In Defence of Repugnance

Michael Huemer

I defend the 'Repugnant' Conclusion that for any possible population of happy people, a population containing a sufficient number of people with lives barely worth living would be better. Four lines of argument converge on this conclusion, and the conclusion has a simple, natural theoretical explanation. The opposition to the Repugnant Conclusion rests on a bare appeal to intuition. This intuition is open to charges of being influenced by multiple distorting factors. Several theories of population ethics have been devised to avoid the Repugnant Conclusion, but each generates even more counterintuitive consequences. The intuition opposing the Repugnant Conclusion is thus among the best candidates for an intuition that should be revised.

1. The Repugnant Conclusion and the Benign Addition Proof

1.1 The 'Repugnant' Conclusion

The following ethical theorem was proved a number of years ago by Derek Parfit, who, however, recoiled from his own progeny, cruelly naming it 'the Repugnant Conclusion':

(RC) For any world full of happy people, a world full of people whose lives were just barely worth living would be better, provided that the latter world contained enough people.[1]

Upon proving this proposition, Parfit introduced the name 'the Mere Addition Paradox' to denote the intellectual problem occasioned by the existence of the proof and its conflict with our intuitions.

Since Parfit's unwelcome discovery, several moral philosophers have taken (RC) under their wing, and several discoveries have been made that further strengthen (RC).[2] Despite all this, most philosophers continue to despise (RC), citing (what else?) their intuitive sense of repugnance.[3] At times, it seems that (RC) will never earn acceptance, no

[1] Parfit 1984, Ch. 17. I have slightly altered the principle from Parfit's formulation.

[2] See Anglin 1977; Sikora 1978; 1981; Ng 1990, pp. 191–3; Attfield 1991, pp. 127–30; Ryberg 1996; Fotion 1997; Tännsjö 2002; and Broome 2004, pp. 210–14.

[3] See Parfit 1984; Temkin 1987; Locke 1987; Boonin-Vail 1996; Arrhenius 2000; Rachels 2001; Hurka 2003; Cowen 2004; and Blackorby, Bossert, and Donaldson 2004.

Mind, Vol. 117 . 468 . October 2008
doi:10.1093/mind/fzn079
© Huemer 2008

matter how strong the arguments in its favour. Because of this ongoing injustice, I have undertaken in this piece to offer a comprehensive defence of (RC) and response to its critics, in the hopes that (RC) will at last come to be accepted for what it is: one of the few genuine, nontrivial theorems of ethics discovered thus far.

1.2 The assumptions of population axiology

The Repugnant Conclusion, like other theses in population ethics, asks us to compare possible worlds in terms of their overall value. I assume that such comparisons are possible, and that we may rely on our ethical intuitions in making such comparisons—whether directly or through reasoning based on abstract principles.

These assumptions are nontrivial. In addition to those who doubt the validity of intuition as a source of ethical knowledge (Mackie 1977; Sinnott-Armstrong 2006), some philosophers reject the notion of the overall goodness of an event or state of affairs, or of improving or worsening the world as a whole. Geach (1956) holds that a thing can be *a good F*, for some particular sortal term 'F', but not good *simply*. Nor is there such a thing as *a good event* in his view, since 'event' is too broad of a category for there to be standards for good events in the way that there are standards, for example, for good pens. Thomson holds that all goodness is goodness *in a way*—where being good in a way includes being *a good F*, being *good for a given person*, and being *good for a given purpose*, among other things. Like Geach, she denies that a thing can be good *simply*, that there are unqualifiedly good and bad events, or that there is such a thing as improving the world as a whole (for this reason, she sees consequentialism as incoherent). She characterizes the contrary assumption as reflecting a confusion about the use of language (Thomson 2001, pp. 17–19).

I do not attempt to disprove their views here; however, I shall assume that these philosophers are mistaken. When I ask myself whether it would be better for there to be one billion barely-worth-living lives or one million wonderful lives, it seems to me that I clearly understand the question, and thus that it is not incoherent or meaningless. I am not asking which would be better for some particular person or group, nor which would be better for some particular purpose, nor, in general, which would be better in some particular way. I am asking which would be better in the generic, agent-neutral sense—'from the point of view of

 © Huemer 2008

the universe', in Sidgwick's phrase.[4] In asking this, I am not falling prey to a simple oversight about the proper use of language: I have explicitly considered whether the words 'good' and 'better' function only as Geach and Thomson describe, and it seems to me that they do not.

Likewise, I do not attempt to refute sceptics about ethical intuition here.[5] I assume that we have some *prima facie* justification for believing what our intuitions tell us regarding generic, agent-neutral value. These intuitions, however, are fallible, and a process of reasoning may often be needed to correct wayward intuitions.

These assumptions are commonly taken for granted in population axiology, whether one accepts the Repugnant Conclusion or not. I believe, for example, that a world containing a billion people with slightly valuable lives is better than a world containing a million people each with lives 100 times better than in the first world. I base this conclusion on reasoning from certain intuitive axioms described below. My arguments are addressed to those who think, on the basis of intuition, that the first world would be overall *worse* than the second.

1.3 The Benign Addition Argument

I begin with a variant of Parfit's proof.[6] Assume that there are levels of well-being, which may be represented by numbers. Positive numbers represent desirable levels of well-being, levels of well-being that render life worth living. Negative numbers represent states in which life is worth *not* living. 'o' represents a neutral state, in which it is a matter of indifference whether one continues in that state or ceases to exist. A possible world's *total utility* is the sum of all its inhabitants' levels of well-being. A world's *average utility* is its total utility divided by the population size. We start with the following ethical axioms:

[4] Sidgwick 1907, p. 382. See Moore (1903), whom Thomson mistakenly accuses of linguistic confusion, for more on the concept of generic, agent-neutral goodness.

[5] I defend the epistemic value of ethical intuition elsewhere (2005, Ch. 5; 2008).

[6] The argument following in the text is inspired by Parfit (1984, Ch. 19; see also his 1986, pp. 14–17), but I have taken considerable liberties. I have simplified the argument so that only two world-comparisons are required, and I have substituted for Parfit's (1984, p. 420) 'mere addition' a case of 'benign addition' (my term). Mere addition occurs when a group of people with positive welfare is added to the world without changing the welfare of any of the original people; benign addition occurs when a group of people with positive welfare is added while *increasing* the welfare of all of the original people (as in Parfit 1986, pp. 15–16; Tännsjö 2002, pp. 358–9). The purpose of the latter change is to avoid objections stemming from the Person-Affecting Principle (see Sect. 3 below) and from Parfit's (1984, pp. 430–2) view that A^+ might fail to be worse than A *without* being either as good as or better than A.

 © Huemer 2008

The Benign Addition Principle: If worlds *x* and *y* are so related that *x* would be the result of increasing the well-being of everyone in *y* by some amount and adding some new people with worthwhile lives, then *x* is better than *y* with respect to utility.[7]

Non-anti-egalitarianism: If *x* and *y* have the same population, but *x* has a higher average utility, a higher total utility, and a more equal distribution of utility than *y*, then *x* is better than *y* with respect to utility.[8]

Transitivity: If *x* is better than *y* with respect to utility and *y* is better than *z* with respect to utility, then *x* is better than *z* with respect to utility.

The qualifier 'with respect to utility' indicates that we are only considering the value that a world has in virtue of the levels of well-being enjoyed by its inhabitants; we are bracketing questions about such values as justice, freedom, knowledge, virtue, and so on. We are to assume, then, that all the worlds discussed are comparable in all those other dimensions. There remains an interesting question as to how we should evaluate worlds on the basis solely of their distributions of utility. Hereafter, I shall take these qualifications as read.

To see how these principles necessitate the Repugnant Conclusion, consider three possible worlds (figure 1):

World A: One million very happy people (welfare level 100).

World A$^+$: The same one million people, slightly happier (welfare level 101), plus 99 million new people with lives barely worth living (welfare level 1).

World Z: The same 100 million people as in A$^+$, but all with lives slightly better than the worse-off group in A$^+$ (welfare level 3).

[7] The notion of 'adding' people to a world need not be taken to denote a temporal process; rather, when we have imagined a possible world, we 'add' people to it by imagining another world just like the first but with additional people. A similar interpretation may be applied to the notion of 'increasing' people's utility in a world.

