Transitivity, Majority Rule, and the Repugnant Conclusion

Brian Hedden

to appear in *Conditionals, Probability and Decision: Essays in Honour of Alan Hájek*

1 Almost Betterness and the Repugnant Conclusion

Alan Hájek is one of the foremost contributors to formal epistemology and decision theory in recent decades. But recently, in a paper with Wlodek Rabinowicz (Hájek and Rabinowicz 2022; henceforth H&R), he has turned his attention to the biggest problem in population ethics, namely how to avoid the Repugnant Conclusion, the claim that for any population consisting of many people all with excellent lives, there is some much larger population that is better, even though its members have lives only barely worth living.

They focus on this spectrum paradox from Parfit (1984): Consider a sequence of possible populations, A through Z. A consists of many people (10 billion, say), all of whom live excellent lives. Each successive population consists of many more people than its predecessor, all with only slightly worse lives. To be concrete, assume that the percentage difference in population size and welfare between a population and its predecessor is the same throughout the sequence. For instance, each population might consist of ten times as many people as its predecessor, each with 95% of the welfare of the people in the preceding population. Z consists of an utterly astronomical number of people, all with lives barely worth living. Intuitively, B is better than A, for the vast increase in population more than compensates for the slight decrease in welfare. For the same reason, C is better than B, D is better than C, ..., and Z is better than Y. By the transitivity of 'better than,' it follows that Z is better than A.

This conclusion is one instance of the Repugnant Conclusion, which is the universally quantified claim that for *any* A-like population, there is some Z-like population which is better. But it seems that we could run an analogous argument for any A-like population, considering a sequence of populations of rapidly increasing size and slowing decreasing quality of life, each of which is intuitively better than its predecessor, and leading to a Z-like population. If so, then the transitivity of 'better than' would secure the full-blown Repugnant Conclusion.

We face a paradox, then: It seems that each population in the sequence is better than its predecessor, betterness is transitive, and yet the last population is not only not better, but indeed worse, than the first. H&R give a novel solution to this paradox. Their solution proceeds in three steps. The first step is to embrace the possibility of incommensurability, whereby one thing can be neither better than, nor worse than, nor equally good as another. They ground this commitment to the possibility of incommensurability in a fitting-attitudes account of value. On this account, A is better than B just in case, and because, one ought to prefer A over B, and A and B are equally good just in case, and because, one ought to be indifferent between A and B (i.e. one ought to equi-prefer them). A and B are incommensurable just in case, and because, it is permissible for one to prefer A over B, but it is also permissible for one not to, and instead to be indifferent between them or to prefer B over A. (They assume that permissible preferences must be complete, such that one must either prefer A over B, or prefer B over A, or be indifferent between them.[1])

The second step is to claim that each better population in our sequence is incommensurable with its predecessor, rather than strictly better. The relation of incommensurability is non-transitive, and so each population can be incommensurable with its predecessor, even though the first is strictly better than the last.[2]

[1] But see their footnote 17, where they consider the possibility of relaxing this assumption.

[2] This is their 'first pass' proposal. But they note (p. 8) that they do not need *each* population to be

This appeal to incommensurability as a response to the spectrum paradox is not original to H&R; Parfit (2016) makes the same move. But as H&R note, this bare appeal to incommensurability fails to explain why we are inclined to judge, of each population, that it is better than its predecessor. If a given population is incommensurable with its predecessor, why are we inclined to judge that it is better? Why don't we instead feel flummoxed, or confidently judge that they're incommensurable, or sometimes judge that they're equally good, or sometimes judge that the population is actually worse than its predecessor? As H&R (8) put it, 'Parfit's solution cries out for an error theory.'

The third step in H&R's solution, then, is to give such an error theory. According to them, there are 'degrees of incommensurability.' If most permissible preference orderings rank A over B (or if most rank them equally), then A and B are almost commensurable. But if the permissible preferences are wildly divergent, such that e.g., a third of them rank A over B, a third rank B over A, and a third rank A and B equally, then A and B are highly incommensurable. More specifically, they introduce the technical notion of 'almost better than,' where A is almost better than B just in case a majority of permissible preference orderings rank A over B.[3]

Just as incommensurability is non-transitive, so is almost betterness, so defined. This is because majority rule is non-transitive. It is possible for a majority to rank A over B, a different majority to rank B over C, and a majority not to rank A over C. Indeed, majority rule is not even acyclic; it is possible for a majority to rank A over B, a different majority to rank B over C, and a different majority to rank C over A. This opens up the possibility that each population in our sequence is incommensurable with, but 'almost better than,' its predecessor, even though the last is not almost better than the first. In fact, H&R claim

incommensurable with its predecessor. Instead, they just need some populations to be incommensurable with their predecessors and the others to be better than their predecessors. Nothing in the present paper turns on this, and so I will continue to work with their first pass proposal, for ease of exposition.

[3] Strictly speaking, they leave open the possibility that A's being almost better than B requires A being ranked above B on some supermajority of permissible preference orderings. But nothing hangs on whether a simple majority or instead a supermajority is required. And so for ease of exposition, I will use the simple majoritarian definition of almost betterness.

something slightly stronger: each population is almost better than its predecessor, and yet the last is strictly worse (dispreferred on all permissible preference orderings) than the first.

This yields the error theory: we are apt to confuse 'almost better' with 'strictly better.' We are apt to confuse its being the case that *most* permissible preference orderings rank A over B with its being the case that *all* of them do so. That's why, despite their incommensurability, we are inclined to judge each population to be better than its predecessor, rather than judging it to be worse than or equally good as its predecessor, or realising that they're incommensurable, or simply feeling flummoxed.

