Value Holism

Richard Yetter Chappell*
Princeton University

June 28, 2011

Abstract

This paper defends and relates two forms of value holism: (1) that the welfare value of a moment in one's life depends on what happens at other moments, and (2) that the contributory value of an adding another life to the population depends on what other lives already exist. In the first section, I bolster the standard normative arguments for welfare holism by also exploring neglected empirical and metaphysical considerations. In the second section, I offer a broad overview of how value holists can escape Parfit's 'Repugnant Conclusion', and address the most powerful arguments commonly offered in support of the alternative, 'atomistic', approach.

*Thanks to Nick Beckstead, Dan Halliday, Peter Singer, Gerard Vong, and Helen Yetter Chappell for helpful comments on earlier drafts of this paper.

How does the value of a whole (person, society) relate to the value of its parts (timeslices, individuals)? Utilitarians have traditionally treated the parts as axiologically fundamental, and held that we may simply sum the intrinsic values of the parts to obtain the intrinsic value of the whole. But this presupposes what we may call 'value atomism', or the claim that the contributive value of a part depends only on its intrinsic (non-relational) features. The value of a part is, in other words, taken to be independent of what else exists. In this paper, I wish to challenge this 'independence assumption', and explore the prospects for an holistic axiology that takes extrinsic or relational properties into account. We may thus take *value holism* to be the claim that the contributive value of a part cannot be assessed in isolation; it depends on how the part stands in relation to the whole, and in particular on what other parts there are. To speak even more generally: value holists treat the whole as axiologically fundamental, and hold that parts get their value in virtue of how they influence the 'shape' (so to speak) of the whole. This rather abstract picture should become clearer when applied to the particular examples discussed below.

The first section discusses holism as it applies within a life. I begin by noting some intuitive cases of extrinsic or relational properties affecting the value of our momentary experiences. I then discuss some surprising evidence suggesting that additional moments of pain may be preferable to the immediate cessation of pain. This leads to a closer examination of hedonic duration, or the sense of 'duration' that is relevant to the assessment of extended ex-

periences as more or less pleasant. I offer a speculative reconceptualization of hedonic duration that strikes an appropriate balance between subjectivity and objectivity, and which reinforces the holist's idea that we cannot determine the value of a momentary state of affairs in isolation.

The second section of this paper then applies value holism to the larger question of how individual lives contribute to the value of the world as a whole. This discussion focuses on a central problem of population ethics known as 'the repugnant conclusion'. I will suggest two distinctively 'holistic' solutions to the problem: the first taking the position that 'mere addition' — adding intrinsically good lives to a world — may make the world worse, and the second exploring the value of diversity. In each case, I argue that the contributory value of a life cannot be assessed in isolation, but instead depends on what else there is. I conclude by considering some objections to this position.

1 Whole Lives and their Individual Moments

1.1 Directional Trends and Narrative Structure

We care about not only our net happiness but also its distribution. Better to start miserable and die happy than the reverse, many of us would think. This may be partly explicable through contrast effects: the order in which you eat brussel sprouts and icecream may influence how tasty each seems individually, i.e. changes in order might also lead to changes in intrinsic hedonic quality.

Similarly, the memory of better days may render present drudgery all the worse, and memories of past difficulties may sweeten today's success. But even if we correct for all of that, so that the momentary experiences are truly identical in intrinsic hedonic quality, and differ *only* in order, we may still prefer an upward trend in its own right (Velleman 1991). For in addition to moment-to-moment experiences, many of us also care greatly about the overall 'narrative structure' of our lives. We prefer to live out a story that improves with time rather than declines. (As an extreme case, consider the — perhaps apocryphal — tales of Japanese lovers who leap to their deaths mid-coitus, so as to end their lives in bliss.)

Value holism has important implications for how we think about longevity and life-extension. Paradigmatic atomistic views (e.g. classical utilitarianism) imply that making a life n times longer also makes it n times better (assuming the additional moments are intrinsically as good as the earlier ones).[1] But this seems implausible. A twenty-five year old should prefer dying at age 75 over an even gamble of either living to 125 or dying instantly. This is not just a matter of risk aversion; rather, it seems that the latter option has a lower expected value than the former. Living an extra hundred

[1] Dan Halliday pointed out to me that not all atomists need think this. For example, rather than *summing* the values of the parts, an atomist might obtain the value of the whole by summing the *square roots* of the values of the parts. This would imply that it's better to spread some fixed amount of utility over more parts (i.e., a greater period of time), since, for example, $\sqrt{2x} < (\sqrt{x} + \sqrt{x})$. An alternative atomistic view could model the opposite pattern. But no such atomistic transformation escapes my basic objection: that the significance and value of a good (e.g. having a child) cannot be determined in isolation — it may depend, for example, on whether you have already achieved a like good in the past.

years (even assuming perfect health, etc.) simply isn't *twice as good* as living an extra fifty. That's not to say that life-span always has diminishing marginal utility. If I am fated to die in one year, this might be too short a time to achieve anything of great value. Twenty additional years of life would plausibly be *more* than twenty times as valuable to me, if this would enable such important goods as raising a family. I might reasonably risk instant death for a chance $< \frac{1}{20}$ of such life-extension. It all depends on what I most want out of my life as a whole, and how long is necessary to achieve this.

