Philosophical Studies
https://doi.org/10.1007/s11098-024-02191-x

Incommensurability, the sequence argument, and the Pareto principle

Gustaf Arrhenius[1,2] · H. Orri Stefánsson[1,2,3]

Accepted: 5 July 2024

Abstract
Parfit (Theoria 82:110–127, 2016) responded to the Sequence Argument for the Repugnant Conclusion by introducing imprecise equality. However, Parfit's notion of imprecise equality lacked structure. Hájek and Rabinowicz (2022) improved on Parfit's proposal in this regard, by introducing a notion of *degrees of incommensurability*. Although Hájek and Rabinowicz's proposal is a step forward, and may help solve many paradoxes, it can only avoid the Repugnant Conclusion at great cost. First, there is a sequential argument for the Repugnant Conclusion that uses weaker and intuitively more compelling assumptions than the Sequence Argument, and which Hájek and Rabinowicz's proposal only undermines, in a principled way, by allowing for implausible weight to be put on the disvalue of inequality. Second, if Hájek and Rabinowicz do put such implausible weight on the disvalue of inequality, then they will have to accept that a population A is not worse than another same sized population B even though *everyone* in B is better off than anyone in A.

Keywords Incommensurability · Population ethics · Pareto principle · Levelling down objection · Repugnant conclusion

1 Introduction

Here's a simple and general formulation of Derek Parfit's infamous "Repugnant conclusion":

✉ Gustaf Arrhenius
gustaf.arrhenius@iffs.se

✉ H. Orri Stefánsson
orri.stefansson@philosophy.su.se

[1] Institute for Futures Studies, Stockholm, Sweden

[2] Stockholm University, Stockholm, Sweden

[3] Swedish Collegium for Advanced Study, Uppsala, Sweden

Published online: 19 September 2024

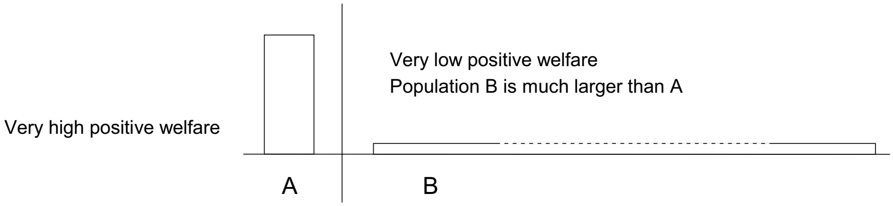
Very high positive welfare
Very low positive welfare
Population B is much larger than A
A
B

Fig. 1 The Repugnant Conclusion

> *The Repugnant Conclusion*: For any population consisting of people with very high positive welfare, there is a better population in which everyone has a very low positive welfare, other things being equal.[1]

In Fig. 1, the width of each block represents the number of people whereas the height represents their lifetime welfare. Dashes indicate that the block in question should be much wider than shown, that is, the population size is much larger than shown.

These populations could consist of all the past, present and future lives (a possible world), or all the present and future lives, or all the lives during some shorter time span in the future such as the next generation, or all the lives that are causally affected by, or consequences of a certain action or series of actions, and so forth.[2]

All the lives in the figure have positive welfare, or, as we also could put it, all the people have lives worth living. Both populations are perfectly equal populations, that is, all the individuals in the respective populations have the same welfare. The A-people have very high welfare whereas the B-people have very low positive welfare.[3] The reason for this could be that in the B-lives there are, to paraphrase Parfit, only enough ecstasies to just outweigh the agonies, or that the good things in those lives are of uniformly poor quality, e.g., eating potatoes and listening to Muzak.[4] However, since there are many more people in B, the total sum of welfare in B is greater than in A. Hence, a theory like Total Utilitarianism, according to which we

[1] For Parfit's original formulation, see Parfit (1984), p. 388. Our formulation is more general than his. For early sources of the Repugnant Conclusion, see Arrhenius (2000), (2016), (forthcoming).

[2] More exactly, a population is a finite set of lives in a possible world. A, B, C, …, A_1, A_2, …, A_n, A∪B, and so on, denote populations of finite size. We shall adopt the convention that populations represented by different symbols (including "+", "*", and the like), or the same letter but different indexes, are pairwise disjoint. For example, $A \cap B = A_1 \cap A_2 = A' \cap B' = \phi$. We shall assume that for any natural number *n* and any welfare level X, there is a possible population of *n* people with welfare X (for a discussion of this *No-Limit Assumption*, see Arrhenius (2000) ch. 3, (forthcoming)).

[3] For a discussion and definition of positive, negative, and neutral welfare, see Arrhenius (2000), (forthcoming) ch. 2 and 9 (for a short summary, see Arrhenius (2016)). Cf. Broome (1999), (2004), Bykvist (2007), p. 101, and Parfit (1984), pp. 357–358 and appendix G. Notice also that we don't need an analysis of a neutral welfare in the present context but rather just a criterion, and the criterion can vary with different theories of welfare.

[4] See Parfit (1984), p. 388 and Parfit (1986), p. 148. For a discussion of different interpretations of the Repugnant Conclusion see Arrhenius (2000), (forthcoming) and Parfit (1984), (2014), (2016).

should maximize the welfare in the world, ranks B as better than A—an instance of the Repugnant Conclusion.[5]

Notice that the Repugnant Conclusion is not just a problem for total utilitarians or those committed to welfarism—the view that welfare is the only value that matters from the moral point of view—since the *ceteris paribus* clause in the formulation implies that the compared populations are equal in all possibly axiologically relevant respects apart from individual welfare levels. Hence, other values and considerations are not decisive for the value comparison of populations A and B. Thus, the Repugnant Conclusion is a problem for all moral theories according to which welfare matters at least when all other things are equal, which arguably is a minimal adequacy condition for any moral theory.[6]

As the name indicates, Parfit found the Repugnant Conclusion very counterintuitive and most philosophers seem to agree. However, there is a well-known and tempting argument for the Repugnant Conclusion, which Parfit called the "Continuum" Argument. That is an unfortunate misnomer, since the argument does not in fact require a continuum. Therefore, we shall instead refer to it as the "Sequence Argument". In section II we explain the Sequence Argument in more detail, but in short, the argument starts with a population like A, where everyone has very high positive welfare, and then introduces a sequence of populations, where each population is much bigger but offers slightly lower individual welfare than the previous population in the sequence. One might hold that for any two consecutive populations in this sort of sequence, the latter, if sufficiently large, is better than the former much smaller one, since the reduction in individual welfare is so small. But then, since "better than" is a transitive relation, we sooner or later get the Repugnant Conclusion, that is, we find that a population like B, in Fig. 1, must be better than population A in Fig. 1.