It's an ingenious proposal, and there is much to learn from it. In the sections that follow, I will raise some challenges for it, having to do with the precise nature of the permissible preference orderings that they appeal to in order to generate the non-transitivity of almost betterness. I will then propose an alternative response to the paradox, one which appeals not to the non-transitivity of almost betterness, but rather to the soritical nature of the paradox: we judge populations early in the sequence to be better than their predecessors and populations late in the sequence to be worse than their predecessors, but we cannot identify a point in the sequence where things stop getting better and start getting worse.

Before moving on, however, I end this introductory section by suggesting that H&R's commitments to incommensurability and to the fitting-attitudes account of value are extraneous. These commitments can, and should, be dropped. First, H&R appeal to incommensurability as opposed to vagueness. Take a case where A is better along some dimensions of value than B but worse along others, and there seems to be no uniquely privileged way to weight and aggregate these competing dimensions of value. Hence we are not inclined to judge that A is better than B, nor are we inclined to judge that A is worse than B. But we are perhaps also not inclined to judge that A and B are equally good, for if we slightly improved A along one dimension of value, this would not lead us to judge that A is now better than B. H&R say that in such cases, A and B are incommensurable. But many other theorists think that in such case, it is determinately the case that A is either better than, worse than, or equally good as

B, but it is indeterminate which of these three value relations holds. Of course, there is an extensive literature debating whether apparent incommensurability is really just vagueness (Broome 1997; Chang 2002; Elson 2017; Dorr et al. 2023). I am inclined to think that it is, but I won't argue the point here. I simply want to note that H&R don't need to take a stand on the question. For their proposal could be reformulated in terms of vagueness without loss, as they acknowledge in §9. Just replace all talk of permissible preference orderings with talk of admissible precisifications of the vague term 'better than.' I'll adopt the latter terminology in what follows.

Second, H&R ground their view of incommensurability in a fitting attitudes account of value. Indeed, they claim that the fitting attitudes account of value points in favor of incommensurability over vagueness. I am skeptical. A fitting attitudes theorist could well deny incommensurability and instead endorse vagueness by claiming that the reason it's indeterminate which (complete) betterness ordering is correct is that it's indeterminate which (complete) preferences one ought to have. Moreover, it is problematic to put much weight on a fitting attitudes account here, because the phenomenon in question is ubiquitous, showing up with all sorts of multidimensional concepts, not just normative ones. Take *biodiverse*. Suppose one ecosystem has a couple more species than another, but those species are slightly less morphologically and genetically diverse. We're not inclined to judge that the one is more biodiverse than the other, but nor are we inclined to judge that they're exactly equally biodiverse, since adding one species to one of the ecosystems wouldn't lead us to judge that it is now more biodiverse than the other. We again face the question of whether we have incommensurability or vagueness. But here, a fitting attitudes account would be out of place. We certainly don't want to be fitting attitudes theorists about biodiversity! Similarly for other multidimensional concepts, like *democratic* or *strong*.

There's an important lesson here: the concept *good* is multidimensional (at least if we're value pluralists), but it's not the only such concept. And so many of the thorny philosophical issues that arise from the multidimensionality of *good*—like worries about incommensurability or about vulnerability to a repugnant conclusion—arise with other multidimensional concepts as well. It would be too quick to insist that all such concepts must be treated in exactly the same way. But a uniform treatment is a natural default to aim at, and insofar as a response to some problem arising from the multidimensionality of *good* doesn't generalize, that should give us pause. I'll return to this point briefly in §4.

2 Thresholds

The core of H&R's account, as I said, is that is population is almost better than its predecessor, while the last is strictly worse than the first. Understanding almost betterness as betterness on a majority of admissible precisifications, their account thus relies on the fact that majority rule can yield non-transitive, and indeed cyclic, orderings, in the sense that it's possible for a majority to rank x over y, a majority to rank y over z, and for a majority not to rank x over z, and, indeed, to rank z over x. We know this from Condorcet's 'paradox.' Suppose there are three voters. The first ranks x over y over z; the second ranks y over z over x; and the third ranks z over x over y. Then, a 2/3 majority ranks x over y, a 2/3 majority ranks y over z, and yet a 2/3 majority ranks z over x. We can also get cases where a majority ranks x over y, a majority ranks y over z, a majority ranks z over w, and yet *everyone* ranks w over x. Just take the Condorcet paradox and add option w, with each of our three voters ranking w just above x. That's needed if we want to maintain that the last population in Parfit's spectrum paradox is strictly worse, and not merely almost worse, than the first. So the structure of admissible precisifications of betterness that H&R want is certainly *possible*.

But is it plausible? What must the admissible precisifications look like in order for majority rule to yield the results about strict betterness and almost betterness that H&R want? Helpfully, they say quite a bit about this. They say that each admissible precisification must have the following peculiar structure.[4] Each one must be such that when we move along the

[4] In fact, this is a bit too strong. Provided that most of the precisifications look more or less like this, we'll get the same result. But if, e.g., we add in a few that don't, we can still get the same result about what happens on a majority of precisifications.

sequence from A to Z, populations keep getting better for a while until we hit a threshold. The first population after this threshold is the worst one of all, and subsequent populations again get better and better, though never as good as the first one in our sequence. Picture a graph with the populations arranged in sequence along the x-axis and the y-axis representing increasing value. H&R require that each admissible precisification appear as an upward-sloping line segment, followed by a plummet down the y-axis, followed by another upward-sloping line segment which still never reaches as high on the y-axis as where we started. But the admissible precisifications differ in terms of where they place the threshold. For this reason, the last population is ranked below the first on all admissible precisifications, but each population is ranked above its predecessor on a majority, namely all of them except the ones which place the threshold right before that population.

In fact, the idea that betterness has this sort of structure has a venerable pedigree. It was suggested by Parfit (1984) himself, and since then, it's been taken up by Kitcher (2000), Parfit (2016), Thomas (2018), and it has received its most detailed development from Nebel (2022).