So we cannot just assume that each additional century (say) adds as much value to a life as the first. Even if the additional centuries are no worse in themselves, they may be less important *overall*. Atomists cannot accommodate this vital insight. Their axiological method is to define the value of a life as a function of the values of its moments. But we have seen that this approach is inadequate. We should thus prefer the holist's axiological method of *directly* evaluating the life as a whole. 'Global' preference theories of wellbeing (such as Parfit's *success theory*) exemplify this holistic approach.[2] Rather than assessing each moment in isolation, we can instead ask ourselves, 'How would I most like my life as a whole to go?' Or: 'What is it that I most want in life?' It is entirely possible that when we categorize our global preferences or 'life goals' into those that could be achieved

[2] Parfit (1984, Appendix I). This is in contrast with an atomistic 'desire satisfactionist' account, whereby we determine the value of each moment simply in virtue of the strength and quantity of presently-held fulfilled desires, and then sum all these values to obtain the value of the whole life.

in a hundred years, and those that could only be achieved with a second hundred, the former group will outweigh the latter. This would justify our unwillingness to risk the former for an even chance at the latter. Even when atomists can't tell the difference, holism affords us the resources to recognize that some years contribute more to our lives than others.

Common sense thus allows that extrinsic features may influence the magnitude of the contributory value of some life experience. We might further question whether even the *valence* of an experience (i.e. as good or bad) can be determined in isolation (Dancy 2004; Kagan 1988; McNaughton 1988). An intrinsically painful experience might be considered good rather than bad if it is embedded within the right context, e.g., the strivings of a marathon runner to stretch her endurance to its limits, or a repentant wrongdoer enduring his just punishment. (A false memory causing the same subsequent feelings of satisfaction would arguably not have the same value, at least if we value the *challenge* — or the *penance* — itself.) When assessing their life as a whole, an agent may reasonably judge a moment of pain to have been non-instrumentally desirable, in light of its relations to other parts of their life, and the elegance of the overall pattern this gives rise to. These examples help to bring out the prima facie plausibility of value holism, and may even convince some readers that they are implicitly committed to the view. But I now want to consider the more surprising empirical phenomenon of 'duration neglect'.

1.2 Preferring the Longer Pain

Kahneman et al. (2003) found that subjects sometimes prefer additional pain, if the appended moments are less painful than what has gone before. In the short episode, subjects are required to hold their hands in very cold water for sixty seconds. In the longer version, this sixty seconds is followed by an additional thirty seconds in which the water temperature gradually rises, though it's still unpleasantly cold. After experiencing both the short and long episodes, and being offered a choice of which to repeat, most subjects choose the long episode. More generally, the duration of an experience seems to have little effect on our retrospective evaluations. Instead, these value judgments are largely determined by some combination of the 'peak' and 'end' moments. Kahneman et al. consider these judgments erroneous on the grounds that they violate *temporal monotonicity*, an independence rule to the effect that "adding moments of pain to the end of an episode can only make the episode worse." (p.401) But perhaps we should instead reject temporal monotonicity in favour of value holism. We should take subjects at their word when they tell us that the added moments of lesser pain made their overall experience better. Or so I will argue.

To begin, let us note the default presumption that each individual is the best authority when it comes to determining the hedonic quality of their own phenomenal experiences. To override this presumption, the atomist might offer two arguments.[3] First, they could point out that the subjects themselves

[3] Versions of each can be found in Kahneman et al. (2003).

would presumably want the pain to end sooner if offered the choice *during* the episode. After all, it is in the nature of painful/aversive experience that it be accompanied by a local preference for its own cessation. So at each moment when we are experiencing pain, we wish it to stop. True enough. Yet when making an overall judgment from 'above the fray', so to speak, the subjects express a conflicting preference, and merely noting the conflict does not tell us how to resolve it. As a general rule, we tend to privilege (reflective) global preferences over (momentary) local ones: such a hierarchy is, after all, essential for the exercise of self-control. So, again, some further reason is required to override the presumption in favour of the subjects' expressed global preference.