[8] The name 'Non-anti-egalitarianism' derives from Ng (1989, p. 238), who uses the principle in an argument much like the Benign Addition Argument (Ng 1989, p. 240).

 © Huemer 2008

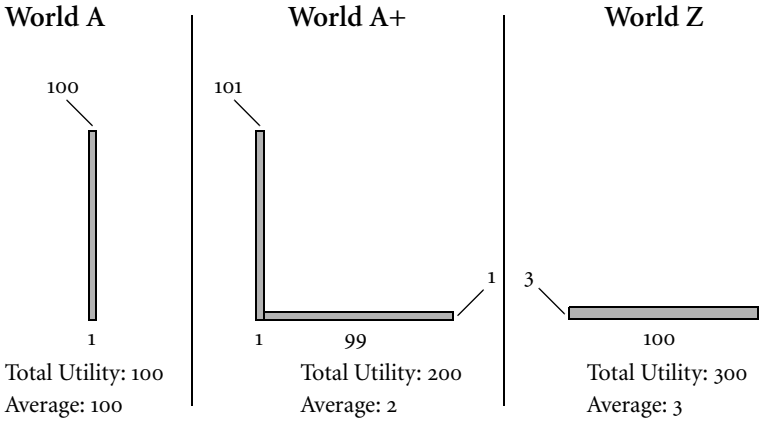

Fig. 1

A^+ is better than A by the Benign Addition Principle, since A^+ could be produced by adding one unit to the utility of everyone in A and adding some more lives that are (slightly) worthwhile. Z is better than A^+ by Non-anti-egalitarianism, since Z could be produced by equalising the welfare levels of everyone in A^+ and then adding one unit to everyone's utility. Therefore, by Transitivity, Z is better than A. Analogous arguments can be constructed in which world Z has arbitrarily small advantages in total utility; as long as Z has even slightly greater total utility than A, we can construct an appropriate version of A^+ that can be used to show that Z is better than A. This suggests that we should embrace not only (RC), but the logically stronger

> *Total Utility Principle:* For any possible worlds *x* and *y*, *x* is better than *y* with respect to utility if and only if the total utility of *x* is greater than the total utility of *y*.

Since the Repugnant Conclusion is counterintuitive, suspicion naturally falls on the three premisses. Must we accept these premisses?

1.4 The premisses of the Benign Addition Argument

The Benign Addition Principle is the most popular target for opponents of (RC). The Benign Addition Principle, however, is supported by the almost irresistible

> *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.

 © Huemer 2008

Modal Pareto expresses a very weak general condition of benevolence—roughly speaking, we should favour outcomes that are good for everyone. This supports the Benign Addition Principle, since in cases of benign addition—where people with good lives are added to a world while benefitting all of the original people—both the original people and the new people would, from the standpoint of rational self-interest, be glad of the change. All the inhabitants of world A would prefer A^+ over A, since A^+ gives them an extra point of utility. Even the worse-off group in A^+ would rationally prefer A^+ to A from the standpoint of self-interest, since they would rather live at a utility level of 1 than not live at all. If asked whether they would like the world to be converted to A, they would decline. If asked whether they were glad the world was like A^+ rather than A, they would answer in the affirmative.

Some would question whether we should count the utility of the worse-off inhabitants of A^+ as an advantage that A^+ has over A—more generally, some doubt that the utility of people who exist in only one of the outcomes should be taken into account (Narveson 1967; 1973). All agree that we should *at least* count the welfare of the people who exist in both outcomes, while only some think we should also count the welfare of the possible individuals who exist only in A^+. But this dispute does not matter here: *either* view delivers the verdict that A^+ is the better of the two worlds, since it is better for the people who are common to both worlds, and, if we are to count the possible individuals existing only in A^+, then it is also better for those possible individuals.

What of the Non-anti-egalitarian principle? This principle holds that equalising everyone's utility while also slightly increasing the total and average utility of the world makes the world better in terms of utility. Presumably, increasing average and total utility makes the world pro tanto better. Therefore, Non-anti-egalitarianism could be false only if equalisation of utility could make the world *worse*. How could this be? The most obvious way would be if equality in the distribution of utility were intrinsically bad, as anti-egalitarianism maintains. But almost no one believes anti-egalitarianism. Most believe equality is intrinsically good, while nearly everyone else believes it is intrinsically neutral.[9]

Finally, Transitivity is among the most widely accepted and intuitive principles in all of ethics, not to say all of philosophy.[10] It can be further

[9] See my 2003 for an argument that equality is intrinsically neutral. For theories that conflict with Non-anti-egalitarianism, see Sider 1991; Parfit 1986; Rachels 1998a; 2001.

[10] Temkin (1987; 1996), Rachels (1998a; 2001), Andreou (2006), and Quinn (1990) have denied Transitivity. However, even Temkin (1996, pp. 175, 177) and Rachels (1998a, p. 71) acknowledge its powerful and widespread intuitive appeal.

 © Huemer 2008

supported by at least two arguments. First is the Money Pump Argument:[11] Suppose that A is better than B, which is better than C, which is better than A. It seems that a rational person might then prefer A to B, B to C, and C to A. Suppose you are such a person. You presently have A. I offer to let you pay a small amount of money to be allowed to trade A for C. Since you prefer C, you accept. I then let you pay a small amount of money to trade C for B. Again you accept. I then let you pay a small amount of money to trade B for A—the very same A that you started with. You accept. And so on. I have found a way to pump money out of you indefinitely, just by relying on your allegedly rational preferences. Surely this shows that those preferences are not in fact rational.[12]

The second argument for Transitivity relies on the following two premisses:

> *Dominance*: For any states of affairs x_1, y_1, x_2, and y_2, if *(i)* x_1 is better than y_1, *(ii)* x_2 is better than y_2, and *(iii)* there are no evaluatively significant relationships among any of these states, then the combination of x_1 and x_2 is better than the combination of y_1 and y_2.
>
> *Asymmetry*: If x is better than y, then y is not better than x.

To illustrate the Dominance principle, suppose that I am deciding whether to buy a Honda or a Ford. I am also deciding whether to live in California or Texas. Assume there are no evaluatively significant relationships between these choices: where I live has nothing to do with what kind of car is best, and vice versa. Finally, suppose that the Honda is better than the Ford, and living in California is better than living in Texas. Then it seems that buying the Honda and living in California would be better than buying the Ford and living in Texas.

Now suppose that Transitivity is false, and that there is a series of unrelated values, A, B, C, and D, where A is better than B, which is better than C, which is better than D, which is better than A. I shall denote the combination of A and C, 'A+C' (and similarly for other combinations). If A and C are two states of affairs, A+C is the state of affairs that obtains when A and C both obtain. Now consider which is better: A+C, or B+D? By Dominance, A+C is better than B+D, because A is better than B and C is better than D. But at the same time, B+D is better than

[11] This argument is taken from Davidson, McKinsey, and Suppes (1955), who credit Norman Dalkey.

[12] Nozick (1993, p. 140 n.), Rachels (1998a, pp. 82–3), and Andreou (2006) question the implicit premiss that it is always rational to choose the acknowledged better of two options. None, however, appear to offer grounds for doubting this premiss that are independent of the assumption of Intransitivity.

 © Huemer 2008

C+A, because B is better than C, and D is better than A. The following diagram depicts these relationships:

A + C
∨ ∨
B + D
∨ ∨
C + A

In short, merely by rearranging the elements of one combination, we can make either combination come out looking better than the other, thus violating Asymmetry.

In sum, Transitivity, Non-anti-egalitarianism, and the Benign Addition Principle each appear obviously correct. Collectively, they entail (RC). Yet most philosophers still deny (RC), calling the argument for (RC) a 'paradox'. The literature on the Repugnant Conclusion has often taken for granted that the philosophical problem is to explain why the argument for (RC) is wrong. Can this attitude be rational?

1.5 Intuitions and paradoxes

Sometimes initially compelling arguments are wrong. And sometimes we know an argument to be wrong, despite our inability to identify a specific flaw in it. The term 'paradox' is often used in such cases. For instance, one may know that motion exists despite being unable to resolve Zeno's Paradox. Or one may know that we have knowledge of the external world, despite lacking a satisfying response to the sceptic's arguments. We should grant, therefore, that it is *sometimes* rational to reject the conclusion of a seemingly compelling argument, despite inability to identify what is wrong with the argument. And it is sometimes rational to reject a premiss of an otherwise compelling argument, simply because the conclusion is very implausible.