The most intuitive development of this view is a lexical threshold totalism that appeals to Mill's famous distinction between higher and lower pleasures. The idea is that welfare is itself multidimensional, with at least two dimensions: the first has to do with how good that life is with respect to certain 'higher goods' like listening to Mozart and appreciating poetry, and the second has to do with how good that life is with respect to 'lower goods' like listening to muzak and playing pushpin. We now suppose that higher goods are lexically prior to lower goods. An implausibly crude lexical priority view says that any amount of the higher goods is better than any amount of the lower goods; in the terminology of Arrhenius and Rabinowicz (2005), higher goods would then be *strongly superior* to lower goods. A more plausible alternative says that there is *some* amount of the higher goods that is better than any amount of the lower goods; higher goods would then only be *weakly* superior to lower goods.[5]

[5] Note, however, that weak superiority collapses into strong superiority in the presence of transitivity,

Weak superiority is enough for H&R's purposes. It naturally yields the sort of betterness rankings that H&R are after. In populations early on in the sequence, the people all enjoy some amount of the higher goods (along with, perhaps, some amount of the lower goods). For a while, subsequent populations involve more and more people all still enjoying some amount of the higher goods, but less of them in each subsequent population. But then we reach a threshold at the point where the people's lives have diminished to the point where they no longer enjoy any of the higher goods, though they still enjoy the lower goods. The first population after this threshold is worse than all the ones that came before. But afterwards, populations keep getting better, since we still have more and more of the lower goods in total, but they never reach the level of goodness of the first population.

The different admissible precisifications of betterness all have this structure, but they differ in where they place the threshold, corresponding to different admissible ways of placing the boundary between higher and lower goods—where to place the boundary as we shade gradually from poetry into pushpin, from Mozart into muzak, and so on. This is plausible, since it's doubtful that higher and lower goods correspond to natural kinds that carve nature at its joints. Instead, the distinction between higher and lower goods is vague.

It's an attractive picture. But I don't find it fully convincing. In the remainder of this section, I raise two problems for it. The first has to do with lexical threshold totalism itself, while the second has to do with the specific purpose to which H&R put it.

The first problem is that there may be sequences of populations leading to the sort of population referenced in the Repugnant Conclusion, but where we never cross a threshold in terms of the quality of the goods enjoyed by people. There may be sequences of populations which do not differ in terms of the quality of goods enjoyed by their members, but rather in the quantity of a given quality (high or low) of goods enjoyed, or in the balance of goods over

completeness, and a separability principle stating that if one population is at least as good as another, then the former added to any other third population is at least as good as the latter added to that same third population. See Jensen (2008) and Nebel (2022). Note that H&R assume that each admissible precisification of 'better than' is transitive and complete, so they must reject separability in order to avoid weak superiority collapsing into the more implausible strong superiority.

bads.

Consider various interpretations of the lives in population Z. The case where lexical threshold totalism is most obviously relevant is one where Z consists of what Portmore (1999) calls *drab lives* (see also Parfit 2016). These are oyster-like lives with the barest flickers of pleasure, or lives spent listening to muzak, or lives spent playing pushpin. But Portmore also mentions the idea of *short-lived lives* and *roller-coaster lives*. The former are only just long enough to enable enjoyment of some positive quantity of the higher goods. The latter fluctuate between periods where they are good overall (including the enjoyment of higher goods) and periods where they are bad overall, with the good periods only just outweighing the bad periods. Venkatesh (2020) also considers *Job lives*, *Cindarella lives*, and *chronically irritated lives*. Job lives consist of a life just like one of those in A, but followed by a period of suffering that is only just short enough so that the good initial period barely outweighs the bad later period. Cindarella lives are the reverse. Chronically irritated lives are just like A-lives, except with a constant level of background pain, whose badness is only just barely outweighed by the goods in that life.

A lexical threshold totalism based on the distinction between higher and lower goods cannot yield the result that A is better than Z for all these different interpretations of Z-lives. While drab lives don't contain any of the higher goods, short-lived lives, roller-coaster lives, Job lives, Cindarella lives, and chronically irritated lives all do contain some amount of the higher goods, and there are enough of them in Z that there will be more of the higher goods in total in Z than in A.

Of course, insofar as we want to distinguish between higher and lower goods, we might also want to distinguish between higher and lower bads. As Kagan (2014) notes, there has been almost no discussion of such a view. But it is not implausible. Perhaps the higher bads (higher being worse, here) are exceptionally intense pains. Or perhaps, as Kagan prefers, they might be bads with a social element, like torture as opposed to naturally-occuring pain of equal intensity. If we add the distinction between higher and lower bads and say that the latter are lexically prior to the former, this will help with the case of chronically irritated lives. For in order to be only barely worth living, the constant level of background pain might have to be sufficiently intense or sufficiently social in nature that it constitutes a higher bad, whereas the lives in A contain no higher bads. But the distinction between higher and lower bads won't help lexical threshold totalism secure the verdict that A is better than Z when the Z-lives are interpreted as roller-coaster, Job, or Cindarella lives. For in those cases, we can plausibly get lives that are only barely worth living just by increasing the duration of the bad periods. In the sequence from A to Z, we don't need to cross any threshold whereby people lose all higher goods where previously they had some, or where they start to suffer higher bads where previously they suffered none.

Of course, the repugnance of the Repugnant Conclusion might well vary depending on how we're understanding the Z-lives. It is most repugnant to think that there is some huge population of drab lives that is better than A. It is perhaps less repugnant to think that there is some huge population of the other sorts of Z-lives that is better than A. So it is open to a lexical threshold totalism to just appeal to the distinctions between higher and lower goods, and between higher and lower bads, and thereby rank A over Z only when the Z lives are drab lives or perhaps chronically irritated lives, while embracing the conclusion that Z is better than A when the Z lives are roller-coaster, Job, or Cindarella lives. I suspect that most theorists would not wish to embrace this more concessive option, however.