Parfit (2016) responded to the Sequence Argument by suggesting that adjacent populations are actually "imprecisely equally good". In section II we briefly explain Parfit's response, but the important observation about imprecise equality is that it is not transitive. Therefore, it is possible that each population in the Sequence Argument is imprecisely equally good as the population that comes before it, even though the last population is *worse* than the first population.

However, Parfit's notion of imprecise equality lacked structure. Hájek and Rabinowicz (2022) improve on Parfit's proposal in this regard. In section III we discuss their argument in detail, but in short, their contribution consists in introducing and formalising a notion of *degrees* of incommensurability. An important benefit of their proposal is that they offer a plausible explanation why people would *erroneously* (in Hájek and Rabinowicz's view) judge that each population in the Sequence Argument is better than a previous population, even if they are in fact incommensurable.

[5] Throughout this paper "better" means "better, all things considered" if not otherwise indicated.

[6] Note that this holds for *deontic* views too. Plausible deontic views hold that, when all other moral considerations are equal, individual welfare levels are relevant when considering what population to bring about. For a discussion of deontic population ethics, see Arrhenius (2022), (forthcoming).

Although Hájek and Rabinowicz's proposal is a step forward, and may help solve many paradoxes, it can only avoid the Repugnant Conclusion at great theoretical cost. First, as we explain in section IV, there is a sequential argument for the Repugnant Conclusion—the Sequential Dominance Addition Argument—that uses weaker and intuitively more compelling assumptions than the original Sequence Argument. We assume that Hájek and Rabinowicz want to satisfy what we call the *Non-Redundancy constraint*: any response on their behalf, to an argument for the Repugnant Conclusion, should not make their introduction of (degrees of) incommensurability redundant, in the sense that the response itself blocks the argument for the Repugnant Conclusion without appealing to incommensurability. And we show that Hájek and Rabinowicz's proposal can only undermine the Sequential Dominance Addition Argument, while satisfying the Non-Redundancy constraint, by allowing for seemingly implausible weight to be put on the disvalue of inequality.

Second, if Hájek and Rabinowicz do put such seemingly implausible weight on the disvalue of inequality, then they will have to accept that a population A is not worse than another equally large population B even though everyone in B is better off than anyone in A. So, their proposal then violates the traditional Pareto principle, which concerns same sized-populations, and thus faces the 'levelling down objection' (Parfit (1997)).

In a sense, what we are pointing out is not surprising: one cannot avoid the Repugnant Conclusion without having to accept some counterintuitive implication or make some intuitively implausible assumption. That has been known for decades; hence, the Repugnant Conclusion is often seen as a *paradox* of population ethics. However, what we take to be interesting about the above result is that in order to avoid the Repugnant Conclusion, Hájek and Rabinowicz have to violate a *same-size* population condition that most would want to accept, namely, the Pareto principle. Giving up the Pareto principle is a pretty hefty price to pay to avoid the Repugnant Conclusion, and Hájek and Rabinowicz have not, as far as we can tell, given us an independent justification for giving up that principle, rather than, say, giving up avoidance of the Repugnant Conclusion.

2 The sequence argument for the repugnant conclusion and Parfit's response

Consider first the following condition:

> *Quantity*: For any pair of positive welfare levels, A and B, such that B is slightly lower than A, and for any number of lives *n*, there is a greater number of lives *m*, such that a population of *m* lives at level B is better than a population of *n* lives at level A, other things being equal.[7]

[7] A welfare level is an equivalence class on the set of all possible lives with respect to the relation "has at least as high welfare as". For an exact statement of this principle, see Arrhenius (2000), (forthcoming) where this condition is formulated in terms of "at least as good as".

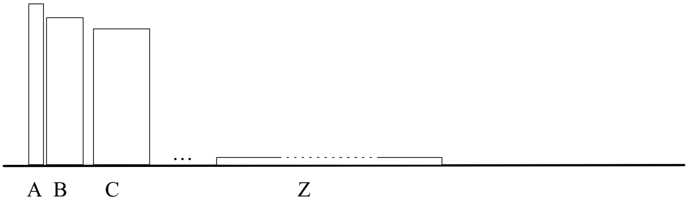

Fig. 2 The sequence argument

Quantity has some intuitive plausibility and should appeal to those who find some truth in the saying "the more good, the better". However, it implies the Repugnant Conclusion together with a reasonable assumption about the structure of welfare:[8]

> *Finite Fine-grainedness*: There exists a finite sequence of slight welfare differences between any two welfare levels.