But it is not *always* rational to reject a seemingly compelling argument for a counterintuitive conclusion. Intuitions can be wrong, and we can be justified in revising our intuitive judgements in the light of reasons. Which intuitions should be revised? While many cases will remain controversial, the following criteria may help in deciding whether a particular intuition should be revised:

(1) Does the intuition conflict with firm and widely-shared other intuitions?

(2) Is there a plausible error theory, explaining what sort of biases or mistakes might generate the intuition?

 © Huemer 2008

(3) Is there more than one independent line of argument against the intuition?

(4) Is it a 'bare' intuition, lacking significant support from other intuitions, and lacking a satisfactory theoretical explanation?

(5) Is there a natural theoretical explanation for a contrary view?

A 'yes' answer to any of these questions counts in favour of revising an intuitive belief. None of these criteria are decisive, but each can contribute to the case for revision.

In the present case, we must assess the intuition that situations like Z are worse than situations like A. Call this 'the Unrepugnant Intuition'. We have already seen that the Unrepugnant Intuition conflicts with a trio of firm and widely-shared intuitions, namely, the Modal Pareto Principle, Non-anti-egalitarianism, and Transitivity. In the next section, I offer explanations for why intuition leads us astray about (RC). In section 3, I review several alternative theories in the literature that aim to explain why (RC) might be false, along with some of the reasons why these alternatives are unsatisfactory. In sections 4, 5, and 6, I present three further arguments for (RC), the last of which provides a simple, straightforward explanation for why (RC) would be true. We shall find in the end that the Unrepugnant Intuition satisfies all of the above criteria for meriting revision.

2. Distrusting unrepugnant intuitions

A number of factors may distort our judgements about (RC), leading to the unreliable sense that Z is worse than A. These factors include:

2.1 The egoistic bias

When comparing worlds A and Z, we may find ourselves imagining what it would be like to live in each, and asking ourselves which we would prefer.[13] Even if we consciously realise that this is not the relevant question, our intuitive evaluation may still be influenced by our preferences. We would prefer a world in which we are ecstatically happy to one in which we are barely content. Thus, we tend to evaluate A more positively than Z. But the fact that we would prefer to occupy world A hardly shows that A is better. Our preference on this score takes account only of the level of well-being we would enjoy if we occupied each world. It takes no account of the vast *numbers* of other people who, in

[13] Locke (1987, p. 144) appeals to this sort of consideration explicitly. Rachels (2004, p. 180) mentions this way of judging as a possible source of bias.

 © Huemer 2008

world Z, are given the chance to live (slightly) worthwhile lives. Some philosophers would deny that the numbers matter. This is exactly what is in dispute. My present point is not to directly resolve that dispute; my present point is that, as long as it is in dispute whether the numbers matter, we cannot hope to resolve the dispute by appealing to a method of judging that, by its nature, is designed to ignore the numbers. To the extent that we have reason to suspect that our intuitions reflect such a method, we should distrust those intuitions.[14]

2.2 The large numbers bias

As Broome (2004, pp. 57–9) observes, we should be wary of intuitions whose reliability turns on our appreciating large numbers. This is because, beyond a certain magnitude, all large quantities strike our imagination much the same. A popular joke illustrates this:

> An astronomer giving a public lecture mentions that the sun will burn out in five billion years. An audience member becomes extremely agitated at the news. The lecturer tries to reassure him: 'No need to worry, it will not happen for another five billion years.' The audience member breathes a sigh of relief, explaining, 'Oh, five billion years. I thought you said five *million* years!'

When we try to imagine a billion years, our mental state is scarcely different, if at all, from what we have when we try to imagine a million years. If promised a billion years of some pleasure, most of us would react with little, if any, more enthusiasm than we would upon being promised a million years of the same pleasure. Intellectually, we know that one is a thousand times more pleasure than the other, but our emotions and felt desires will not reflect this.

Worlds A and Z both contain sufficiently large numbers of people that we cannot clearly imagine these numbers, let alone sympathize fully with all of the imagined people. This psychological limitation produces a bias in favour of world A when we intuitively evaluate the two worlds. Our evaluations are influenced by our emotional states when we imagine different worlds, where these emotional state are strongly influenced by the average welfare of each world, but, above a relatively low population level, only weakly influenced by the population size of each world due to our inability clearly to imagine large populations.

[14] Tännsjö (2002) proposes that we consult, not our preference between living in A and living in Z, but our preference between (i) a small probability of living in A and a large probability of instead ceasing to exist and (ii) a certainty of living in Z. Narveson (1973, p. 85) considers such an approach but finds it unhelpful for deciding between different population policies.

 © Huemer 2008

Since—according to the advocates of (RC)—the key to world Z's great value lies particularly in its enormous population, world Z gets the worse of our intuitive evaluation process.

The dispute between the proponents and the opponents of (RC) centres on whether or not the sheer size of world Z's population is a great advantage. We have independent grounds for expecting our intuitions to more or less ignore that factor when we imagine worlds A and Z, whether or not that factor is really morally relevant. For this reason, we cannot trust a direct appeal to intuition to tell us whether Z is better than A.

2.3 Compounding small numbers

In many cases, we make intuitive errors when it comes to compounding very small quantities. In one study, psychologists found that people express greater willingness to use seatbelts when the *lifetime* risk of being injured in a traffic accident is reported to them, rather than the risk *per trip* (Slovic, Fischhoff, and Lichtenstein 1978). This suggests that, when the very small risk per trip is presented, people fail to appreciate how large the risk becomes when compounded over a lifetime. They may see the risk per trip as 'negligible', and so they neglect it, forgetting that a 'negligible' risk can be large when compounded many times.

For an especially dramatic illustration of the hazards of trusting quantitative intuitions, imagine that there is a very large, very thin piece of paper, one thousandth of an inch thick. The paper is folded in half, making it two thousandths of an inch thick. Then it is folded in half again, making it four thousandths of an inch thick. And so on. The folding continues until the paper has been folded in half fifty times. About how thick would the resulting paper be? Most people will estimate that the answer is something less than a hundred feet. The actual answer is about 18 million miles.[15]

For a case closer to our present concern, consider the common intuition that a single death is worse than any number of mild headaches. If this view is correct, it seems that a single death must also be worse than any amount of inconvenience. As Norcross observes, this suggests that we should greatly lower the national speed limit, since doing so would save some number of lives, with (only) a great cost in convenience.[16] Yet few support drastically lowering the speed limit. Indeed, one could imagine a great many changes in our society that would save at least one life at some cost in convenience, entertainment, or other similarly

[15] $(2^{50})(.001 \text{ inches})(1 \text{ foot}/12 \text{ inches})(1 \text{ mile}/5280 \text{ feet}) = 1.78 \times 10^7$ miles.

[16] Norcross (1997, pp. 159–60) discusses lowering the limit to 50 mph, but the same considerations would presumably support lowering it even more.

 © Huemer 2008

'minor' values. The result of implementing all of these changes would be a society that few if any would want to live in, in which nearly all of life's pleasures had been drained.

In all of these cases, we find a tendency to underestimate the effect of compounding a small quantity. Of particular interest is our failure to appreciate how a very small *value*, when compounded many times, can become a great value. The thought that *no* amount of headache-relief would be worth a human life is an extreme instance of this mistake—as is the thought that *no* number of low-utility lives would be worth as much as a million high-utility lives.

2.4 Underrating low-quality lives

When we imagine a low-quality life, even if we fill in a great many factual details, we may easily be unsure what its utility level is. When we imagine any realistic sort of life, we must be able to weigh complex combinations of goods and bads of various different kinds in order to arrive at any overall assessment of the life's utility level. Because of difficulties involved in judging such things as the weighting of values of very different kinds and whether and how values combine to form organic unities,[17] we may easily mistake a life with welfare level –1, for example, for one with welfare level 2. According to the advocates of (RC), the ability to distinguish such alternatives would be crucial for intuitively evaluating an imagined world of low average utility.

To avoid this problem, we might try imagining unrealistically simple lives, such as a life containing no evaluatively significant experiences or activities other than a uniform, mild pleasure. However, even the evaluation of a very simple life may be a complex matter. Our sense that we would be *bored* by experiencing a lifetime of such uniform, mild pleasure; that such a life would be *meaningless*; and that we would have to be seriously *mentally defective* to have no evaluatively significant other activities or states than this single pleasure, all may combine to give us a negative reaction to what we intended to be a slightly *positive* state.