Alternatively, the atomist may grant the authority of (informed) global preferences, but suggest that subjects' retrospective assessments here are ill-informed. The lesser pain, being more recently experienced, is more salient in their memory, so perhaps subjects misremember or fail to appreciate how bad the initial period of pain really was. This is a difficult claim to assess. There's something dialectically suspicious about it, insofar as the claim is wholly motivated by our prior reluctance to credit the substantive assessment offered by the subjects, when the credibility of this assessment is precisely the question under dispute. Independent evidence of error might be sought in the fact that most subjects claimed that the longer trial caused "less overall discomfort", which Kahneman et al. (2003, p.403) insist is "simply wrong".[4]

[4] I should note that some subjects made more straightforward factual errors, but that

But for this to qualify as independent evidence of factual error, we must assume that subjects were interpreting 'overall discomfort' to mean 'aggregate momentary discomfort'. This seems unlikely. It's far more plausible to think that subjects were simply reiterating their holistic judgment that the longer trial was less unpleasant on the whole. So these considerations leave us at a dialectical impasse.

1.3 Subjective Time

So far, I have assumed that the third-personal perspective yields accurate insights into the aggregate momentary qualities of experience. For example, I have assumed (with Kahneman et al.) that an episode of pain that has longer *physical* duration must thereby have longer *experienced* duration. But this arguably presupposes a false picture of consciousness — what Dennett (1991, p.107) calls 'the Cartesian Theatre', or a special place in the brain "where the order of arrival equals the order of 'presentation' in experience because *what happens there* is what you are conscious of."

An important consequence of rejecting this picture of consciousness is that our subjective experience of time may not match up with the objective timeline. We experience the *content* of conscious representations, not the vehicle doing the representing. It is tempting to assume that the temporal representation in consciousness is somehow transparent — that if our brains represent to us that A occurs before B, this must be because a vehicle repre-

still leaves many others whose preference for the longer pain is yet to be 'debunked'.

senting A was followed, in objective time, by a vehicle representing B. But of course there is no logical reason why this must be so. As Dennett points out, we can represent time using a medium other than time itself. We can say "B occurs after A", and it represents the ordering $\langle A, B \rangle$ even though the sentence mentions them in the order $\langle B, A \rangle$. Similarly, you can subjectively experience the represented order of events $\langle A, B \rangle$, even if the brain regions doing the representing actually process the events in the opposite order. And indeed several experiments have been conducted which demonstrate precisely this effect, at least on very short time scales.[5]

At this point, one may worry that our understanding of phenomenal consciousness has become *too* subjective. There is surely a difference between actually experiencing some temporally extended episode and merely *believing* (representing) this to be so. But can we maintain this distinction without falling back on physical duration as our objective standard? To bring out the problem, compare the following three 'experience machines':[6]

Machine A gives you pleasant experiences for 100 years.

Machine B (allegedly) gives you all the same experiences, feeling exactly the same from the inside, but packed into just a single physical day.

Machine C gives you one day of pleasure, and then simply implants in you the (presumably false) belief that it felt like it lasted for 100 years.

[5] See Dennett (1991, p.143) on the 'phi phenomenon', 'cutaneous rabbit', etc.

[6] Adapted from Ben Bradley and Troy Jollimore's blog discussion at *PEA Soup*: http://peasoup.typepad.com/peasoup/2008/02/objective-and-s.html#comment-101237434

The challenge is to make sense of how Machine B could be anything other than a fanciful redescription of Machine C — which we don't want to say *really* gives us 100 years of experienced pleasure. The solution, I think, is found by reflecting on the relative paucity of the C-implanted representation. It's one thing to write a story which says "100 years passed", and quite another to fill out the details for 100 years' worth of fictional events. So, if we think of Machine A as taking a long time to impart rich informational content (lots of pleasure), and Machine C as taking a very short time to impart a very thin representation (very little pleasure), the question whether Machine B imparts a lot or only a little pleasure comes down to the richness of the representations it implants. We can thus make the necessary distinctions without appeal to the dubious notion of a 'Cartesian Theater' in our minds where consciousness 'occurs'.

This suggests that a fairly radical reconceptualization of hedonic value may be in order. The spirit of hedonism seems more consonant with a concern for subjectively experienced duration than mere physical time. But this in turn is best analysed in terms of *representational richness*, or so I have proposed. This may have significant practical implications — let me note just two.

Firstly, evidence suggests that younger, inexperienced brains tend to lay down denser and richer memories, explaining why time seems to pass more quickly as we age (Stetson et al. 2007). So if my above claims are correct, utilitarians should be disproportionately (relative to their merely physical du-

ration) concerned to increase the hedonic quality of children's experiences — possibly counterbalancing our prior preference for a life that improves with age.