The idea here is that one can get from one welfare level to another in a finite number of steps of intuitively slight welfare differences. Examples of such welfare differences could be some minor pain or pleasure or a shortening of life by a minute or two.[9] These differences don't have to be of the same size or type. Let's say that a life of type *a* has higher welfare than a life of type *b*, and suppose that you are successively making *a* slightly worse, perhaps by shortening it by a minute or two or by adding some minor pain. Finite Fine-grainedness implies that there is a finite (but possibly great) number of such slight worsening from *a* to another type of life *c* such that a life of this type will have the same welfare or lower welfare than a life of type *b*. It is quite hard to deny the intuitive force of this assumption.[10]

Consider the following sequence of populations for an informal demonstration that these two conditions together imply the Repugnant Conclusion[11]:

Assume that A in the Fig. 2 above is a population with very high welfare and that Z is a population with very low positive welfare (again, the width of the blocks represents the number of lives in the population, the height represents their lifetime welfare; dashes indicates that the block in question is much wider than shown). According to Quantity, there is a population B with slightly lower welfare than A

[8] It also implies, and thus presupposes, *the No-Limit Assumption*: For any possible population consisting of lives with a certain welfare, there is a larger possible population consisting of lives with the same welfare. For a discussion, see Arrhenius (2000), (forthcoming).

[9] For a precise definition of "slight welfare difference" see Arrhenius (2000), (forthcoming).

[10] Notice that Finite Fine-grainedness doesn't imply that all sequences of slight welfare differences between two welfare levels are finite, just that there exist at least one such sequence. It is compatible with the welfare ordering being continuous as well as discreet. It just rules out that there are, so to speak, big "jumps" or "holes" in the order of welfare levels. For a discussion of Finite Fine-grainedness and possible theories of welfare that violate this condition, see Arrhenius (2005), (forthcoming); Arrhenius & Rabinowicz (2015). For an effort to challenge Finite Fine-grainedness (in light of the impossibility theorems in population ethics), see Thomas (2018) and Carlson (2022). These efforts fail for several reasons, one is that they missed that lotteries can make lives better or worse and the difference in value of lotteries can be made arbitrarily small by making small changes to the probabilities.

[11] For a proof, see Arrhenius (2000), (forthcoming).

and which is better than A; a population C with slightly lower welfare than B and which is better than B; and so forth. We can assume that the welfare levels in this sequence of populations satisfy Finite Fine-grainedness. Hence, we will finally reach population Z with very low positive welfare. By transitivity, Z is better than A. Since A is an arbitrary population with very high welfare, this shows that for any population with very high welfare, there is a population with very low positive welfare which is better, that is, the Repugnant Conclusion. Consequently, assuming Finite Fine-grainedness, any theory which avoids the Repugnant Conclusion has to violate Quantity.

As previously mentioned, Parfit (2016) suggests a way of avoiding the sequence derivation of the Repugnant Conclusion by introducing what he calls "imprecision" in value comparisons.[12] He suggests that in a range of important cases, outcomes are only imprecisely comparable.[13] In such cases, transitive relations such as "equally as good as" are not applicable. Instead, we have to make use of imprecise concepts that are non-transitive.

In the Sequence Argument, Parfit suggested that each population is "imprecisely equally good" to adjacent populations in the sequence. However, since imprecisely equally good is not a transitive relation, he could still maintain that the last population in the sequence is worse than the first population in the sequence. In other words, he had an answer to the Sequence Argument for the Repugnant Conclusion.

Our aim in this article is not to assess how plausible Parfit's answer was (for an assessment of that, see Arrhenius (2021b)). Instead, we shall assess Hájek and Rabinowicz's additions to Parfit's proposal, to which we now turn.

3 Hájek and Rabinowicz's addition to Parfit's proposal

Hájek and Rabinowicz's basic observation is that cases that involve incommensurability can differ in *how far from* comparable the relevant options are:

> Sometimes, when attempting to compare two alternatives, we are totally flummoxed, regarding them as not really comparable at all. In other cases, we are more inclined to form a preference one way or another, or to regard them with indifference, but we do so with some hesitancy. And in many of these cases, the hesitancy comes in degrees because incommensurability comes in degrees. (2022: 899)

So, contrary to what Parfit's remarks may have suggested, (in)comparability is not a binary—an either/or—property. Sometimes two options are really incomparable,

[12] Parfit (2014), (2016). Here we are just summarizing his argument, drawing on Arrhenius (2021b) where a detailed discussion can be found, to contrast it with Hájek and Rabinowicz theory.

[13] This imprecision is not due to any cognitive or epistemic limitations, Parfit thought, but a fact about the value comparisons of certain types of outcomes.

and sometimes they are really comparable.[14] But sometimes they are somewhere in between these two extremes, say, close to being comparable. Hájek and Rabinowicz's illustrate their idea with the following example:

> Who was more of a genius: Einstein or Bach? Plausibly, they are incommensurable—one was a great scientist, the other a great composer. How about Einstein or Chopin? Plausibly, they are still incommensurable, but perhaps it is easier to favor Einstein: while Chopin was undoubtedly a genius of piano composition, he arguably did not quite have Bach's range. How about Einstein or Schumann? This comparison is arguably easier again—while brilliant, Schumann was not quite as original as Chopin, let alone Bach. How about Einstein or Salieri, the mediocre composer made famous by Amadeus? That's easy—Einstein was the greater genius, period. We have proceeded by steps to closer and closer approximations to the 'better' relation with regard to genius. (ibid)

Hájek and Rabinowicz's focus is on value comparisons, analysed in terms of *fitting attitudes* (Brentano (1969)). On this view, alternative A is better than B if it is *fitting to prefer* A to B and no other attitude is fitting. In that case one ought to prefer A to B. A and B are equally good if it is fitting to be indifferent between them and no other attitude is fitting. But sometimes, Hájek and Rabinowicz suggest, there may be more than one fitting attitude one could have when comparing A and B. There would, then, be more than one permissible preference ordering of A versus B. It could be fitting to prefer A to B, making it permissible to rank A over B, while it at the same time being fitting to prefer B to A, making it permissible to rank B over A (and indeed it might be permissible to also be indifferent between A and B). In that case, A and B are *incomparable*, since they contain (or realise) *incommensurable* values.