For these reasons, it is not clear that our intuitions can be expected to reliably distinguish very slightly good lives from neutral or slightly bad lives. Thus, again, we should not trust the direct appeal to intuition to evaluate world Z.[18]

[17] An organic unity exists when the value of a whole exceeds the sum of the values of its parts. See Moore 1903, pp. 27–31.

[18] Ryberg (1996) and Tännsjö (2002) have suggested, in addition, that we tend to underrate what a life barely worth living is like, and that in fact privileged members of prosperous societies typically have lives only barely worth living. Some people find this claim much more plausible than do others.

 © Huemer 2008

In sum, there are several reasons to distrust the Unrepugnant Intuition: this intuition may be produced by a bias towards the world we would prefer to live in, a difficulty in grasping large numbers, a tendency to underrate the effect of compounding small quantities, and a difficulty in accurately picturing low-quality lives.

3. The failure of unrepugnant accounts

I turn now to six theories designed to explain why (RC) might be false. As each has been effectively criticized elsewhere, I shall only note briefly a few of the most damaging implications that have been drawn out of these theories, referring the reader to the literature for details.

First: the Average Utility Principle holds that the value of a world is determined solely by its average utility, rather than its total utility. Since world Z has a much lower average utility than A, Z is far worse. The Average Utility Principle has several counterintuitive consequences, one of the more striking of which is the following:

> *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 is because, starting from a high average utility, adding a large number of slightly happy people can lower the average by more than would adding a small number of unhappy people.[19] The same result follows from any *nonzero weighting* assigned to average utility. To see this, let ε be some small number and n some large number, and imagine three possible worlds (figure 2):

World B: n people at welfare level 100

World C: As in B, but with an extra unhappy person (level $-\varepsilon$)

World D: As in B, but with an extra n^2 slightly happy people (level ε/n^2)

[19] Arrhenius 2000. See Parfit (1984, pp. 420–2) for further criticisms of the Average Utility Principle.

 © Huemer 2008

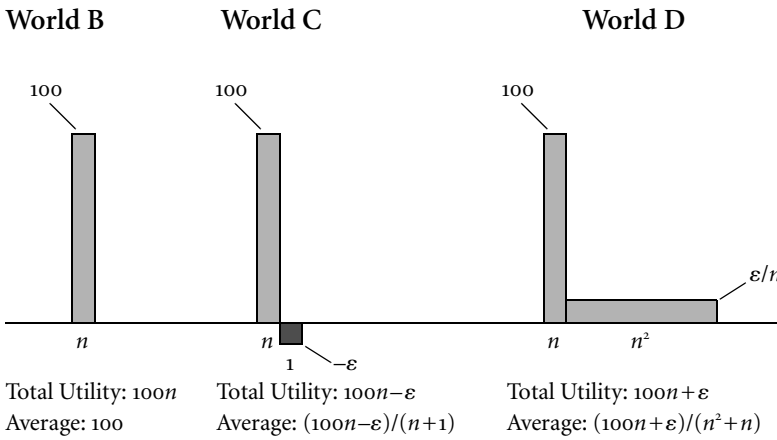

Fig. 2

As ε approaches 0 and n increases, the total utility of both C and D approaches $100n$, the average utility of C approaches 100, and the average utility of D approaches 0. Thus, by increasing n and decreasing ε, worlds C and D can be brought arbitrarily close in total utility, with an arbitrarily large ratio difference in average utility. Therefore, any nonzero weight assigned to average utility in the evaluation of worlds can be made to dictate a preference for world C over D, thus implying that, starting from world B, it would be better to add the unhappy person of world C than the many (slightly) happy people of world D.

Second: Critical Level principles hold that there is some threshold above zero at which lives begin to contribute to the world's value. Perhaps only lives above, say, welfare level 10 make the world better. Between 0 and 10, a life has value *to the subject* of the life but does not have impersonal value, that is, its existence does not improve the world. This hypothesis would enable us to avoid (RC) by holding that lives at welfare level 3 contribute no value to the world. Broome (2004, pp. 141–2) argues that, given a Critical Level theory, one should take lives below the critical level to *detract* from the world's impersonal value, rather than merely failing to augment it. As a result, the Critical Level theory leads to

> *The Strong Sadistic Conclusion:* For any world full of tormented people, a world full of people with lives barely worth living would be *worse*, provided that the latter world contained enough people.

 © Huemer 2008

This conclusion is much less plausible than the Repugnant Conclusion: (RC) holds that a sufficient number of slightly good lives are better than, say, a million wonderful lives. The Strong Sadistic Conclusion holds that a sufficient number of slightly good lives are *worse* than, say, a million *horrible* lives.[20]

Third: Narveson argues that world Z is better than world A only if there is someone *for whom* Z is better than A.[21] But there is no one for whom Z is better than A. Z is much worse than A for all the people who exist in A. Nor is Z better than A for the other 99 million people, who exist in Z but not in A, because *those* people have no welfare level at all in the world in which they do not exist. Since they have no welfare level at all in A, world A cannot be either better than, worse than, or as good as Z for them. Since A is better than Z for some people and worse for no one, A must be better overall.

Narveson's view has counterintuitive consequences. Suppose it were possible to slightly increase the welfare of presently-existing people while creating ten billion new people all of whom would lead lives of constant agony. On Narveson's view, this would not be worse than the actual world, for it could be worse only if it were worse for someone. By hypothesis, it would be better for the presently-existing people. And it would be neither better nor worse for the ten billion new sufferers, for they have no welfare level at all in the actual world. Since, on Narveson's view, the proposed change would benefit some actual people while harming no one, we would have to view it as an improvement.

Fourth: Variable-Value theories hold that lives have diminishing marginal value: the more people there already are, the less a life at a given welfare level contributes to the overall value of the world. The Variable Value theories described by Ng and Hurka are designed to approximate total utilitarianism for small populations, but to approximate average utilitarianism for large populations.[22] This enables them to avoid the Repugnant Conclusion; however, they fall prey to the same objections as the Average Utility Principle, including that they engender the Sadis-

[20] Blackorby, Bossert, and Donaldson (2004) embrace this amazing conclusion on the basis of a critical level theory. See Broome 1996, pp. 189–92, for criticisms of their view.

[21] Narveson 1967, p. 67; 1973. See Parfit 1984, p. 394, and Temkin 1987, pp. 166–7, for more discussion of the Person-Affecting Principle. Broome (1996, p. 179) endorses Narveson's argument. Boonin-Vail (1996, p. 268) appears to agree, stating that we should aim 'to produce more happiness for people, not to produce more people for happiness'. Both Narveson and Boonin-Vail are more concerned with what one ought to do than with what is good. To make it relevant to my present concern, I have recast Narveson's view as a view about goodness.

[22] Ng 1989; Hurka 1983. Ng does not endorse the Variable Value theory that he describes; he prefers the Total Utility Principle.

 © Huemer 2008

tic Conclusion. Sider's Variable Value theory avoids this consequence, but it is anti-egalitarian, apparently leading Sider himself to reject it.[23] That is, Sider's theory implies that there could be two worlds, such that one of them has a higher average utility, a higher total utility, and a more equal distribution of utility than the other, yet the former is worse than the latter with respect to utility.

Furthermore, all Variable Value theories face the Egyptology objection: these theories imply that the value that a person's life adds to the world can depend upon how numerous and/or how prosperous are the members of some remote society that has no interaction with the person in question. Thus, on a Variable Value theory, how much reason I now have to produce children depends in part on how happy the ancient Egyptians were and how many of them existed—even when the facts about the ancient Egyptians have no bearing on how my children's lives would go, nor on how anything else in the future would go. This seems absurd; as Parfit observes, 'research in Egyptology cannot be relevant to our decision whether to have children'.[24]

Fifth: Parfit's preferred alternative is *Perfectionism*, the view that 'even if some change brings a great net benefit to those who are affected, it is a change for the worse if it involves the loss of one of the best things in life' (Parfit 1986, p. 19). This view is highly anti-egalitarian. Taking Mozart's music (following Parfit) as an example of one of the best things in life, Perfectionism implies that enabling a few people to hear Mozart's music might be more important than providing food, shelter, and medical care to millions. At times, Perfectionism strikes even Parfit as crazy (1986, p. 20). As he acknowledges, there are artists, such as Haydn, who are only slightly inferior to Mozart, other artists who are only slightly inferior to them, and so on, continuing all the way down to Enya. Parfit suggests that it is the move from having Mozart to having only Haydn that marks an unbridgeable value gap, a loss that could not be compensated by any quantity of sub-Mozart goods. Yet he acknowledges that Mozart's music is only *slightly* better aesthetically, not infinitely better, than Haydn's. It also seems that a person who hears Mozart's music is only slightly better off, not infinitely better off, than a person who has heard only Haydn's. Since aesthetic value and contributions to human welfare seem to be the only dimensions relevant to

[23] Sider 1991, p. 270, n. 10. Arrhenius (2000, pp. 252–4) explains the problem.