Secondly, we may be led to the conclusion that repetitive experiences may be discounted, if merely repeating the same old information over and over again does nothing to enrich the representation. (Suppose all you need is a ditto mark, or the cognitive equivalent of "times a million".) For example, if I am groggily experiencing a long, painful operation,[7] it's possible that after a while the moments will begin to blur together, to the point where it makes no subjective difference to me whether the operation lasts (say) two hours or four. Of course, the mere *retrospective* inability to tell the two scenarios apart does not by itself imply that there was no hedonic difference. But if the intervening momentary experiences are indeed qualitatively indiscernible in fact, and not broken up by any (perhaps later forgotten) distinguishing features, then this provides some reason to think that we should count the hedonic (dis)value as the same in either case. On the other hand, providing a 'signpost' that dispels temporal blur may cause an experience to count for a lot more.

I should emphasize that most of the time, even intrinsically identical pains are embedded in discernibly different experiences (with different background thoughts running through our heads, etc.), and so count as recognizably distinct. But we can at least imagine a case where extending the duration

[7] Compare the thought experiments in Parfit (1984, chapter 8).

doesn't introduce sufficient qualitative differences. After a while, many moments of hospitalized agony all blur together, and we may think the reason for this is precisely that there is truly nothing in the experiences to distinguish them. And so, on this account, they count for just one.

1.4 Taking Stock

I began this section by bringing out our implicit commitment to value holism as it applies within an individual life. There are many plausible cases in which it seems the value of a momentary experience cannot be assessed in isolation. Then, by distinguishing physical and 'subjective' (or consciously represented) time, I suggested, in addition, a *metaphysical* basis for something that closely resembles holistic practice: for even insofar as the *intrinsic* qualities of experience are concerned, these may yet be underdetermined by a physical moment in isolation, instead depending on physically later (or earlier) events to precisify their phenomenal character. Developing this idea further, I offered a speculative reconceptualization of hedonic evaluation, whereby the 'duration' dimension is replaced by an atemporal notion of 'representational richness'. This provided a reason to think that duplicate experiences may count for only one, assuming that duplication may be represented cheaply. In the following section, I will seek to establish some corresponding normative conclusions on the societal level, but without relying on such contentious metaphysical assumptions.

2 Whole Populations and Individual Lives

2.1 The Repugnant Conclusion

Derek Parfit (1984) famously observed that total utilitarianism implies the 'Repugnant Conclusion' that for any finite flourishing population A, we can imagine some vastly larger population Z of lives barely worth living, which ends up counting as "better". This is partly because total utilitarianism is an atomistic theory according to which the contributory value of a life is simply its welfare value for the person living it — a value that can be determined in isolation, simply by looking at that life in itself. Since worthwhile lives presumably have some positive contributory value, atomism implies that astronomically more lives yield astronomically greater value. Yet when we consider the imagined world Z as a whole, this conclusion no longer seems plausible. Holistic judgment enables us to recognize the Repugnant Conclusion *as* repugnant; we may accordingly expect value holism to provide a fruitful starting point in our search for a solution.

It is not easy to avoid the Repugnant Conclusion, as Parfit's 'Mere Addition Paradox' demonstrates.[8] 'Mere addition' is when we add additional lives — all above the baseline of lives worth living — to a world, without affecting the prior inhabitants in any way. Parfit claims that this process cannot make a world worse. This seems prima facie plausible: after all, where's the harm? How could it be bad to add intrinsically good lives, to no ill effect

[8] Parfit (1984, chapter 19); see also Arrhenius (1999).

for anyone else? This suggests the following principle:

> **Mere Addition:** If the only difference between worlds A and A+ is that the latter contains additional lives above the baseline, then A+ is no worse than A.

Next, note that it can only improve a world to reduce inequality in such a way as also increases total welfare, while holding all else equal. Call such a shift '**beneficial equality**'. Beneficial equality licenses the move from A+ to a world B where the worse-off group in A+ is benefited more than the well-off group is harmed by the shift. If B is better than A+, which in turn is no worse than A, it follows — by transitivity — that B (a world of greater total but lesser average utility) is likewise at least as good as A. We may iterate this process until we reach the repugnant world Z, with astronomic total utility but miniscule average utility.

These implications may lead us to examine the Mere Addition principle more closely. Indeed, to a value holist, the justification for the principle will seem immediately suspect, as it practically assumes atomism from the start. Recall that the justification appeals to the idea that the contributory disvalue of a life must be due to some badness in its intrinsic qualities: the life must be bad in itself, in a way that lives above the baseline are not. But this is just to assume atomism. To a holist there is nothing contradictory about the idea that adding an intrinsically good part may make the whole worse. (Laughter and merriment are good in themselves, but not at a funeral.

