each at any positive well-being level worse than the threshold added life at 1 (or any other fixed, positive well-being level x) is better than
  - 10 billion lives, each 100 times as good as the threshold added life (or the other fixed, positive well-being level x).
- If $\gamma = 10$:
  - 10 trillion lives each at a well-being level $y > 0$ is better than
  - 10 billion lives, each 10 times as good as y.
- If $\gamma = 2$:
  - 10 trillion lives at any positive well-being level z is better than
  - 10 billion lives, each twice as good as z.

Clearly the implications for large γ constitute the very same repugnance that motivates some population ethicists to avoid Total Utilitarianism. To be sure, our argument uses an empirical premise about the size and well-being of the future population, formalized in the assumption that $\gamma > 1$. We believe that it is a strength of our argument that it speaks to a plausible calibration of the actual world. After all, that means that the implications we derive are not merely theoretical possibilities, but rather results that we would get if we were to apply the theories under examination in actual policy evaluation. But any reader uncomfortable with these empirical premises can read our argument as a *conditional* one, where the results are conditional on a plausible and relevant hypothetical future.[21]

[21] See, e.g., Ord (2020), Greaves and MacAskill (2019).

5 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.[22]

Why has the fact that Variable-Value views imply many instances of the Repugnant Conclusion been overlooked? We suggest that the reason is that standard formalizations of the Repugnant Conclusion use *universal* quantification ("For *any* perfectly equal population of very well-off people..."). But that quantification is not, we think, necessary to capture the intuition that there is something repugnant about views that suggest we choose a population consisting of lives that are barely worth living over a (smaller) population of excellent lives. That is, the fact that a view recommends we give up *many* populations of excellent lives for larger populations of lives that are barely worth living will strike those who worry about the Repugnant Conclusion as repugnant, even if the view in question does not make this suggestion for *all* populations. And as we have seen, Variable-Value population axiologies avoid *some* instances of such repugnant choices, but not other instances — just like any other population axiology (Spears and Budolfson, 2021).

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 — Variable-Value views substantially agree with Totalist views on how to rank policies that affect a relatively

[22]In fact, given Tännsjö's principle of unrestricted instantiation (recall fn. 6), these implications *must* be deemed repugnant if the Repugnant Conclusion is to be used as an argument against Total Utilitarianism.

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). The only escape would be for these Variable-Value views to be strikingly anti-natalist, such that they do not even satisfy 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. Variable-Value axiologies are commonly taken to avoid the Repugnant Conclusion. However, these views cannot avoid supporting repugnant-type judgments; not only in theory, but also, as we have seen, when these views are calibrated to the real world. So, one lesson from our paper is that Variable-Value views have been excluded from the set of population axiologies understood to imply repugnance only because of how the Repugnant Conclusion is typically formalized and quantified — not because they would rank populations in a way that would seem satisfactory to those who find the Repugnant Conclusion repugnant.

Population ethicists have long understood that escaping undesirable or unintuitive implications is impossible. But 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.[23]

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.

Atkinson, Anthony B. 1970. "On the measurement of inequality." *Journal of Economic Theory*, 2(3): 244–263.

Blackorby, Charles and David Donaldson. 1984. "Social criteria for evaluating population change." *Journal of Public Economics*, 25(1-2): 13–33.

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

[23]The authors of Zuber et al. (n.d)—a recent statement of agreement by many authors from diverse perspectives—argue that avoiding the Repugnant Conclusion has been overemphasized by population ethics research.

Bossert, Walter. 2017. "Anonymous welfarism, critical-level principles, and the repugnant and sadistic conclusions." working paper, University of Montréal.

Franz, Nathan and Dean Spears. 2020. "Mere Addition is equivalent to avoiding the Sadistic Conclusion in all plausible variable-population social orderings." *Economics Letters*, 196: 109547.

Greaves, Hilary and William MacAskill. 2019. "The case for strong longtermism." working paper, Global Priorities Institute.

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

Nebel, Jacob M. and H. Orri Stefánsson. 2020. "Calibration Dilemmas in the Ethics of Distribution." working paper, University of Southern California and Stockholm University.

Nebel, Jacob M. forthcoming. "Totalism Without Repugnance." in Tim Campbell, Jeff McMahan, and Ketan Ramakrishnan eds. *Festschrift for Derek Parfit*.

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. n.d. "What should we agree on about the repugnant conclusion?" working paper, UT Austin.