It is still possible for lexical threshold totalists to formulate their view in such a way that it avoids the Repugnant Conclusion, no matter how we interpret the Z-lives. They can draw a distinction not between higher and lower goods, or between higher and lower bads, but directly between great lives and merely good lives. And they can say that there's some number of the former that is better than any number of the latter. The distinction between great lives and merely good lives then needn't have to do only with the quality of the goods and bads they contain, but also (or instead) with the balance of goods over bads therein.

But this more generally-applicable version of lexical threshold totalism is, I think, less well-motivated than the original version based on the Millian distinction between higher and lower goods (and, perhaps, an analogous distinction between higher and lower bads). There does seem to be a difference in kind, albeit a vague one, between the good of enjoying Mozart and the good of enjoying muzak, or between the bad of suffering torture and the bad of suffering from a headache. It is much less intuitive to think that there is a difference in kind, even a vague one, between lives with above some level of net goods over bads and lives with below that level of net goods over bads. And so it is much less intuitive to think there is some lexical priority for lives of the former sort over lives of the latter. I concede, though, that I have no knock-down argument here.

While the first problem had to do with the possibility that there might be too few thresholds in our sequence of populations, namely none, the second problem has to do with the possibility that there might be too many. Take the original version of lexical threshold totalism based on a Millian distinction between higher and lower goods. It's crucial to H&R's account both that there is some line between higher and lower goods (along with the claim that some amount of the former outweighs any amount of the latter) *and* that it is a vague matter where exactly to that line lies. If it weren't vague, for instance because the higher and lower goods each constituted some natural kind, then different admissible precisifications couldn't differ in terms of where they place the relevant threshold, and whichever population in our sequence is just after the threshold (such that the people therein enjoy no positive amount of the higher goods) would be strictly worse than its predecessor, not almost better, as H&R want.

But the vagueness of higher vs. lower goods is a double-edged sword. For if there are no natural kinds here, then insofar as it's admissible to divide goods into two kinds—higher and lower—it should presumably be admissible to divide goods into three kinds, or four, and so on. Just as we can divide music into Mozart and muzak, so we could divide it into Mozart, Madonna, and muzak, or into Mozart, Madonna, Mariah, and muzak, and so on.

The point is even more obvious when we consider a version of lexical threshold totalism that distinguishes not between types of goods (or bads) but between types of lives, with the distinction drawn in terms of the level of net goods over bads therein. Surely great and merely good lives do not constitute natural kinds. But then, insofar as we can divide lives into two kinds—great and good—with the boundary between them being vague, we could also divide lives into three kinds—great, very good, and good—or into four kinds—awesome, great, very good, and good—and so on.

We can carve up kinds of goods, or kinds of lives, in more fine-grained ways or in more coarse-grained ways, and there is no level of grain that is privileged. But if all such carvings are admissible, then it will no longer be the case that each population is better than its predecessor on a majority of admissible precisifications. In fact, so long as we insist that the last population must be worse than the first on all admissible precisifications, we'll get the exact opposite result, namely that each population is worse than its predecessor on a majority of precisifications—'almost worse,' in H&R's terminology—which would scuttle H&R's error theory.

Here's why. For any given carving, we can consider its 'mirror image.' Wherever the original draws a line between kinds of lives, the mirror image doesn't, and wherever the original doesn't draw a line, the mirror image does. In terms of our original sequence of populations, we can model this by grouping certain populations together using parentheses. When two populations are in the same group, the lives therein are of the same kind; when they are not, the lives in the populations to the right of the parentheses are of a worse kind than, and lexically posterior to, the lives in the populations to the left. So, when two populations are in the same group, the one to the right is better than the one to the left, since it contains many more of the same kind of lives, albeit slightly worse ones. But when two populations aren't in the same group, the one to the right is worse than the one to the left, since it contains only lives of the worse kind.

Here, then, is the key: For any two adjacent populations, whenever one admissible carving yields a precisification on which the one to the right is better than the one to the left, its mirror image yields a precisification on which the one to the right is worse than the one to the left, and vice versa. This is because whenever the original carving has them in the same group, and so the one to the right is better, the mirror image has them in different groups, and so the one to the right is worse. Hence, a given carving and its mirror image cancel each other out in calculating how adjacent populations are ranked by a majority of precisifications.

But there's one carving—namely the maximally fine-grained one—whose mirror image is inadmissible. The maximally fine-grained carving has each population in its own group, yielding the precisification on which each population in the sequence is worse than its predecessor. This precisification is admissible. Its mirror image, the maximally coarse-grained carving, has each population in the same group, yielding the precisification on which each population in the sequence is better than its predecessor. But this maximally coarse-grained mirror image must be inadmissible, else we would not get H&R's desired result that the last population is worse than the first on *all* precisifications.

So suppose maximal vagueness reigns. It's a vague matter not only *where* to draw distinctions between kinds of lives, but also *how many* such distinctions are to be drawn. Then, every admissible carving has an admissible mirror image which together cancel out when we're calculating majority betterness, with one exception. The maximally fine-grained carving yields a precisification on which each population is worse than its predecessor, and its mirror image is inadmissible. Hence each population in the sequence will be worse on a majority—half plus one—of the admissible precisifications, and hence almost worse rather than almost better.

3 Mountains

Take a step back and consider again the original sequence of populations from A to Z. Suppose I asked you to draw a graph, with the x-axis representing the different populations arranged in sequence and the y-axis representing how good those populations are. What sort of shape would you draw? As we've seen, H&R think you should draw a line segment with positive slope, followed by another line segment, again with positive slope, but where all points on the latter line segment are below all the points on the former. I think this sort of shape is strange. At first blush, I'd be inclined to draw a shape like a mountain, with the curve sloping steadily upward for a while before reaching a peak and then sloping downward, eventually reaching a point below where we started. Of course, I would have no strong view about where the peak should be located, and so I wouldn't favor one mountain-shaped graph over a different one which differed in the location of the peak. This might be indicative of vagueness: all the admissible precisifications of betterness yield mountain-shaped graphs, but different ones place the peak in different places.