[24] Parfit 1984, p. 420 (discussing the Average Utility Principle). I have assumed that the variable value theory applies to the world's total population over all time. For theories that apply only to the population existing at a given time, replace the Ancient Egyptians with some remote contemporary society.

 © Huemer 2008

assessing the value of Mozart's music, and Mozart only slightly exceeds Haydn on these dimensions, it is hard to see where the unbridgeable value gap could come from.[25]

Sixth: Stuart Rachels and Larry Temkin hold that some intense pleasures are 'lexically better than' any mild pleasure, meaning that no amount of mild pleasure is as good as a given, short duration of the intense pleasure.[26] But they do not think that an unbridgeable value gap arises at some particular point as we move gradually from ecstasy down to the mildest pleasure: at every point, a small decrease in intensity could be made up for by a large increase in duration.[27] Rather, they think betterness is non-transitive: even though one could move through a series of cases, from a short duration of ecstasy to a very long duration of mild pleasure, where each case was better than the previous one, the final case would be *worse* than the first.

I have discussed the arguments for Transitivity in section 1.4. Here I limit myself to assessing the central intuition behind Rachels's theory, the intuition that no amount of mild pleasure, however protracted, is as good as, say, fifty years of ecstasy (Rachels 1998a, p. 76). Some people, including myself, have no such intuition. Moreover, people's intuitions seem to shift when the length of the ecstatic experience is shortened: a mere second of ecstasy seems inferior to a thousand years of mild pleasure.[28] This can be explained by a particular error theory: we have difficulty grasping very long time periods. The duration of a mild pleasure that is really superior to fifty years of ecstasy is too long for us to adequately grasp; hence, we fail to appreciate its superiority. To alleviate this problem, we may replace the fifty years of ecstasy with a very short (but still clearly graspable) period of ecstasy—say, one second—and then ask whether we can imagine a superior experience consisting of protracted mild pleasure. When we thus change the example to improve the reliability of our intuition, the ecstatic experience no longer seems categorically better.

[25] For further arguments against Perfectionism, see Rachels 2001, p. 230; 2004, p. 178.

[26] See Rachels 1998a; 2001. Temkin's (1996) argument is based on Rachels.

[27] That is, for any given pleasure, it is always possible to imagine a better pleasure that is longer but slightly less intense. However, as Rachels (personal communication) has pointed out, his view *does* imply that there is a threshold level of intensity below which a pleasure is lexically inferior to a given ecstatic experience: pleasures just above the threshold can, if sufficiently protracted, be superior to, say, an hour of ecstasy; but pleasures just below the threshold are always inferior to an hour of ecstasy, no matter how long they last. This creates an implausible sort of value gap between what may be nearly indistinguishable experiences.

[28] Rachels (1998a, p. 77; 2001, pp. 219–20; personal communication) seems to acknowledge this shift of intuitions.

 © Huemer 2008

In sum, every attempt to escape the Repugnant Conclusion lands us in worse trouble. If the situation were merely that every anti-(RC) theory anyone had devised *so far* had some implausible consequence or other, then we might hold out hope for some as-yet-undiscovered theory that would rescue us from the 'paradoxes' of population ethics. But in fact, we know there is no such theory, because any theory that avoids (RC) must reject the Modal Pareto Principle, Transitivity, or Non-anti-egalitarianism—and any of these options would make the theory strongly counterintuitive. The Average Utility Principle, the Critical Level Theory, and Ng's and Hurka's Variable Value Principles all conflict with the Modal Pareto Principle. Perfectionism and Sider's Variable Value Principle are anti-egalitarian. Lexicality and Narveson's theory violate Transitivity.[29] And none of these theories have a compelling motivation beyond the desire to avoid (RC) and related conclusions. It is time to stop searching for a solution to the 'paradoxes' of population ethics and simply embrace the Repugnant Conclusion.

4. The actualist bias

In this and the following two sections, I discuss three further arguments in support of (RC), each of which may succeed even if the other arguments for (RC), including the Benign Addition Argument, should fail.

Our intuitions about (RC) seem to reflect a bias in favour of present actual people, as opposed to potential future people. One manifestation of actualist or presentist bias is found in the contrast between our prospective and our retrospective attitudes towards creating people. When it comes to creating a new person, we accept relatively weak considerations as showing that creation is undesirable; but when it comes to evaluating a present actual person's existence, we demand very strong grounds before concluding that it would have been better had that person never existed. Thus, consider two cases:

> *Jon and Mary's Potential Child:* Jon and Mary are considering whether to have a child. They are confident that any such child would have a life well worth living. But they already have two children, and raising another would entail a fair amount of inconvenience; overall, Jon and Mary would be slightly worse off. Knowing all this, they ask a friend for advice. The friend advises them: 'It would be better that you not have a third child.'

[29] Temkin (1987, pp. 152–3) cites this as a reason to reject Transitivity.

 © Huemer 2008

> *Jon and Mary's Actual Child:* Jon and Mary have disregarded the friend's advice in the above scenario and had a third child, Sally. Twenty years later, the same friend is having dinner with Jon, Mary, and their three children. As anticipated, Jon and Mary's welfare was slightly lowered overall by their having Sally, but Sally has and will continue to have a life well worth living. Remembering his assertion of twenty years ago, which he has seen no reason to revise, the friend announces: 'It would have been better had Sally never been born.'

I think that the friend's remark would widely be regarded as perfectly reasonable in the first case, but as both unreasonable and offensive in the second. How can this be? The retrospective statement, 'It would have been better had Jon and Mary not had a third child', is just the past tense of the prospective statement, 'It would be better if Jon and Mary did not have a third child', so presumably the two statements have the same truth-value. And if the friend was *justified* in making the prospective assertion, then surely he would later be justified in making the retrospective assertion, given that everything turned out exactly as expected.

Perhaps the prospective and retrospective statements have different truth-values due to a hidden context-sensitive element in their meanings. Perhaps by 'It would be better that you not have a child', the friend meant merely, 'It would be better *for you* that you not have a third child'. And perhaps 'It would have been better had Sally never been born' would normally be interpreted to mean that it would have been better *for Sally*, or for the family (including Sally), or for society as a whole, had Sally never been born. But this does not fully capture our attitudes. For in addition to thinking that it might be better *for the parents* that a particular pair of parents refrain from having a third child, most of us also have the attitude that what is good for the parents, itself, should carry more weight in the parents' deliberations than what would be good for the prospective child, and perhaps even that the interests of this potential child should not count at all. If the friend had said only, 'It would be better *for you* that you not have a third child', Jon and Mary might have replied, 'Yes, we know that. But, if we have a third child, it will be better *for her* that we had her. So what do you think is best, overall?' Here, the friend might have said, 'Overall, it is best that you refrain from having the child', and this judgement would likely be accepted, if not as obviously correct, at least as a reasonable position.

I take it, then, that our relatively greater sympathy with the friend's remark in *Jon and Mary's Potential Child*, as compared with his remark in *Jon and Mary's Actual Child*, indicates that we assign more weight, in

 © Huemer 2008

evaluating the world as a whole, to the interests of people who actually, presently exist than to people who merely might come into existence. At the time of the friend's first remark, Sally is only a potential future person. At the time of his second remark, she is a present actual person. Consequently, we are much more inclined to consider Sally's interests in the second than in the first case. But as I have suggested, this must be a mistake: the world with Sally in it is either better than, worse than, or exactly as good as the world that would have obtained if Sally had not been created. Which of these is correctly said to be the case cannot depend upon the time at which one is speaking.