Given the above understanding of incommensurability, there is a natural way of conceptualising *degrees* of incommensurability:

> We now add that the degree of commensurability can be higher or lower depending on the extent to which different permissible orderings agree or disagree in their ranking of the items. If in nearly all permissible orderings A and B are ranked in the same way, their degree of commensurability is very high—for example, if A is almost always ranked above B, or they are almost always equal-ranked. But if there is more divergence in how A and B are ranked, their degree of commensurability is lower. (Equivalently, their degree of incommensurability is higher.) (2022: 900)

Hájek and Rabinowicz add that if *almost* all permissible preference rankings of A versus B have A higher than B, then A is *almost better* than B. In that case, A and B are commensurable to a high degree, but still incommensurable as long as *some* permissible preference ranking has B higher than A.

[14] We take it that Hájek and Rabinowicz are here not referring to our abilities to compare, even though their choice of terminology may suggests otherwise, but rather whether the options are in fact comparable. (Thanks to Nir Eyal for making us see the need to clarify this.).

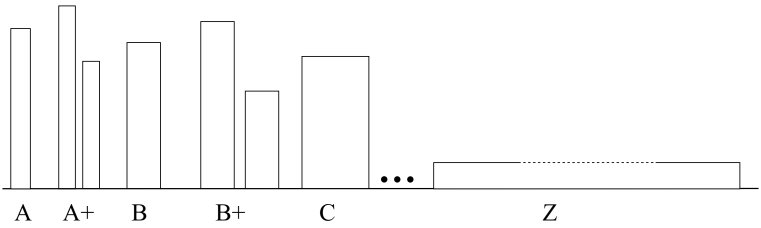

Fig. 3 The sequential dominance addition argument

Hájek and Rabinowicz suggest ways of making these degrees precise; most simply, in the finite case, one can equate degrees with proportions. The exact details of Hájek and Rabinowicz's proposal are however not all relevant for our purposes. What is relevant is how they apply their general idea to counter the Sequence Argument for the Repugnant Conclusion, while at the same time adding important details to Parfit's similar argument. As Parfit, Hájek and Rabinowicz suggest that it is false that each population in the Sequence Argument is better than its immediate predecessor. Instead, they are incommensurable. And unlike the better-than relation, the incommensurable-to relation is not transitive. Thus, the Sequence Argument for the Repugnant Conclusion is undermined.

In addition, Hájek and Rabinowicz suggest that each population is *almost better* than its immediate predecessor. That would explain why so many people get 'tricked' by the Sequence Argument into endorsing the Repugnant Conclusion, and why very few people say that some (or all) populations in the Sequence Argument are *worse* than their immediate predecessor. So, Hájek and Rabinowicz may have found a plausible *error theory* of people's judgement.

> Each [population] is not better than its predecessor, but it is almost better. In fact, it is so close to being better that we mistake the one relation for the other. We do not notice or we ignore the reasonable weighings that do not favor the second population over the first, because they are overwhelmed by those that do. But it is a minor mistake: *almost better* is almost *better*! Our intuitions are wrong, but almost right. This is the error theory that Parfit needed. (2022: 904)

An important question that the above remarks raise is how one should *choose* when one option is *almost better* than another. It does not seem implausible that if, say, A is better than B according to all permissible preference rankings except one, then we ought to choose A over B. But that would mean that Hájek and Rabinowicz cannot avoid a *deontic* version of the Sequence Argument for the Repugnant Conclusion, that is, an argument that is formulated in terms of 'more choiceworthy than' rather than in terms of 'better than'.

Nevertheless, we grant that Hájek and Rabinowicz have suggested an important improvement on Parfit's response to the Sequence Argument. Moreover, the notion of degrees of incommensurability is fruitful outside of population ethics, for instance, promising to solve—or shed light on—paradoxes and puzzles in other areas of philosophy. Unfortunately, however, Hájek and Rabinowicz's proposal can only avoid the Repugnant Conclusion at considerable cost. To appreciate these costs, it is helpful to consider a different (and, in our view, more convincing) sequential argument for the Repugnant Conclusion.

4 The cost of Hájek and Rabinowicz's attempt to avoid the repugnant conclusion

Now instead of the sequence in the original Sequence Argument, consider the following:

All the lives in population A in Fig. 3 enjoy very high welfare. In A+, we have one population that is equally as large as the A-population but with lives that enjoy even higher welfare.[15] In addition, A+contains a second population with positive welfare but a bit lower than in A. However, we assume that the welfare of the better-off lives in A+is sufficiently high to make the average welfare in A+greater than that in A. It seems to us hard to deny that A+is better than A, and not only almost better. In B, which is of the same size as A+, we have equalized the welfare at a level higher than the+-lives but lower than the A-lives, in a way that increases aggregate (and thus also average) welfare. Unless one has anti-egalitarian intuitions, it seems hard to deny that B is better than A+. And similarly for other consecutive populations in this sequence. But then we are again faced with the Repugnant Conclusion: Z is better than A.

In a moment we will explain the cost of introducing incommensurability to undermine the above "Dominance Addition" argument for the Repugnant Conclusion. But first, let's make the argument more precise, by introducing the two conditions that we implicitly appealed to above when deriving the Repugnant Conclusion. Here's the first one:

> *Dominance Addition*: An addition of lives with positive welfare and an increase in the welfare in all the lives in the rest of the population makes the population better, other things being equal.[16]

One way to motivate Dominance Addition is that you don't make a population worse by adding lives worth living, and you make a population better by increasing the welfare of the lives in it, so all together you get a better population if you do both.

One could make Dominance Addition even more compelling by assuming that the non-added people are the same in the two compared populations. Then one could also appeal to so-called *person-affecting view* for judging A+better than A since then the A-people will benefit in the move from A to A+. We shall not avail ourselves of this possibility here, however, since the person-affecting view has been

[15] Notice, as we stated in fn. 2, that populations represented by different symbols (including "+", "*", and the like) are pairwise disjoint. So, for example, A and A+are disjoint.