If that's right, then the set of admissible precisifications of betterness is *single-peaked* with respect to the set of populations in our sequence. As the name suggests, this means that there is some way of arranging the populations—in this case, it's just the natural, intuitive, alphabetical arrangement we've been considering all along—such that each admissible precisification has some highest-ranked population, and populations are ranked lower as they get further away from that highest-ranked population. More formally, a set of orderings satisfies single-peakedness just in case there is some arrangement of the alternatives such that: for each ordering i in the set and for all population, if $x \succcurlyeq_i y$ and y lies between x and z in the arrangement, then $y \succ_i z$. (Note that single-peakedness doesn't actually require all orderings to be mountain-shaped, for it's compatible with orderings on which the most preferred alternative comes first or last in the arrangement; these orderings would yield graphs that look like constant slopes, either upward or downward.)

Black (1948) famously proved that if a set of orderings over some set of alternatives satisfies single-peakedness, then majority rule yields a transitive 'overall' ordering.[6] In the present terminology, this means that if the set of admissible precisifications satisfies single-peakedness with respect to the populations in our sequence, then majority betterness is transitive, contra H&R. This, in turn, rules out the possibility that each population in the

[6] Single-peakedness is sufficient but not necessary for majority rule to yield transitive orderings. There are other domain conditions which likewise suffice for the transitivity of majority rule. See Sen (2017, 10.3) for details.

sequence is almost better (better on a majority of precisifications) than its predecessor while the last is not almost better than the first, and indeed is strictly worse.

We noted that H&R are motivated by the thought that we're inclined to judge, of each population in the sequence, that it is better than its predecessor. They give an error theory in which it's really just that each population is almost better than its predecessor, and we are apt to mistake almost worse for strictly worse. But if single-peakedness holds for admissible precisifications of betterness, then this error theory is unavailable. So what about the original datum that we're inclined to judge each population to be better than its predecessor?

I'm skeptical that this really is a datum. Instead, here's what I think: Near the start of the sequence, we're inclined to judge each population to be better than its predecessor. After all, the lives are still great lives, albeit slightly worse, and there are so many more people enjoying them! But near the end of the sequence, we're inclined to judge each population to be worse than its predecessor. After all, the lives are not at all great, and they're slightly worse than in the preceding population, so it's worse if there are so many more people enjoying these slightly worse, and very mediocre, lives! But there's no particular population in the sequence at which we definitively switch from being inclined to judge it better than its predecessor to being inclined to judge it worse. It's just like in a forced march Sorites. And it's the pattern of judgments that's nicely captured by the claim the set of admissible precisifications is single-peaked: It's determinately the case that populations get better for a while and then get worse and worse, but it's indeterminate exactly where the switch happens.

Single-peakedness is not only intuitive; it is also generated by a number of popular axiologies. Of course, totalism and averagism are like this. Totalism yields a graph that is always sloping upwards, while averagism yields a graph that is always sloping downwards. (Totalism is determinately ruled out on the assumption that Z is determinately worse than A.)

Single-peakedness would also follow if critical-level utilitarianism is determinately true but it is a vague matter where the critical level is. On critical-level utilitarianism (Blackorby et al. 1997; Broome 2004), lives may not contribute positively to the overall value of an population, even though they are good enough to be worth living. Such lives have personal value but not moral value. More specifically, this view says that there is some critical level of welfare which is above the level at which a life becomes worth living for the person who lives it, and which is such that the value of a population is found by adding up, across all people, the difference between their welfare and the critical level.

This straightforwardly yields the desired result that the last population in our sequence is worse than the first, at least assuming that the lives in the last population are below the critical threshold. This is plausible, since they are said be to barely worth living, and we can assume that the critical level is significantly, and not just barely, above the level at which a life becomes worth living. Critical level utilitarianism thereby avoids the Repugnant Conclusion.

Provided that we do cross that critical level at some point in our sequence of population, then the result will be a mountain-shaped graph. For populations that are before that point in our sequence, each one is better than its predecessor, for the vast increase in population size more than compensates for the slight reduction in everyone's welfare and, hence, in the difference between everyone's welfare and the critical level. But as soon as we cross that point, populations get worse. Indeed, if there's a population where everyone's welfare is right at the critical level, that population will be neutral in value: neither good nor bad overall. Subsequent populations get worse and worse as we get more and more people whose welfare levels are further and further below the critical level. But if it's vague where the critical level is, it's vague where we reach a peak, and different admissible precisifications can put it in different places.

Single-peakedness would also follow if some 'compromise' or 'variable value' view is determinately true. Compromise views are inspired by the thought that averagism looks plausible when comparing very large populations with each other, while totalism looks plausible when comparing small populations with each other. Compromise views therefore try to mimic totalism for small population sizes and averagism for large population sizes. It's easy to see that compromise views will yield mountain-shaped betterness graphs. Early in the sequence, populations are relatively small, and since each population has greater total welfare than its predecessor, it is better. Late in the sequence, populations are large, and since each population has lower average welfare than its predecessor, it is worse.

Some prominent compromise theories achieve this result by holding that the value of a population depends on two dimensions of value, average welfare and total welfare, which have diminishing marginal rates of substitution with respect to each other. This means that as you have less and less of one, you're willing to trade less and less of it to gain a given amount of the other. This is, roughly speaking, how things work with Hurka's (1982) variable value view and Ng's (1989) Theory X'. I'll just discuss the latter. Whereas totalism says that the value of a population is the product of average lifetime welfare and population size, Ng's Theory X' says that the value of a population is the product of average lifetime welfare and *some strictly increasing, concave, and upper-bounded function of* population size. Provided this upper bound is sufficiently low, we get the result that no Z-like population with lives barely worth living, not matter how big, is better than our population A. Plausibly, it is vague or indeterminate which strictly increasing, concave, and upper-bounded function is the right one, and different choices will different mountain-shaped betterness graphs, differing in the location of the peak.