I suggest that our intuitive evaluation of Jon and Mary's Actual Child is the correct one. Our evaluation of Jon and Mary's Potential Child is skewed because of our difficulty in sympathising with people who are not present—in this case, Sally is non-present in the particularly strong sense of not existing at the time. If I am right, the lesson is that a decrease in the utility of present people can be made up for by the addition of new people with worthwhile lives. This lends support to the Repugnant Conclusion. If Sally's birth was good despite its slightly lowering the utility of some previously-existing people, then presumably the births of many more people might be good, even if they each slightly lowered the utility of some pre-existing people. If that is so, then it is possible to improve the world by increasing the population while decreasing the world's average utility.[30] This at least suggests that a series of improvements taking us from an A-like world to a Z-like world would be possible. The Repugnant Conclusion could be blocked by either a Critical Level principle or an intransitive theory, but we have seen reasons for rejecting such theories above.

5. The Equivalence Argument

I shall speak of a *dimension* or variable as being 'at least equivalent' to another, just in case an increase in the former is at least as good as a proportionate increase in the latter. For instance, if the intensity of a pleasurable experience is at least equivalent to the duration of the experience, then a doubling of intensity is at least as good as a doubling of duration. If two variables are each at least equivalent to the other, then I call the two variables 'equivalent'. The relation '*x* is at least equivalent to *y*' is reflexive and transitive, given that '*x* is at least as good as *y*' is reflexive and transitive.

[30] I assume that Sally's utility is at or below that of her parents after the parents have had her.

 © Huemer 2008

The Equivalence Argument posits two instances of the at-least-equivalent-to relation:

(1) Duration of a benefit is at least equivalent to intensity of benefit

(2) Number of recipients of a benefit is at least equivalent to duration of benefit

(3) Therefore, for populations with positive utility, population size is at least equivalent to average utility (from (1), (2))

(4) If (3), then the Repugnant Conclusion is true

(5) So the Repugnant Conclusion is true (from (3), (4))

The 'intensity' of a benefit is a matter of how much it raises one's level of well-being during the time one enjoys the benefit. Premiss (1) tells us, for example, that experiencing a welfare level of 10 for ten minutes is at least as good as experiencing a welfare level of 20 for five minutes. To illustrate, suppose there are two benefits, E_1 and E_2, either of which, by itself, would confer on one a welfare level of 10 while one enjoyed it. E_1 might be, say, the pleasure of watching an Enya music video, and E_2 the pleasure of eating a cucumber sandwich. Suppose that neither benefit either enhances or interferes with the other, regardless of the timing (for example, one's enjoyment of Enya is completely unaffected by one's eating of a cucumber sandwich, and vice versa). It seems that one way of having a welfare level of 20 would be to have E_1 and E_2 at the same time. Furthermore, it seems that having E_1 and E_2 in sequence is, given our stipulations, as good as having E_1 and E_2 simultaneously (where each benefit lasts the same length of time either way); it does not matter whether you eat the sandwich while watching the video, or eat the sandwich first, then watch the video. Also, it does not matter whether one enjoys a welfare level of 20 by having two or more simultaneous benefits, or by having a single, better benefit, as long as one's overall welfare level really is the same in either case. Therefore, having a welfare level of 10 for some length of time is just as good as having a welfare level of 20 for half as long. Using '$\geq_v$' to denote the at-least-equivalent-to relation, we can summarize the argument for (1) as follows:

(1a) [Having welfare 20 for five minutes through having two simultaneous, level-10 benefits] is as good as [having welfare 20 for five minutes in any other manner] (Premiss)

(1b) [Having welfare 10 for ten minutes through having two benefits sequentially] is as good as [having welfare 10 for ten minutes in any other manner] (Premiss)

 © Huemer 2008

(1c) [Having welfare 10 for ten minutes through having two level-10 benefits in sequence, each for five minutes] is at least as good as [having welfare 20 for five minutes through having the same benefits simultaneously] (Premiss)

(1d) Therefore, [having welfare 10 for ten minutes] is at least as good as [having welfare 20 for five minutes] (from (1a), (1b), (1c))

(1e) If (1d), then duration of welfare $\geq_v$ intensity of welfare (Premiss)

(1) Therefore, duration of welfare $\geq_v$ intensity of welfare (from (1d), (1e))

Premiss (1e) is plausible, since there seems to be nothing special about welfare level 20, about the duration 'five minutes', and so on.

I turn to premiss (2) of the Equivalence Argument. (2) tells us, for example, that a state of affairs in which two people each experience a welfare level of 10 for five minutes is at least as good as one in which a single person experiences a welfare level of 10 for ten minutes.[31] For example, if Sue watches Enya and Mary eats a cucumber sandwich, this is at least as good as if Sue watches Enya and then eats a cucumber sandwich (provided Sue and Mary are relevantly similar—for example, neither has greater desert, and neither has a greater moral claim on the sandwich than the other). Why believe this? The intrinsically good-making feature of the sandwich-eating experience—namely, its pleasurableness—is equally present in either case. Neither Sue nor Mary is more important than the other. So it seems that, from an impartial standpoint, it is *at least* as good for Mary to get the sandwich as it is for Sue to get it. (Egalitarians may hold that it is *better* for Mary to get the sandwich, since this would provide a more equal distribution of benefits.)

Some would accept this claim when Sue and Mary both already exist, but resist the claim if Mary is a new person who exists in only one alternative. To illustrate, suppose that Sue and Mary are qualitatively indistinguishable though distinct persons, and consider two possible worlds (see figure 3):

World G: Sue exists for ten minutes, experiencing E_1 for the first five minutes, followed by E_2 for the last five minutes. Mary does not exist.

World H: Sue exists for five minutes, experiencing E_1. Then she ceases to exist and Mary pops into existence, experiencing E_2 for five minutes.

[31] This is an application of Rachels's (2001, pp. 214–15) Conflation Principle.

 © Huemer 2008

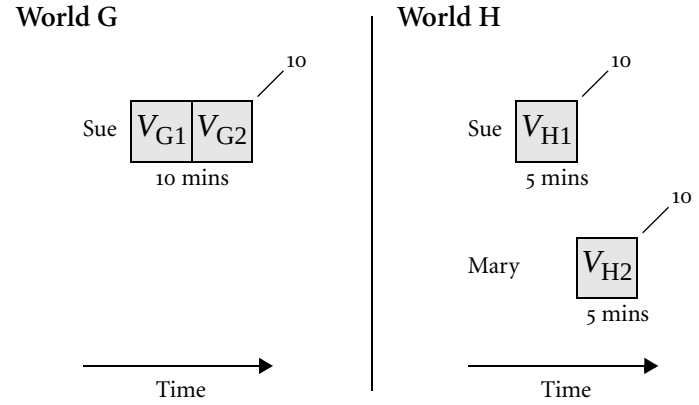

Fig. 3

Some would argue that world G is better than world H, because it is better, all else equal, to give a benefit to an already existing person than to create a new person to receive that benefit; similarly, it seems worse for an existing person to die prematurely than for a potential person not to be created.

I think G and H are equally good. My argument is as follows. Let 'V_{G1}' denote the value of the earlier half of world G, which consists in Sue's enjoying E_1 for five minutes; let 'V_{G2}' denote the value of the later half of G; and let 'V_{H1}' and 'V_{H2}' be understood analogously. It seems that the two temporal halves of world G are equally good, since each consists in Sue's enjoying an equally good benefit for five minutes. It is hard to think of a reason why one of these five-minute events should be better than the other. Similarly, it seems that the two temporal halves of world H are equally good. Sue and Mary are qualitatively indistinguishable people who enjoy equal benefits for equal lengths of time. So it is hard to see how one of their lives would be better than the other. Finally, it seems that the first half of world G and the first half of world H are equally good. Both consist in Sue's enjoying the same benefit for the same length of time. If the intrinsic value of an event supervenes on its intrinsic, non-evaluative properties, then the first halves of G and H must have equal intrinsic value. Thus, we have that $V_{G2} = V_{G1} = V_{H1} = V_{H2}$. Therefore, $V_{G1} + V_{G2} = V_{H1} + V_{H2}$. Furthermore, it seems that the value of world G is $V_{G1} + V_{G2}$, while that of world H is $V_{H1} + V_{H2}$.[32] So worlds G and H are equally good. And since there is nothing special about these particular benefits and these time periods, it seems that duration of welfare—the length of time that some person or group of

[32] Elsewhere (2003, pp. 157–62), I defend the assumption that the value of events adds over time.