[16] For an exact statement of this condition, see Arrhenius (2000), (forthcoming) where it is formulated in a logically weaker manner in terms of "at least as good as". We are using the stronger formulation here to simplify the exposition.

shown to be deeply problematic (e.g., Arrhenius (2000), (2009a), (forthcoming)). We shall instead continue to assume that the compared populations are pairwise disjoint. Those who still think the person-affecting view can be salvaged may however make that assumption which some will find strengthens the intuitive appeal of Dominance Addition.

Dominance addition is an intuitively more compelling version of the more well-known *Mere Addition Principle*: An addition of people with positive welfare does not make a population worse, other things being equal.[17] Yet, although this principle might seem compelling at first glance, it is controversial and several authors have rejected it.[18] One might, for example, object to it on egalitarian grounds since a mere addition can introduce great inequality in an otherwise perfectly equal population.[19] Likewise of course for Dominance Addition although then the disvalue of the introduced inequality has to be weighed against the positive value of the increased welfare of the lives in the original population, not only against the possible positive value of more lives with positive welfare. We shall get back to such objections to Dominance Addition in a moment. But first, we introduce the second condition we appealed to informally above when deriving the Repugnant Conclusion:

> *Inequality Aversion*: For any triplet of welfare levels, A, B, and C, A higher than B and B higher than C, and for any population A with welfare A, there is some larger population C with welfare C such that a perfectly equal population B of the same size as A∪C and with welfare B is better than A∪C, other things being equal.[20]

Another way of stating Inequality Aversion is that for any welfare level of the best off and worst off, and for any number of best off lives, there is some (possibly much) greater number of worst off lives such that it would be better to have an equal distribution of welfare on any level higher than the worst off, other things being equal.

The above is a very weak egalitarian condition since it can be satisfied by a theory which demands that the total welfare must be greater for a population with perfect equality to be better than an unequal population of the same size. Moreover, it is also compatible with principles that give much greater weight to the welfare of the best off as compared to the welfare of the worst off. For example, a theory which requires that to compensate for one life falling from twenty to ten units of welfare, a

[17] Cf. Parfit (2014), p. 420ff, Hudson (1987), Ng (1989), and Sider (1991). Cf. fn. below. Notice that the original formulation of Dominance Addition (see fn. above) is also logically weaker than the Mere Addition Principle.

[18] See e.g. Ng (1989), p. 244; Blackorby et al. (1995), p. 1305, and Blackorby et al. (1997), pp. 210–211; Fehige (1998). From Parfit (2014), p. 420ff, one might get the impression that he thinks a population axiology should satisfy the Mere Addition Principle (see Ng (1989), p. 238) but in personal communication, Parfit has expressed doubts about the Mere Addition Principle in cases where the added people are much worse off than the rest of the population. See also Feldman (1997) ch. 10, Kavka (1982), and Carlson (1998), pp. 288–289.

[19] See Arrhenius (2009b), (2013), (forthcoming).

[20] For an exact statement of this principle, see Arrhenius (2000), (forthcoming) where this condition is formulated in terms of "at least as good as". We've here formulated it in terms of "better than" to simplify the exposition.

hundred lives have to be moved from zero to ten units, is compatible with Inequality Aversion. In that sense, its name is a bit misleading since it is compatible with quite non-egalitarian theories. Roughly, Inequality Aversion only rules out theories that imply that we should always or sometimes give some kind of “lexical priority” to the best off. A simple example of such a theory is “Maximax”: Maximise the welfare of the best off (some more subtle examples will be discussed below when we address some objections to Inequality Aversion).

Let’s consider now what Dominance Addition and Inequality Aversion imply for the sequence in Fig. 3. Dominance Addition implies that A+ is better than A. We can assume that A+ and B fulfil the antecedent of Inequality Aversion.[21] So, Inequality Aversion implies that B is better than A+. Likewise for populations B, B+, and C, and so forth until we finally reach population Z with very low positive welfare. By transitivity, Z is better than A, that is, the Repugnant Conclusion.

Since Hájek and Rabinowicz’s article concerned the Sequence Argument, in which same number conditions such as Inequality Aversion play no role, we have no textual evidence that they want to use their proposal to question such same number conditions. More importantly: we are sceptical that Hájek and Rabinowicz’s proposal has resources to plausibly deny Inequality aversion.

Admittedly, there are some theories that violate Inequality Aversion and that could be made more plausible with the help of Hájek and Rabinowicz’s proposal, in particular, theories that invoke some form of superiority in value. (See Arrhenius (2005); Arrhenius and Rabinowicz (2005), (2015) for a discussion.) Some such views would imply that there is some ‘higher’ value that, first, is necessarily lost when a life drops below some wellbeing level (which in itself is a controversial assumption), and that, second, cannot be made up for by any quantity of some ‘lower’ value. Hájek and Rabinowicz’s proposal might make such views more plausible with regards to the second (but not the first) assumption, by, first, implying that great quantities of the lower quantity could make things *almost* better despite a loss in the higher quantity, and, second, by allowing for incommensurability in where to draw the line below which the higher value is lost.[22]

However, as discussed in Arrhenius (2000), (2016), (forthcoming), Inequality Aversion can be derived from an even more intuitively compelling condition, *Non-Elitism*. Informally put, Non-Elitism ensures that there is some (possibly great) number of worst off people such that a slight decrease in welfare for *one* of the best off persons can be compensated for by an at least as great increase in welfare for all those worst off people. And of course, theories that violate Inequality Aversion, by invoking superiority in value, will also violate Non-Elitism.

It seems to us that even if Hájek and Rabinowicz’s proposal could make the denial of Inequality Aversion more plausible, it is very hard to see how it could make the rejection of Non-Elitism less implausible. For the proposal would then have to be that there is *no* number of worst off people such that a slight decrease in

[21] If welfare is measurable on at least an interval scale, we could also assume that the total and average welfare in B is higher than in A+.

[22] We owe this point to a reviewer for this journal.