To sum up this section: When confronted with our sequence of populations, I find it natural to judge that initially, each population is better than its predecessor, but toward the end of the sequence, each population is worse than its predecessor. That is, when we array our populations in sequence, it is natural to graph their goodness as a mountain-shaped curve, getting better and better for a while before reaching a peak, and thereafter getting worse and worse, and ending lower than where we started. This sort of 'single-peaked' graph is yielded by some popular axiologies, including critical-level utilitarianism and compromise views like those of Hurka and Ng. (Trivially, totalism and averagism likewise yield single-peakedness.) If all admissible precisifications of betterness have this sort of structure, then the set of admissible precisifications is single-peaked with respect to our set of populations, and so almost betterness must be transitive, contra H&R. If single-peakedness holds for admissible precisifications of betterness, we can give an alternative error theory to that proposed by H&R: We judge that populations near the start of our sequence are better than their predecessors, and we judge that populations near the end of the sequence are worse than their predecessors, but we cannot identify a privileged point where the switch happens and populations change from being better than their predecessors to being worse.

4 Uniformity and Diversity

So who is right, me or H&R? ('Neither' is also a possibility.) Here is one possible argument in favor of my approach. I noted earlier that *good* is not the only multidimensional concept around. Multidimensionality is ubiquitous. And just as we can get a spectrum paradox for goodness qua population, so we can get analogous spectrum paradoxes for other multidimensional concepts.

Some of these analogous cases still lie within the normative realm. Suppose we are pluralists about welfare, such that there are multiple kinds of goods which make a life go well. For instance, welfare might have perfectionist goods as constituents, where these perfectionist goods come in different kinds, like theoretical excellence and practical excellent. Now start with one life, which has a substantial amount of theoretical excellence and a substantial amount of practical excellence. The next life in the sequence has slightly lower theoretical excellence (95% of that of its predecessor, say) but much greater practical excellence (10 times that of its predecessor, say). It seems better overall, for the vast increase in practical excellence more than compensates for the slight reduction in theoretical excellence. The next life again has slightly lower (95%) theoretical excellence and much greater (10x) practical excellence. Again it seems better overall. And on and on until we reach a life that has a huge amount of practical excellence but only the slightest amount of theoretical excellence. It seems worse than the first life. We could do the same thing with moral value pluralism.

Hare (ms) gives several analogous cases that lie outside the normative realm. Consider strength, with underlying dimensions of tensile strength and compressive strength. Start with a metal that has high tensile strength and high compressive strength. The next metal in the sequence has slightly lower tensile strength (95% of that of its predecessor, say) but much greater compressive strength (10 times that of its predecessor, say). It seems stronger overall, for the vast increase in compressive strength more than compensates for the slight reduction in tensile strength. The next metal again has slightly lower (95%) tensile strength and much greater (10x) compressive strength. Again it seems stronger overall than its predecessor. And on and on until we reach a metal that has unbelievably high compressive strength but only the slightest amount of tensile strength. It seems weaker than the first metal, for a slight tug would pull it apart.

Next consider baldness, with underlying dimensions of hair distribution and hair number. We start off with someone with a pretty high number of hairs which are pretty evenly distributed on his head. The next man in the sequence has vastly more (10 times more, say) hairs, but they are slightly less evenly distributed (95% as evenly distributed, say). He seems less bald, or more hirsute, than his predecessor. The next man again has vastly more hairs, which are again slightly less evenly distributed. Again he seems less bald, or more hirsute, than his predecessor. And on and on until we reach a man who has an unbelievably large number of hairs, but they're all localized in a tiny portion of the crown of his head. He is balder, and less hirsute, than the first man in our sequence.

This is just the tip of the iceberg. We can consider analogous sequences, yielding analogous paradoxes, for biodiversity (varying the number of species and their morphological diversity), size (varying volume and mass), and a host of other multidimensional concepts.

There are two observations I want to make: First, a lexical threshold view like that embraced by H&R seems completely inapplicable to these other paradoxes. We would have to posit some difference in kind between metals with 'high' tensile strength and those with 'low' tensile strength, which is innocuous enough. But then we'd have to say that, holding fixed that we have high tensile strength, slight reductions in tensile strength can be compensated for by sufficiently large increases in compressive strength, and that the same is true when we hold fixed that we have low tensile strength. But we'd have to say that crossing the vague threshold from high to low tensile strength outweighs any increase in compressive strength, so that it always yields a material that is weaker overall. I find this wildly implausible.

Second, single-peakedness seems very plausible as a constraint on admissible precisifications of 'good qua life,' 'strong,' 'bald,' and the like in these sequences. I think that the lives get better and better (greater overall perfection) for a while as we move through our sequence, but then the reach a peak and start to get worse and worse. Similarly, the metals get stronger and stronger for a while as we move through our sequence, but then they reach a peak and start to get weaker and weaker thereafter. And the men get more and more hirsute for a while as we move through our sequence, but then they reach a peak and start to get balder and balder thereafter. And so on.

Moreover, I think something like a compromise view is exactly the right thing to say about these other cases. (There seems to be no plausible analogue of critical level utilitarianism, either, for these other cases.) Hurka (1993, 88) explicitly endorses such a view for trade-offs between perfectionist goods: 'if one excellence has been achieved more than another, the second is more important.' This vindicates the idea that a well-rounded life is particularly valuable. Similarly, I think that the more tensile strength we have, the more tensile strength we'd be prepared to trade for a given increase in compressive strength, if we want an increase in overall strength. The more hairs we have, the more hairs we'd be prepared to trade for a given increase in their evenness of distribution, if we want an increase in overall hirsuteness. And so on.