 © Huemer 2008

people enjoys some benefit—is evaluatively equivalent to widespreadness of welfare—the number of people enjoying the benefit.

We have just seen the plausibility of premisses (1) and (2) of the Equivalence Argument. It seems to follow that

(3) For populations with positive utility, population size is at least equivalent to average utility

To see this, consider the four possible worlds depicted in figure 4. In F, one person has welfare level 20 for five minutes. In G, one person has welfare 10 for ten minutes. In H, two people each have welfare 10 for five minutes, one after the other. And in I, two people each have welfare level 10 for five minutes, simultaneously. It seems that world I is at least as good as H. Given what I have said above in this section, H is at least as good as G, which is at least as good as F. Therefore, I is at least as good as F. World I amounts to a doubling of population with a halving of average utility, relative to F. If all this is right, then—since there is nothing special about the welfare levels, durations, and numbers of people in this example—the size of a group of people is at least equivalent to the average utility of the group. If so, then a drop in average utility can be compensated for by a proportional increase in population—and so the Repugnant Conclusion is true.

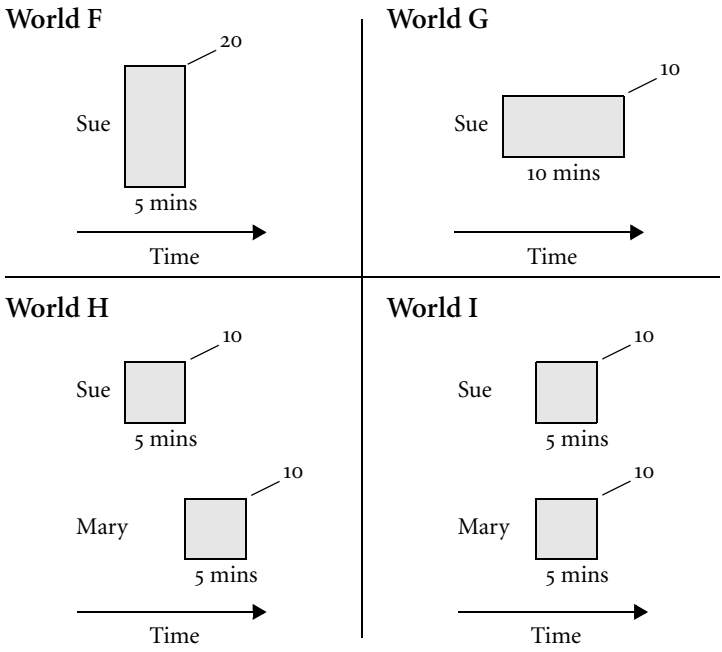

Fig. 4

 © Huemer 2008

6. The More-is-Better Argument

A very simple, natural explanation for the truth of (RC) is available: Worthwhile lives are good. More of a good thing is better. Therefore, increasing the number of worthwhile lives makes the world better. If we add enough such lives, we can produce arbitrarily large value. In sum, we can argue:

(1) It is better for there to be more lives with positive welfare

(2) The marginal value of such lives does not diminish so as to create an upper bound to the value of such lives

(3) If (1) and (2), then the Repugnant Conclusion is true

(4) So the Repugnant Conclusion is true (from (1), (2), (3))

Let us turn to the reasons for accepting each of the above premisses.

6.1 Life: more is better

There are at least three reasons for thinking that, when it comes to happy people, more is better. First, consider the symmetry between time and space. Most will agree that, as long as human beings have worthwhile lives, it is better for the human species to survive longer, and that our continued survival does not have diminishing marginal value as the age of the species increases. In particular, there does not come a point, after the species has existed so long, at which continued survival becomes of little or no value. If anthropologists should discover that the human species is older than previously thought, or that more people have existed than previously thought, we would not thereupon have less reason to avoid a catastrophic nuclear war. But it seems that the value that a life with a given welfare level contributes to the world's total value should be independent of *when* that life occurs. So adding more people *elsewhere in space* should be as good as adding more people *later in time*. Adding people elsewhere in space is just increasing the population. So additions to the population also have non-diminishing marginal value. This supports (1) and (2).

Second, consider the symmetry between pleasure and pain. Nearly everyone will agree that, if a potential person would have an overall *painful* life, then it would be *bad* to bring him into existence. If so, then by analogy, it seems that if a person would have an overall *pleasurable* life, it would be *good* to bring him into existence.[33]

[33] This argument derives from Rachels (1998b, pp. 104–5).

 © Huemer 2008

Finally, premiss (1) can be supported by the following sub-argument. Let x be any positive welfare level.

(1a) Existence at welfare level $x/2$ is at least as good as non-existence

(1b) The existence of some number of people at welfare level x is better than the same people's existing at welfare level $x/2$

(1c) Therefore, the existence of some number of people at welfare level x is better than those people's not existing at all (from (1a), (1b))

Some would deny (1a), on the grounds that when one does not exist, one has no welfare level at all, and therefore, such an alternative is *incomparable* in terms of one's own interests to any alternative in which one exists (Narveson 1967; 1973). To see why this is a mistake, consider an analogy. Suppose you are asked to choose between two possible futures:

Future F^+: You continue to exist, with positive utility

Future X: You immediately cease to exist

Prudentially, you should prefer F^+ to X. We should reject the quasi-Epicurean claim that, since you have no future welfare level in option X, that future is for you prudentially incomparable to any future in which you exist. By analogy, it seems that we should reject the claim that, since you have no welfare level in a possible world in which you never exist, such a world is incomparable in terms of your interests to any world in which you exist. Just as the future in which you no longer exist is worse, from the standpoint of your interests, than one in which you continue at a positive welfare level, a possible world in which you never exist is worse, from the standpoint of your interests, than one in which you exist at a positive welfare level.[34] It therefore seems that, other things being equal, the world is better when you exist with positive utility than it would be without you; at least it is surely no worse.

Thus, we may consider the following possible worlds (figure 5):

World J: No people

World K: One person at welfare level ½

World L: One person at welfare level 1

[34] Parfit (1984, pp. 487–90) and Rachels (1998b, pp. 105–6, 107) have pressed the analogy between parts of lives and whole lives.

 © Huemer 2008

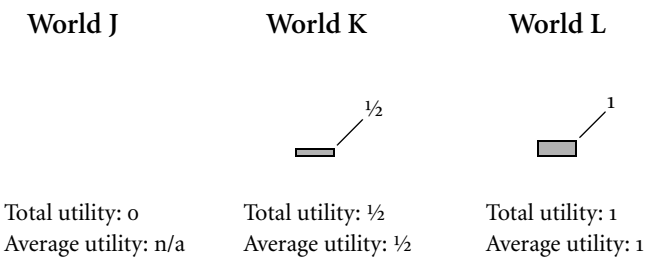

Fig. 5

L is better than K, which is at least as good as J. Therefore, L is better than J. Therefore, creating happy people, even slightly happy people, is good.

6.2 The value of happy people does not diminish

Many goods have diminishing marginal value: the more of the good one already has, the less value there is in receiving a given addition to that good. For a person with no money, $100 is very valuable. But for a millionaire, $100 is almost worthless. The same holds for less tangible goods: having spent three hours talking to a friend, I value a fourth hour of conversation on the same day much less than I did the first. Having visited Paris twelve times, I will likely get less out of the thirteenth visit. Likewise, perhaps *lives* have diminishing marginal value: perhaps, once we already have a billion people, the next happy person to be born adds less to the value of the world than the first person, even if his life is otherwise the same (Hurka 1983; Sider 1991; Ng 1989).

This supposition is not *obviously* false. But why would it be true? We know why the principle of diminishing marginal value holds for most goods. It is because most goods are instrumental, and instrumental goods tend to contribute less to what is intrinsically valuable, the more of the instrumental good one already has. Wealth has diminishing marginal value because the more wealth one already has, the less enjoyment or well-being one can derive from a fixed-size addition to that wealth. Roughly speaking, a poor man, given $100, may spend it on life-sustaining food, where a rich man might spend that same $100 on a silver-plated back scratcher. Similarly, visits to Paris have diminishing marginal value because subsequent visits (especially within a short time period) tend to produce smaller amounts of enjoyment, learning, and whatever else one goes to Paris for. This *need* not be the case, but the fact that it is usually the case explains why trips to Paris usually have diminishing marginal value.