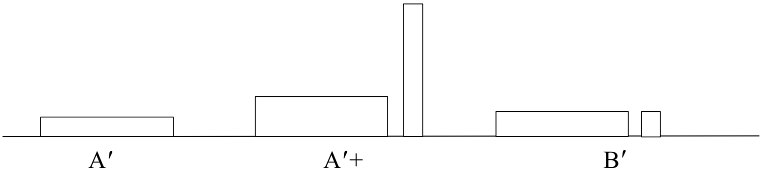

Fig. 4 Levelling down

welfare for one of the best off persons can be compensated for by an at least as great increase in welfare for all those worst off people: making all the worst off people that much better off would only *almost* compensate for making one of the best off people worse off. That seems too implausible to us. Still, to simplify the discussion, we shall continue to use Inequality Aversion rather than the more complex Non-Elitism condition.

So, let's suppose then that Hájek and Rabinowicz want to resist the Sequential Dominance Addition Argument by rejecting Dominance Addition. They do in fact have the resources to do so. For they could claim that there is a permissible preference ordering that ranks A above A+, for instance, a preference ordering that puts very high weight on the disvalue of inequality. (In a moment we shall consider another reason for why there could be a permissible preference that ranks A above A+.) However, the preference in question really would have to put *a lot* of weight on the disvalue of inequality. After all, we can make the population that gets better lives when we move from A to A+ arbitrarily large, and we can similarly make the additional people in A+ (whose lives are worth living) arbitrarily numerous.[23] So, to avoid saying that A+ is better than A by appealing to the permissibility of valuing equality, Hájek and Rabinowicz have to say that it is permissible to give what seems to us to be implausibly high importance to equality. And while their framework makes room for such judgements, nothing in their paper gives us *good reasons* for such judgements. Let's however set that issue aside.

Now, it is worth noting that denying Dominance Addition while accepting Inequality Aversion may not quite suffice to deliver the result that Hájek and Rabinowicz are after.[24] For instance, one could construct a sequence where the first step would replace A, from Fig. 3, with a population A* consisting in one subpopulation equally numerous to A but much better off and another huge subpopulation of lives that are barely worth living. After that one could apply Inequality Aversion repeatedly, until one gets to population Z from Fig. 3. By Inequality Aversion and Transitivity, Z would be better than A*. But then if A* is *almost* better than A, then

[23] We are assuming that Hájek and Rabinowicz do not deny that the number of people enjoying very high levels of welfare is of *some* moral importance. After all, if they denied that, say, total welfare is of *any* moral importance, then that would suffice to block the Sequence Argument (without appealing to incommensurability). So, denying that total welfare is of any moral importance would violate the Non-Redundancy constraint.

[24] We thank a referee for very fruitful correspondence about this issue.

the conclusion is that Z too is almost better than A. (For a similar argument, using Non-Elitism instead of Inequality Aversion, see Arrhenius (2016)). But that Hájek and Rabinowicz want to deny: they want to say that Z is commensurable with *and worse than* A.

So, *only* denying Dominance Addition might not quite deliver everything that Hájek and Rabinowicz are after. It would however strictly speaking allow them to avoid the sequential arguments for the (stronger, and traditional) Repugnant Conclusion, since it would no longer follow that Z is *better than* A. Moreover, the above argument (based on Inequality Aversion) that Z is almost better than A requires accepting the A* is almost better than A even though the former (but not the latter) contains a lot of lives that are only *barely* worth living. Let's therefore focus on an issue that arises if they do want to resist the Sequential Dominance Addition Argument by rejecting Dominance Addition.

Consider Fig. 4. We assume that the number of people in A′ is *n*, which is the same as the number of the worse-off people in A′+. The *n* worse-off people in A′+are better off than the people in A′. In addition, A′+contains some even better off people. Population B′ however contains exactly the same number of people as population A′+, but in B′ everyone is worse off than the worse-off people in A’+but still better off than the people in A′.

Now compare population A′ with population A′+. Here it would seem that Hájek and Rabinowicz would have to say that the latter is only *almost* better than the former; that is, there is some permissible preference according to which A′ ranks higher than A′+, namely, a preference that places a very high weight on the disvalue of inequality. At the very least, there will have to be *some* similar pair of populations for which they will have to say that the population containing both more people and higher welfare for everyone is only *almost* better, if they are to resist the Sequential Dominance Addition Argument for the Repugnant Conclusion by claiming that in Fig. 3 A+ is not better than A due to the added inequality in the former. Moreover, to get their desired conclusion that Z is *worse* than A (rather than only that Z is not better than A), they need this to hold for more than just one pair in the sequence.

What about A′ versus B′? It is hard to see how there could be a permissible preference that does not rank B′ over A′. The difference between the two is that, first, everyone in B′ is better off than anyone in A′, and, second, B′ contains more people with lives worth living. But there is no added inequality in B′ compared to A′; nor is there anything else in B′ but not in A′ that could, in our view, plausibly be of negative value. So, if *either* having more people with lives worth living makes a world at all better, no matter how slight, *or* if everyone being better off makes a world at all better, then we must say that B′ is better than A′. For the purposes of our argument, it however suffices that B′ is at least as good as A′ (as should be apparent below).

In response to the last paragraph, some might point out that there is a respectable view according to which B′ could contain something of negative value that A′ does not. For according to *Critical Level Utilitarianism* (CLU), adding lives with positive welfare under a positive critical level has negative value. So, if the people in both A′ and B′ are below the critical level, then the fact that there are *more* people in B′ might make the former better, according to CLU.