Or, as Hurka (1983) suggests, adding mediocre paintings to a collection of masterpieces may degrade the collection as a whole.) So while mere addition may sound harmless enough, the real test is to directly evaluate the two worlds in question. And here is it entirely open to us to judge that A+ is indeed a worse world than A. Why might we think this? Well, for one thing, the addition of worse (though not bad) lives alters the shape of the world as a whole, and not for the better. Whereas before we had a world full of flourishing, we now find mediocre lives in addition. That's not to say that the mediocre lives are bad in themselves, or considered in isolation. But given how the rest of the world is, their addition may be considered undesirable nonetheless.

None of this is to endorse anything so crude as the 'average utilitarian' principle that it's always bad to lower average welfare. A world sparsely populated with only a hundred people, however well-off they might be, would plausibly be improved by adding more good lives, even if they are not *as* good as those already there. For one thing, a world that's too sparsely populated can be expected to lack the full range of diversity that makes a (typical) larger population so desirable. But note that *diversity* is a distinctively holistic value: whether a particular life or experience adds to the diversity of the whole depends on what else there is.

2.2 Duplication and Holism

The Repugnant Conclusion brings to mind dystopian visions of a universe (Z) tiled with bland mediocrity — the same old "muzak and potatoes" (Parfit 2004), repeated over and over. This is certainly a dreary scenario. Muzak and potatoes were never all that great to begin with, but to simply add more of the same is arguably not to add any more value to the world whatsoever. It would be different if we were to instead imagine a world Z* of very diffuse and diverse excellences, where each life contained only modest value when considered in isolation, but nonetheless offered a *distinctive* contribution to the world as a whole. This no longer seems so repugnant at all; it may even be an open question whether, intuitively, it would be better to condense those diffuse glimpses of excellence into a smaller number of more consistently flourishing lives. The holist may go either way on this question, so I will not attempt to settle it here. What I want to highlight instead is the intuitive significance of diversity. What seems *most* repugnant about the Repugnant Conclusion (as envisioned above) is not just that value is so diffuse, but that there doesn't really seem to be much value in the Z-world at all. We're inclined to think that contributory value just can't be duplicated in this way. Rather, to make the world as a whole more valuable, we must add lives of distinctive value. So claims the 'diversity principle' that I now wish to consider.

DP: Multiple evaluands count for less the less distinct they are.

A fully fleshed-out version of this principle would need to specify what qualitative dimensions count as introducing a normatively relevant 'difference'. For example, diversity of lifestyles and fundamental goals seems more normatively significant than diversity of spatial location or hair colour. As I don't have space to develop the details here, I will simply assume that the reader shares with me a rough sense of what dimensions seem relevant. Even this rough of a grasp of the principle suffices to allow some initial observations and assessments.

However it may be fleshed out, DP has a number of significant implications. We've already seen how it can explain the problem with (the most repugnant version of) the Repugnant Conclusion.[9] For another example: if DP were false, then it would be of momentous importance whether ours was a world of Nietzschean 'eternal recurrence'. So those of us skeptical about whether this would really matter so much must be relying on some DP-like holistic principle for discounting all those duplicate epochs. For a more practical example: suppose it turns out that the experiences of most hens in factory farms are qualitatively identical (or nearly so).[10] DP would then imply that the total disvalue here is less than we might at first expect, since the duplicates are subject to discounting. This would be a very surprising, and somewhat disconcerting, result.

[9] Though it's worth noting that it suggests a new objection to the mere addition paradox: the 'beneficial equality' step may actually be for the worse if it reduces diversity.

[10] I owe this example to Michael Vassar.

This may be turned into an objection: it seems absurd to think that it might make the world no worse were a billion people tortured instead of just one, for example.[11] But such intuitions gain their force from 'ordinary' cases, where different people have different memories, etc., and which thus involve discernibly different experiential contexts. (This background assumption of human cognitive diversity may be at least part of the explanation why this objection seems so much more gripping when applied to humans than hens. The logic of atomism should apply equally to either case, by contrast.) Special care is required to conceive of a situation where DP would actually apply — e.g. a dystopian 'farm' of duplicate brains-in-vats, programmed to have exactly the same series of experiences throughout their existence — but then the intuition is less clear. To ensure that all variables are controlled, imagine a 'digital person' or conscious Artificial Intelligence, whose 'life' is constituted by the running of a computer program. Would it matter how many duplicate copies of the program were run? It isn't obvious. Suppose that, as a digital person, you are about to undergo copying (digital 'fission'), but that not all of your future continuants will be identical. If there were to be 95 copies of one programmed future, and just one copy each of five other, qualitatively distinct futures, would you rather improve the first program or the other five? The latter preference would indicate a strong commitment to DP.