Does that give us a compelling argument in favour of compromise views in population ethics (and, derivatively, in favor of my preferred solution to the original spectrum paradox of populations)? Do we really want a unified solution to all these paradoxes? Or is some piecemeal approach acceptable, whereby we given different solutions to the various different paradoxes?

I am not sure. I think that unification is a theoretical virtue, and this is one consideration in favor of my preferred solution. But my opponent can point out that despite similarities, there are also some differences between the spectrum paradox for populations and the other spectrum paradoxes. First, in the case of populations, there is the possibility of negative values, which have to do with lives that are not worth living. By contrast, there is no such thing as negative tensile (or compressive) strength, a negative number of hairs on one's head, and so on. Second, populations are unlike metals and heads in that they involve multiple points of view which needn't be all the same. We've been considering populations within which everyone lives the same kind of life and has the same level of welfare, but we can also consider populations where there is some inequality. But there is no analogue of inequality in the cases of strength and baldness.

These differences may be important. Together, they enable us to consider cases of Mere Addition, where we hold fixed the original population but just add some more people with lower but still positive levels of welfare. Averagism, critical level utilitarianism, and compromise views like Hurka's variable value view and Ng's Theory X' entail that sometimes Mere Addition makes things worse overall. After all, Mere Addition can lower average welfare, and the views of Hurka and Ng give far greater weight to average welfare than to total welfare for large populations (and, obviously, averagism always gives it far greater weight). And the Mere Addition could involve adding people below the critical level, in which case critical level utilitarianism will condemn the addition.

The two differences also enable us to consider the option of adding many people with positive welfare, or instead adding one person with negative welfare. Averagism, critical-level utilitarianism, and compromise views like those of Hurka and Ng will say that in some such cases, it is better to add the one person with negative welfare than to add the many people with positive welfare. This is the Sadistic Conclusion (Arrhenius 2000). For averagism and the views of Hurka and Ng, this is because they sometimes give exclusive weight—or nearly so—to average welfare, and many people with positive welfare below the average can drag down that average to a greater extent than one person with negative welfare. For critical-level utilitarianism, this is because lives with positive welfare below the critical level are outright disvaluable, and enough of them can be worse than one life with negative welfare.[7]

Where does all this leave us? If we want a unified solution to all of our paradoxes, we must endorse some sort of compromise view in population ethics, for the analogous views are compelling in the cases of good qua life, strength, baldness, and the like. We might stick with the views of Hurka or Ng and try to make their problematic implications more palatable. We might, for instance, insist that we should view populations holistically (Yetter-Chappell ms.), just as we do with careers. There, Mere Addition is not intuitively always good. Muhammad Ali's boxing career would have been better without the last few fights against Larry Holmes and Trevor Berbick (Hurka 1993, 71), which were still good fights and would have added to the quality of lesser fighters' careers. Miles Davis' career would have better without his 1980's albums, even though they were still good and would have added to the quality of lesser

[7] There is at least one compromise view, however, which always endorses Mere Addition and which avoids the Sadistic Conclusion. This is Sider's (1991) Geometrism. Let u_1, u_2,…, u_n be the welfares of the the people in world w with positive welfare, arranged in order of descending welfare, and let v_1, v_2, …, v_m be the welfares of the people with negative welfare, arranged in order of increasing welfare. One population is at least as good as another just in case it has a least as great a geometric value, defined as:

$GV(w) = \sum_{i=1}^{n} \frac{u_i}{r^{i-1}} + \sum_{j=1}^{m} \frac{v_j}{r^{j-1}}$

Geometrism always endorses Mere Addition, since lives with positive welfare always contribute positively to geometric value. And it avoids the Sadistic Conclusion, since in addition to lives with positive welfare always contributing positively to geometric value, lives with negative welfare always contribute negatively to geometric value. And given an appropriate bit of vagueness—perhaps it's vague what the correct value of the discount factor r is—it will yield the result that the set of admissible precisifications of betterness is single-peaked with respect to the populations in our sequence, in line with the kind of error theory I have proposed.

The main objection to Geometrism is its highly inegalitarian nature, which Sider himself takes as a reason to reject it (he treats Geometrism as a proof of concept that a compromise view can endorse Mere Addition and avoid the Sadistic Conclusion). In determining the value of a population, people with higher positive welfare have their welfare weighted more heavily than people with lower positive welfare. For this reason, it would be better to have e.g., one person with welfare 20 and another with welfare 10 than to have one person with welfare 16 and another with welfare 14. Indeed, Geometrism violates what Arrhenius (2000) calls:

> Non-Antiegalitarianism: A population with perfect equality is better than a population with the same number of people, inequality, and lower average (and thus lower total) welfare.

If we're troubled by Geometrism's anti-egalitarianism, then we still haven't found an adequate compromise view. This is no surprise, for Arrhenius (2000) gives an impossibility theory which says that any theory (whether a compromise theory or not) yields at least one of the following problems: it fails to avoid the Repugnant Conclusion, it fails to always endorse Mere Addition, it fails to avoid the Sadistic Conclusion, or it fails to entail Non-Antiegalitarianism.

musicians' careers. For the same reason, viewing populations holistically makes the Sadistic Conclusion more palatable. Perhaps a single outright bad fight would have detracted less from Ali's career than many good-but-not-great ones. Perhaps a single outright bad album (*Decoy*, perhaps) would have detracted less from Davis' career than many good-but-not-great ones.