However, Hájek and Rabinowicz can hardly appeal to CLU in response to our argument. The reason is that if a critical level is allowed, then we already have a response to the Sequence Argument, since once we get below the critical level in the sequence, the populations get worse and worse, according to CLU, the further along the sequence we go. In other words, appealing to CLU would violate the Non-Redundancy constraint.[25]

So, we can safely assume that Hájek and Rabinowicz won't respond to our argument by assuming CLU. Is there some other way to deny that B' is at least as good as A' (in Fig. 4)? In particular, is there some way for Hájek and Rabinowicz do deny this without violating the Non-Redundancy constraint?[26]

Perhaps the most principled way to deny that B' is at least as good as A' (in Fig. 4) is to say that populations of different sizes are *always* incommensurable. In fact, Parfit briefly considered such a view.[27] That however seems to us very implausible (and, in fact, Parfit himself abandoned the view). For instance, it would imply that a population in Stone Age conditions, where nobody has an excellent life and most people lead very miserable lives, is no worse than a population in which a huge number of people live in great luxury thanks to technological and moral advancement.[28]

We can thus assume that B' is at least as good as A'. However, recall that to avoid the Repugnant Conclusion, Hájek and Rabinowicz have to say that $A'+$ is merely *almost* better than A'. Therefore, since better-than is a transitive relation, they have to deny that $A'+$ is better than B'. But that seems highly counterintuitive (even though they can say that $A'+$ is *almost* better than B). These populations contain the same number of people, but everyone in $A'+$ is better off than anyone in B'. In fact, some people in $A'+$ are *much* better off than anyone in B'. So, Hájek and Rabinowicz have to reject a weak version of the widely endorsed Pareto principle, according to which a population A^* is better than a same sized population B^* if everyone in A^* is better off anyone in B^*. For the same reason, they face the levelling down objection.

Is there some way for Hájek and Rabinowicz to resist the above implication while also resisting the Sequential Dominance Addition Argument for the Repugnant Conclusion and satisfying the Non-Redundancy constraint? We can think of one response on their behalf. Consider again Fig. 3. Hájek and Rabinowicz could argue that the reason $A+$ is only almost better than A is that there is a preference that ranks

[25] It is worth acknowledging that Hájek and Rabinowicz's proposal may rationalise a version of CLU, Incomplete CLU according to which there are an interval of critical levels between zero and a positive welfare levelα. See Blackorby et al. (1997), pp. 216–219; Blackorby, Bossert, & Donaldson (2005), ch. 7. For a critical discussion of Incomplete CLU, see Arrhenius (2000), (2021a), (forthcoming).

[26] A reviewer for this journal points out that someone who either puts great weight on average welfare or on the welfare of the worst off might reject Dominance Addition. But appealing to the importance of average welfare or the welfare of the worst off would, in the present context, violate the Non-Redundancy constraint.

[27] See in particular Parfit's Rolf Schock Prize Lecture and his unpublished 2014 manuscript based on the lecture. See also Arrhenius (2016) for a lengthier discussion of this view.

[28] Thanks to Nir Eyal for suggesting to us an example like this.

A over A+, but not in virtue of the inequality in the latter, but rather because in the latter it is not true that *everyone has a fantastic life*.[29] At least, that would plausibly be true for some pair of worlds with the relevant relationship, that is, where one is a "dominance addition" of the other. But if they claim that it is *not* permissible to base one's preference for A over A+ on concern for equality, then they don't have to say that it is permissible to prefer A′ over A′+; so, they don't have to violate the Pareto principle.

But is the above response plausible? We think not. To resist the Sequential Dominance Addition Argument for the Repugnant Conclusion, Hájek and Rabinowicz have to be very liberal about what can be permissibly preferred and what reasons one can permissibly have for one's preferences. In particular, they have to say that it is permissible to prefer A over A+ *because* only in the former world does everyone have a fantastic life. (Or at least, they have to say that of some worlds where the latter is a dominance addition of the former.) But in another sense, they cannot be liberal about what can be permissibly preferred: they have to say that it is impermissible to prefer A over A+ *because* the latter contains inequality.

The above response that we are considering on Hájek and Rabinowicz's behalf therefore strikes us as being rather odd. Equality is a widely recognised value and many people think it is fitting to accept considerable cost to bring it about. But the same doesn't seem true about everyone having fantastic lives. There is, for instance, no traditional distributive view that places a particular significance on *everyone* having fantastic lives. Egalitarians think that it is good that everyone is equally well-off; but if that justifies preferring A over A+, then that is because of the importance of *equality*, not because of the importance of everyone having fantastic lives. Utilitarians by contrast place greater weight on everyone having fantastic lives than on equality; but utilitarian principles do not justify preferring A over A+. More generally, it seems to us that it would be hard to find a principled and ethically sound justification for preferring A over A+ that is not grounded in the value of equality. But then it may not be possible to satisfy the Pareto principle and avoid the levelling down objection.

5 Concluding remarks

Before concluding, we would like to acknowledge again that, first, Hájek and Rabinowicz's proposal is interesting in its own right and may shed light on various paradoxes in philosophy; and, second, that we think their response to the Sequence Argument is an improvement on Parfit's. Nevertheless, their proposal can only help us avoid the Repugnant Conclusion at great cost. For as we have

[29] More generally, Hájek and Rabinowicz could argue that A is better than A+ due to some holistic value that A has but that A+ lacks. (Thanks to a reviewer for this journal for pointing this out.) Lack of inequality may be one such holistic value, everyone having fantastic lives might be another holistic value, and *the majority* having fantastic lives might be yet another such holistic value (cf. Carlson 1998). Perhaps there are other *plausible* holistic values that could make A is better than A+, but we must admit that we cannot think of any.

now demonstrated, it seems that the only principled way in which their proposal can avoid the Repugnant Conclusion (without becoming redundant) is by allowing the desire to avoid inequality to play a seemingly implausibly strong role; so strong that we would sometimes have to say that one population is no better than another population even though everyone in the one population is better off than anyone in the other population. In other words, they then violate the Pareto principle and thus face the levelling down objection. This is a pretty hefty price to pay in order to avoid the Repugnant Conclusion.

Acknowledgements We are grateful for comments and suggestions received when presenting this paper at Princeton University. Additional thanks to Nir Eyal, Anders Herlitz, and two anonymous referees, for very helpful comments and suggestions. Financial support from Riksbankens Jubileumsfond is gratefully acknowledged.