[11] Or, as Bostrom (2006, p.188) puts it: "It would... be odd to suppose that whether one's own brain produces [morally relevant] phenomenal experience strongly depends on the happenings in other brains that may exist in faraway galaxies..."

Clear intuitions are difficult to come by in these cases, so we may prefer to decide the issue on more general theoretical grounds. The first section of this paper noted that we care about the *shape* of our life as a whole. A stronger claim is that this is *all* that matters. Mere multiplication — say where every event extends for twice as long — would then be dismissed as lacking in normative significance, so long as the broad contours of the life remained much the same. The value of an extra year of life depends on what would be achieved with it, and whether it would contribute anything new or significant to the overall structure of the life. If it is all much of a sameness with what has gone before, it may not have any significant impact on the quality of the life taken as a whole. And the same may be true of the relation between individual lives and the value of the world as a whole.

This general holistic picture may cast some doubt on the importance of absolute quantities of mere duplicates. But it doesn't immediately follow that all that matters is the number of distinct evaluands. We may also care about their proportions, or the *relative* quantities of each duplicated kind. For example, in the case of digital fission, perhaps I should care more about the future that occurs in 95% of instances, even if I shouldn't care whether the absolute number of instances is 100 or 1000. This is another issue on which the holist could go either way; I won't attempt to settle the matter here.

2.3 The Alleged Asymmetry

Some claim that there is an important asymmetry in our assessments of merely possible pleasures and pains (or goods vs. bads more generally). They claim that we have little or no reason to bring more good lives into existence, or to regret their absence, whereas we do have (significant) reason to prevent lives of suffering and to feel relief at their absence.[12] For example, we feel that a 'package deal' containing several additional good lives and one bad life would thereby be a *bad* deal. As Christopher Belshaw (2007) writes, "We can't justify starting this bad life by appeal to the good in other, separate, lives." But I propose that we can explain away such intuitions by appeal to value holism.

According to value holism, the value of the above 'package deal' is not independent of what else exists. It might be positive in some circumstances and negative in others, depending on how it impacts the overall 'shape' of the world. As it happens, we have a high normal baseline: we assume that most lives in our society are pretty good. So, creating a good life is nothing exceptionally good, whereas creating a bad life *is* exceptionally bad. Given the more fundamental principle that exceptionally good or bad actions have greater moral significance (in virtue of their impact on the general shape or form of society), we find that a contingent asymmetry in social circumstances leads to the above moral asymmetry.

Note that things could have been different. If we imagine a dystopian

[12] See, e.g., Benatar (2006, p.30); Parfit (1984, p.391).

world where the normal 'baseline' is much lower — i.e., where most lives are rather awful — then it seems to me that the moral asymmetry would be likewise inverted. Given the opportunity to bring about an exceptionally good life, people ought to do so. To prevent another typically bad one would be permissible — good, even — but not required. So, dystopians ought to embrace the package deal of good and bad lives.

On the view I've outlined, there is no fundamental moral asymmetry in the relative weighting of good and bad additional lives. Any asymmetry here instead arises from the application of value holism to our particular circumstances. We may be right to think that we shouldn't want six additional good lives at the cost of an additional bad one, in our current circumstances. But this does not imply that more good lives *couldn't* — in other circumstances — outweigh bad ones so as to make a world better on net.

Some have proposed a more plausibly fundamental asymmetry: although additional good lives may make a world non-instrumentally worse (if they reduce the average welfare, say), additional bad lives cannot make a world non-instrumentally better (even if they are less bad than average). I think it's open to a holist to reject this assumption. But even if we don't go quite that far, Huemer (2008) proves that assigning any non-zero weight to average utility commits us to:

> **The Sadistic Conclusion:** In some circumstances, it would be better with respect to utility to add some unhappy people to the world (people with negative utility), rather than creating a larger

number of happy people (people with positive utility).

This does seem counterintuitive, at least at first glance. But further reflection reveals that it is not much of a move from the claim that adding mediocre lives can make a world worse. For then we may expect that adding a great many mediocre lives could make a world much worse (transforming it from a predominantly flourishing world to a predominantly mediocre one). If this is a harm at all, then it isn't surprising that it could outweigh the modest harm of adding a single moderately bad life. We are tempted to draw a bright line between lives that are worth living and those that aren't, but the absolute difference in utility might be as small as you care to imagine (for arbitrarily small ϵ, compare the welfare values $+\epsilon/2$ and $-\epsilon/2$). So we should not place as much weight on this difference as is relied upon in the above objection.[13]

2.4 Wrapping Up

While my first section explored the application of value holism within a life, this second section has explored how a holistic view might be applied to some of the central problems of population ethics. Value holism naturally suggests two routes to avoiding the Repugnant Conclusion: the first by rejecting Mere Addition, and the second by rejecting any duplication that may sneak in during the 'beneficial equalizing' step. (The latter leaves open a diversified version of the repugnant conclusion, but then it no longer seems

[13] Arrhenius (1999) similarly relies on the 'Non-Sadism Condition' that any number of happy lives cannot be worse than any number of unhappy lives. Despite the forceful name, I think that value holists should feel quite comfortable in rejecting this principle.

so repugnant.) The preceding sub-section addressed some objections that are based upon alleged asymmetries between harms and benefits. I now want to wrap up by drawing out what I consider the two most pressing challenges that emerge from all this.