Many will find it unattractive to view populations holistically. They are unlike careers, in that people have their own points of view, whereas there's nothing it's like to be a boxing match or a jazz album. I am not entirely convinced by my response. We may view lives holistically, even though each time-slice within a life has a point of view. If we care about the shape of a life (Velleman 1991), then we may think that Mere Addition of moments isn't always good. A life of love, happiness, and accomplishment might be made worse by appending a year of mediocrity, even though this year of mediocrity isn't bad in its own right. Similarly, it might be better to append to this very good life an hour of pain rather than a decade of mediocrity. Insofar as we find it attractive to view lives holistically, we might also find ourselves willing to extend this holistic approach to whole populations.

5 Conclusion

H&R give us one axiology—lexical threshold totalism—which, once we sprinkle in some vagueness, yields the result that each population in our sequence is almost better than its predecessor, and yet the last is strictly worse than the first. This yields their preferred error theory for our paradoxical judgments: we mistake almost betterness for strict betterness. After raising some 'internal' challenges to their account, I noted that a host of other axiologies are such that if it's determinately the case that some version of one of them is correct, but indeterminate which one, then the set of admissible precisifications of betterness are single-peaked. This yields my preferred error theory for our paradoxical judgments: we start off judging each population to be better than its predecessor, and late in the sequence we judge each population to be worse than its predecessor, but there's no point at which we're confident in making the switch from judging populations to be better than their predecessors to judging them to be worse.

How can we decide between these axiologies and error theories? I suggested that we might look at analogous spectrum paradoxes for other multidimensional concepts and aim for a unified solution to all of them. This militates against H&R's error theory and its associated commitment to lexical threshold totalism and in favor of my error theory and some version of a compromise view. This is because compromise views seem spot-on as responses to these other paradoxes. Then again, there are also disanalogies between goodness qua population and our other multidimensional concepts, and these disanalogies point toward serious objections to compromise views in population ethics.

I have suggested that these objections might be worth stomaching, since doing so will enable a unified solution to all our spectrum paradoxes. But even if I am wrong, it would be an interesting conclusion in its own right that the spectrum paradox for populations is unique and requires a radically different treatment.

References

Arrhenius, Gustaf. 2000. ‘An Impossibility Theorem for Welfarist Axiologies.’ *Economics and Philosophy* 16(2): 247–66.

Arrhenius, Gustaf and Rabinowicz, Wlodek. 2005. ‘Millian Superiorities.’ *Utilitas* 17(2): 127–46.

Black, Duncan. 1948. ‘On the Rationale of Group Decision-Making.’ *Journal of Political Economy* 56(1): 23–34.

Blackorby, Charles, Bossert, Walter, and Donaldson, David. 1997. ‘Critical-Level Utilitarianism and the Population-Ethics Dilemma.’ *Economics and Philosophy* 13(2), 197–230.

Broome, John. 1997. ‘Is Incommensurability Vagueness?’ in Ruth Chang (ed.) *Incommensurability, Incomparability, and Practical Reason.* Cambridge, MA: Harvard University Press, 67–89.

Broome, John. 2004. *Weighing Lives.* New York: Oxford University Press.

Chang, Ruth. 2002. ‘The Possibility of Parity.’ *Ethics* 112(4): 659–88.

Dorr, Cian, Nebel, Jacob M., and Zuehl, Jake. 2023. ‘The Case for Comparability.’ *Noûs* 57(2): 414-53.

Elson, Luke. 2017. ‘Incommensurability as Vagueness: A Burden-Shifting Argument.’ *Theoria* 83(4): 341–63.

Hájek, Alan and Rabinowicz, Wlodek. 2022. ‘Degrees of Incommensurability and the Repugnant Conclusion.’ *Noûs* 56(4): 987–919.

Hare, Caspar. ms. ‘The Great Spectrum Paradox.’ Massachusetts Institute of Technology.

Hurka, Thomas. 1982. ‘Value and Population Size.’ *Ethics* 93(3): 496–507.

Hurka, Thomas. 1993. *Perfectionism.* New York: Oxford University Press.

Jensen, Karsten Klint. 2008. ‘Millian Superiorities and the Repugnant Conclusion.’ *Utilitas* 20(3): 279–300.

Kagan, Shelly. 2014. ‘An Introduction to Ill-Being.’ *Oxford Studies in Normative Ethics* 4: 261–88.

Kitcher, Philip. 2000. ‘Parfit’s Puzzle.’ *Noûs* 34(4): 550–77.

Nebel, Jacob M. 2022. ‘Totalism without Repugnance.’ In Jeff McMahan, Tim Campbelll, James Goodrich, and Ketan Ramakrishnan (eds.), *Ethics and Existence: The Legacy of Derek Parfit*, Oxford: Oxford University Press, 200–31.

Ng, Yew-Kwang. 1989. ‘What Should We Do About Future Generations?’ *Economics and Philosophy* 5(2): 235–53.

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

Parfit, Derek. 2016. ‘Can we Avoid the Repugnant Conclusion?’ *Theoria* 82(2): 110–27.

Portmore, Douglas. 1999. ‘Does the Total Principle have any Repugnant Implications?’ *Ratio* 12(1): 80–98.

Sen, Amartya. 2017. *Collective Choice and Social Welfare: An Expanded Edition.* Cambridge, MA: Harvard University Press.

Sider, Theodore. 1991. ‘Might Theory X be a Theory of Diminishing Marginal Value?’ *Analysis* 51(4): 265–71.

Thomas, Teruji. 2018. ‘Some Possibilities in Population Axiology.’ *Mind* 127(507): 807–32.

Velleman, J. David. 1991. ‘Well-Being and Time.’ *Pacific Philosophical Quarterly* 72(1): 48–77.

Venkatesh, Nikhil. 2020. ‘Repugnance and Perfection.’ *Philosophy and Public Affairs* 48(3): 262–84.

Yetter-Chappell, Richard. ms. ‘Value Holism.’ Princeton University.