Funding Open access funding provided by Stockholm University.

Declarations

Conflict of interest The authors declare that no conflict of interest.

References

Arrhenius, G. (2000). Future Generations: A Challenge for Moral Theory. Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2:170236

Arrhenius, G. (2005). Superiority in value. *Philosophical Studies, 123*(1/2), 97–114. https://doi.org/10.1007/s11098-004-5223-0

Arrhenius, G. (2009a). Can the person affecting restriction solve the problem in population ethics? In M. A. Roberts & D. T. Wasserman (Eds.), *Harming future persons: ethics, genetics and the nonidentity problem* (pp. 291–316). Springer.

Arrhenius, G. (2009b). Egalitarianism and population change. In A. Gosseries & L. Meyer (Eds.), *Intergenerational justice* (1st ed., pp. 325–349). Oxford University Press.

Arrhenius, G. (2011). The Impossibility of a Satisfactory Population Ethics. In H. Colonius & E. N. Dzhafarov (Eds.), *Descriptive and normative approaches to human behavior, advanced series on mathematical psychology* (pp. 1–26). World Scientific Publishing Company.

Arrhenius, G. (2013). Egalitarian concerns and population change. In O. Frithjof Norheim, N. Eyal, S. A. Hurst, & D. Wikler (Eds.), *Inequalities in health concepts, measures, and ethics* (pp. 74–91). Oxford University Press.

Arrhenius, G. (2016). Population ethics and different-number-based imprecision. *Theoria, 82*(2), 166–181. https://doi.org/10.1111/theo.12094

Arrhenius, G. (2021a). Incommensurability and vagueness in population axiology. In A. Herlitz & H. Andersson (Eds.), *value incommensurability: ethics, risk, and decision- making*. Routledge.

Arrhenius, G. (2021b). Population ethics and conflict-of-value imprecision. In J. McMahan, T. Campbell, & J. Goodrich (Eds.), *Ethics and existence: The legacy of derek parfit*. Oxford University Press.

Arrhenius, G. (2022). Population paradoxes without transitivity. In G. Arrhenius, K. Bykvist, T. Campbell, & E. Finneron-Burns (Eds.), *The Oxford handbook of population ethics* (pp 180–203). https://doi.org/10.1093/oxfordhb/9780190907686.013.11

Arrhenius, G. (forthcoming). Population ethics: The challenge of future generations. Oxford University Press

Arrhenius, G., & Rabinowicz, W. (2005). Millian superiorities. *Utilitas, 17*(2), 127–146. https://doi.org/10.1017/S0953820805001494

Arrhenius, G., & Rabinowicz, W. (2015). Value superiority. In I. Hirose & J. Olson (Eds.), *The oxford handbook of value theory* (pp. 225–248). Oxford University Press.

Blackorby, C., Bossert, W., & Donaldson, D. (1995). Intertemporal population ethics: critical-level utilitarian principles. *Econometrica, 63*(6), 1303–1320. https://doi.org/10.2307/2171771

Blackorby, C., Bossert, W., & Donaldson, D. (1997). Critical-level utilitarianism and the population-ethics dilemma. *Economics and Philosophy, 13*(02), 197–230. https://doi.org/10.1017/S026626710000448X

Blackorby, C., Bossert, W., & Donaldson, D. (2005). *Population issues in social choice theory, welfare economics, and ethics*. Cambridge University Press.

Brentano, F. (1969). *The Origin of Our Knowledge of Right and Wrong* (R. M. Chisholm, Ed.) Routledge

Broome, J. (1999). *Ethics out of economics*. Cambridge University Press.

Broome, J. (2004). *Weighing lives*. Oxford University Press.

Bykvist, K. (2007). The good, the bad, and the ethically neutral. *Economics and Philosophy, 23*(01), 97–105. https://doi.org/10.1017/S0266267107001253

Carlson, E. (2022). On some impossibility theorems in population ethics. In G. Arrhenius, K. Bykvist, T. Campbell, & E. Finneron-Burns (Eds.), The Oxford handbook of population ethics (1st ed., pp 204–225). https://doi.org/10.1093/oxfordhb/9780190907686.013.14

Carlson, E. (1998). Mere addition and two trilemmas of population ethics. *Economics and Philosophy, 14*(02), 283–306.

Fehige, C. (1998). A Pareto principle for possible people. In C. Fehige & U. Wessels (Eds.), *Preferences* (pp. 508–543). W. de Gruyter.

Feldman, F. (1997). *Utilitarianism, hedonism, and desert: essays in moral philosophy*. Cambridge University Press.

Hájek, A., & Rabinowicz, W. (2022). Degrees of commensurability and the repugnant conclusion. *Noûs, 56*(4), 897–919.

Hudson, J. L. (1987). The diminishing marginal value of happy people. *Philosophical Studies, 51*(1), 123–137. https://doi.org/10.1007/BF00353967

Kavka, G. S. (1982). The paradox of future individuals. *Philosophy & Public Affairs, 11*(2), 93–112.

Ng, Y.-K. (1989). What should we do about future generations? *Economics and Philosophy, 5*(02), 235–253.

Parfit, D. (1984). *Reasons and persons* (1991st ed.). Clarendon.

Parfit, D. (1986). Overpopulation and the quality of life. In P. Singer (Ed.), *Applied ethics* (pp. 145–164). Oxford University Press.

Parfit, D. (1997). Equality and priority. *Ratio, 10*(3), 202–221. https://doi.org/10.1111/1467-9329.00041

Parfit, D. (2014). *How we can avoid the repugnant conclusion*. University of Oxford, Faculty of Philosophy.

Parfit, D. (2016). Can we avoid the repugnant conclusion? *Theoria, 82*(2), 110–127. https://doi.org/10.1111/theo.12097

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

Thomas, T. (2018). Some possibilities in population axiology. *Mind, 127*(507), 807–832. https://doi.org/10.1093/mind/fzx047