First, there is the question whether atomists can make a principled stand in defence of Mere Addition. Huemer suggests the following "almost irresistable" principle:[14]

> **Modal Pareto Principle:** For any possible worlds x and y, if, from the standpoint of self-interest, x would rationally be preferred to y by every being who would exist in either x or y, then x is better than y with respect to utility.

As before, this principle is clearly atomistic: it considers each life in isolation ("from the standpoint of self-interest"), and leaves no room for holistic 'big picture' considerations. But this can be defended by appeal to an independently appealing conception of ethics as fundamentally person-centered. It is tempting to think that in order for something to be bad, it must be bad *for someone* (cf. Parfit 1984, p.395). We may further think that our reason to avoid a bad outcome is fundamentally *second-personal*, in the sense that it stems from the normative authority of the individual(s) who would be harmed (Darwall 2006). On this view, moral agents aim simply to act in

[14] Huemer (2008, p.903). This supports what Huemer calls "Benign Addition" — just like Mere Addition except that the original inhabitants of world A are very slightly better off in A+. (The new additions to A+ are, as before, (barely) happy to be alive, and so also prefer A+ to A.)

ways that are justifiable to others, without any ambition to promote impersonal value or make things better from 'the point of view of the universe'. (The universe has no point of view, and if it did we would have no reason to care about it, or so the thought goes.) Call this view **personalism**.

Now, personalists might accept holism within an individual life, since they accept that — morally speaking — the whole person is 'prior' to their temporal parts. Value inheres in people first, and their momentary stages only derivatively. But the personalist will reject my extension of holism to the interpersonal level. They deny that there is any supra-personal collective entity (even 'the world' as a whole) that can serve as an independent value-bearer or source of normatively authoritative reasons. Whereas the holist talks seriously about making *the world* a better place, personalists will insist that this is merely shorthand for making *the individual people* in the world better off. (Compare Frankena's maxim that "Morality is made for man, not man for morality.")[15]

Such a view has significant traction in the western cultural tradition. But even total utilitarians must ultimately reject it, for reasons that emerge from Parfit's work on the 'non-identity problem' (1984, chapter 16). Suppose that we have an opportunity to bring about world Z, but choose to stick with world A instead. Looking back, the total utilitarian will judge that we chose wrongly. But who or what has been 'wronged' or harmed by our choice? The

[15] Frankena (1973, p.116). Note that Frankena wasn't talking about this particular issue; but it's a suggestive turn of phrase all the same.

only available answer, I want to suggest, is 'the world as a whole'. After all, they can't very well insist that we should have brought about world Z *for the sake of those merely possible people* who will now never get to exist. Only existent beings can make claims on us, so there are no second-personal reasons to bring about world Z. In particular, there is nobody who could complain of being harmed, mistreated, or disrespected by our failure to bring Z about. This is not to deny that we may have reason to bring people into existence, or to lament our past failure to do so. But it would be nonsensical to say that the non-existent people are themselves the source of the reason. There are no such people.[16] So if we want to say that there are reasons in this case, we are forced to go beyond personalism. The reasons that we have in these cases will be *impersonal* reasons — which opens the door to holism.

Despite being forced to reject personalism, the total utilitarian may reasonably insist that there's an important sense in which (on their view) the intrinsic value of the additional possible lives is what 'gives us reason' — or explains our duty — to bring about their existence.[17] After all, utilitarianism is a form of **welfarism**: the axiological view that what's good is just

[16] One may object that the people would exist if we acted rightly. So in that case, at least, the total utilitarian could point to these existing people as the reason why the act was right. But this is insufficient, for we need an explanation that will also carry over to the case in which we fail to act rightly — in which case there do not exist any such additional people to ground the claim that it was wrong of us not to bring them into existence. Hence the moral claim must stem from some other source.

[17] That's slightly sloppy wording, however. We arguably can't refer to the imagined 'additional people' individually, if they don't exist and so have no particular identities. To speak more carefully, what gives us reason here is *the fact that the imagined outcome would contain x many additional happy people*. This makes it clearer that we're talking about (a property of) the world, and not about any particular individuals in it.

the welfare of sentient beings. But this should not lead us into metaphysical confusion. In particular, even if we have a duty to procreate so as to bring about more good lives — lives that will be good for the people living them — this is not to say that we have a duty *to* those non-existent people. (Again, there are no such people to whom we might be duty-bound, in the case where we fail to procreate.) Rather, such loose talk merely serves to specify the *content* of the obligation, not its *source*. So the utilitarian may certainly hold that an additional good life would be intrinsically good, and indeed that the right-making feature of the procreative choice is precisely the welfare value of the resulting life: the value it would have for the person living it. But this first-order moral claim about *what* we have reason to choose (and why) must be distinguished from metaphysical claims about the source or nature of the reason. In this case, the reason must be impersonal in nature, even if welfarist in content. And since welfarism is just another first-order axiological theory rivaling value holism, it cannot provide *independent* support for the Modal Pareto Principle, the way that personalism might have.

To wrap up the point: It may be thought impersonally desirable that the world should contain more happy people — in which case we may reasonably lament that this isn't so. We may regret that the world is worse than it could have been. But we cannot sensibly think that non-existent individuals are worse-off than they might have been. Though our world is worse, it is not worse *for them* (or, perhaps, anyone). This means that defenders of the repugnant conclusion cannot appeal to personalism, and so their atomistic

Modal Pareto Principle lacks independent support after all.

A second — more methodological — challenge may be suggested by the various unresolved questions I've raised about how to best flesh out the holistic position. Rejecting atomism opens up a number of new variables, and the sheer range of options here may raise concerns about 'curve-fitting'. Tailoring the view to accommodate our intuitions in particular cases may just seem less 'principled', somehow, than the atomist's bullet-biting resolve. I think that can't be right quite as stated: the method of reflective equilibrium licenses moving back and forth between our judgments of particular cases and general principles, as we seek to bring them all into coherence (Daniels 2008). But the atomist might at least note that their theory has an advantage in terms of simplicity. The question is whether it is *too* simple in its neglect of the 'big picture' relations between lives.

References

Arrhenius, Gustaf. 1999. "An Impossibility Theorem In Population Axiology With Weak Ordering Assumptions." In Rysiek Sliwinski (ed.), *Philosophical crumbs. Essays dedicated to Ann-Mari Henschen-Dahlquist on the occasion of her seventy-fifth birthday*, volume 49 of *Uppsala Philosophical Studies*, 11–21. Uppsala: Department of Philosophy, Uppsala University.

Belshaw, Christopher. 2007. "Review of David Benatar, 'Better Never to Have Been'." URL: http://ndpr.nd.edu/review.cfm?id=9983.

Benatar, David. 2006. *Better Never to Have Been: The Harm of Coming into Existence*. Oxford: Oxford University Press.

Bostrom, Nick. 2006. "Quantity of experience: brain-duplication and degrees of consciousness." *Mind and Machines* 16:185–200.

Dancy, Jonathan. 2004. *Ethics Without Principles*. Oxford: Oxford University Press.

Daniels, Norman. 2008. "Reflective Equilibrium." In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Fall 2008 edition. URL: http://plato.stanford.edu/archives/fall2008/entries/reflective-equilibrium/.

Darwall, Stephen. 2006. *The Second-Person Standpoint*. Cambridge, MA: Harvard University Press.

Dennett, Daniel. 1991. *Consciousness Explained*. Boston: Little, Brown and Co.

Frankena, William. 1973. *Ethics*. Englewood Cliffs, NJ: Prentice Hall, 2nd edition.

Huemer, Michael. 2008. "In Defence of Repugnance." *Mind* 117:899–933.

Hurka, Thomas. 1983. "Value and Population Size." *Ethics* 93:496–507.

Kagan, Shelly. 1988. "The Additive Fallacy." *Ethics* 98:5–31.

Kahneman, Daniel, Fredrickson, Barbara L., Schreiber, Charles A., and Redelmeier, Donald A. 2003. "When More Pain is Preferred to Less: Adding a Better End." *Psychological Science* 4:401–5.

McNaughton, David. 1988. *Moral Vision*. New York: Basil Blackwell.

Moore, G.E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.

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

—. 2004. "Overpopulation and the Quality of Life." In Torbjörn Tännsjö and Jesper Ryberg (eds.), *The Repugnant Conclusion*, 7–22. Springer Netherlands.

Stetson, Chess, Fiesta, Matthew P., and Eagleman, David M. 2007. "Does Time Really Slow Down during a Frightening Event?" *PLoS ONE* 2:e1295. doi:10.1371/journal.pone.0001295.

Velleman, J. David. 1991. "Well-being and Time." *Pacific Philosophical Quarterly* 72:48–77.