 © Huemer 2008

This very satisfying explanation for the law of diminishing marginal value cannot apply to utility itself, nor to any other intrinsic value. Wealth, conversation, and visits to Paris have diminishing marginal value because they make diminishing marginal contributions to utility. But *utility* does not make diminishing marginal contributions to utility. So there is no apparent reason why the marginal value of utility or any other intrinsic good should diminish. These remarks do not prove that worthwhile lives do *not* have diminishing marginal value; it is not *incoherent* to ascribe diminishing marginal value to such lives. The point here is that the standard *reason* for ascribing diminishing marginal value to other things cannot be applied to worthwhile lives, and that no other obvious reason for ascribing diminishing marginal value to worthwhile lives suggests itself.

Furthermore, most people will agree that the marginal *disvalue* of *unhappy* lives does not diminish: no matter how many miserable souls there already are who wish they had never been born, it remains just as bad to create another one. If this is so, then it seems that the marginal value of happy lives must also be non-diminishing, because sufficiently many happy lives can always compensate for a given number of unhappy lives. Thus, suppose you can create one million (at least slightly) happy people and one slightly unhappy person. It seems that this would be good, or at least not bad. (If this is not so, then it must be bad for the human species to continue, since unhappy people make up more than a millionth of each generation.) If so, it seems that it would not be bad to perform many such acts of creation. So in general, it would not be bad to create any number of slightly unhappy people while also creating a million times as many happy people. But this would not be so if the marginal value of happy lives diminished as we created more of them.[35] Consider the following sequence of cases:

World L_1: 1 person at welfare level –10, plus 1 million people at +10

World L_2: 2 people at welfare level –10, plus 2 million people at +10.

…

World L_n: n people at welfare level –10, plus $n \times 10^6$ people at +10

…

If, as we go through the above series, the total disvalue of the unhappy lives increases linearly, while the total value of the happy lives increases

[35] This argument derives from Sikora (1978, pp. 140–5). Cf. Anglin 1977, pp. 752–4; Rachels 1998b, p. 104.

 © Huemer 2008

less than linearly, approaching some upper bound, then the disvalue of the unhappy lives must at some point exceed the value of the happy lives (figure 6). This would give us a variant of what Parfit (1984, pp. 410–11) calls the Absurd Conclusion: that the creation of some number of slightly unhappy people together with a million times as many happy people would be bad. If that conclusion seems absurd, we should avoid it by rejecting the hypothesis of diminishing marginal value for happy people.

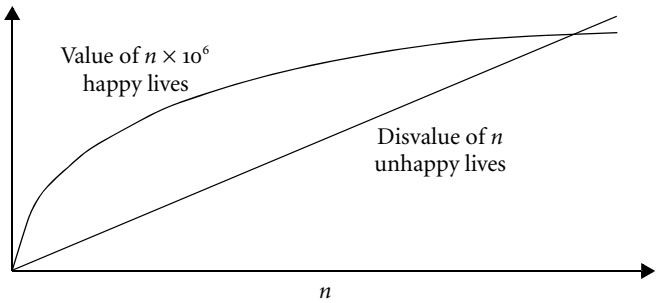

Fig. 6

6.3 Non-diminishing value implies (RC)

A sufficient amount of something very small is very large: A speck of dust is very light; even so, a sufficiently large pile of dust would be heavier than Mount Everest. A piece of rice paper is very thin; even so, a stack of sufficiently many pieces of rice paper would reach to the moon. Similarly, if worthwhile lives have non-diminishing marginal value, then a sufficient number of them can exceed any assigned amount of value, even if each is only slightly worthwhile.

Contrary to what is sometimes suggested in the literature, there is nothing paradoxical in this conclusion. It is a trivial consequence of the most simple and natural account of the value of lives: namely, that the value of lives is additive.

7. Taking stock: the case for revision

It is not always right to revise particular intuitions in the light of theory; sometimes, theory goes wrong. It is not always right to follow the argument where it leads; sometimes, reasoning goes astray. We bear that in mind when dealing with paradoxes such as Zeno's Paradox or the Liar Paradox. Perhaps we are also right to resist arguments for such counterintuitive conclusions as philosophical scepticism or the denial of free will, even when we cannot say where they go wrong.

 © Huemer 2008

But *sometimes* it is right to follow the argument where it leads; *sometimes* starting intuitions should be revised. As I have suggested, the case for revising an initial, intuitive judgement is strongest when:

(1) The judgement conflicts with strong and widely-shared other beliefs

(2) We have a plausible error theory, explaining what sort of biases or mistakes might generate the judgement

(3) There are multiple, independent lines of argument against the judgement

(4) The judgement lacks significant argumentative support or a credible theoretical explanation; and

(5) The contrary view has a simple, natural theoretical explanation

When only condition (1) is satisfied, as we often find, the case for revision is shaky; then reasonable people may disagree about which belief to revise. But when *all five* of these criteria are satisfied, only through dogmatism can we cling to the initial judgement.

As we have seen, the intuition opposing the Repugnant Conclusion satisfies all five criteria for meriting revision. It conflicts with the conjunction of the Modal Pareto Principle, Non-anti-egalitarianism, and Transitivity, each of which seems obviously correct considered on its own. Though the Benign Addition Argument is the strongest argument for (RC), (RC) has at least three auxiliary arguments standing behind it, the Actualist Bias Argument, the Equivalence Argument, and the More-is-Better Argument. Against this, opponents of (RC) can muster only a bare appeal to intuition, and that intuition is under suspicion of contamination by multiple biasing factors. Several theories have been proposed to explain why (RC) might be false, but each rests on one or more dubious assumptions and generates unacceptable consequences. Finally, we have a straightforward, natural explanation for why (RC) would be true, namely, that the value of lives is additive. Other things being equal, adding more of something intrinsically good makes the result just that much better. (RC) is a simple consequence of this.

8. Repugnant in theory, congenial in practice

I have sided not only with (RC) but with its logically stronger brother, the Total Utility Principle. What is the practical import of my conclusion? Should we, in fact, aim at a drab future like world Z, where each of our descendants occupies a single, cramped room and there is just enough gruel to keep them from hunger?

 © Huemer 2008

Given any plausible view about the actual effects of population growth, the Total Utility Principle supports no such conclusion. Those who worry about population growth believe that, as the population increases, society's average utility will decline due to crowding, resource shortages, and increasing strain on the natural environment (Ehrlich and Ehrlich 1990). On this view, the graph of average utility versus population size might look like figure 7. The curve represents how our welfare level will decline as our numbers grow: at low levels of population (the left-hand part of the curve), increases in population will make little or no difference to average utility. But as the population increases, further additions will start to have greater impacts on our average level of well-being, until eventually we are so cramped and are subsisting on such bare resources that average welfare goes negative. According to the Total Utility Principle, the optimum is the point where total utility is greatest. This is shown as point P on the diagram. The total utility is the area of rectangle PO, since this is equal to the population times the average utility. Q, on the other hand, represents a crowded world with low positive welfare. The total utility of this world is the area of rectangle QO, which is much smaller (in moving from P to Q, we lose the area of PR and gain only the area of QS).

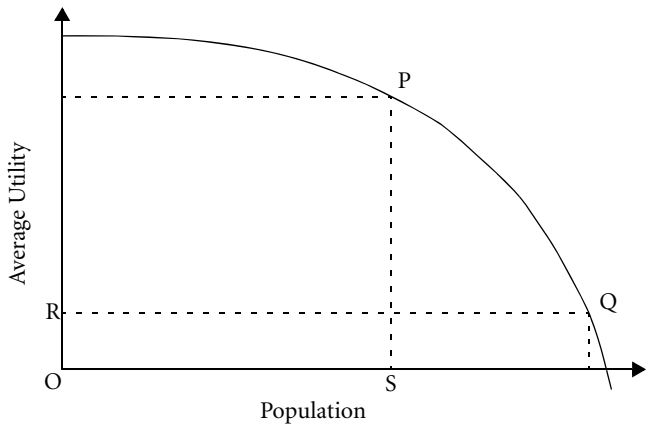

Fig. 7

Some people believe that population increases will lead to *increases* rather than declines in average utility, for the foreseeable future (Simon 1996). Their reasoning is that economic and technological progress will be accelerated as a result of the new people, and that technology will solve any resource or environmental problems that we would otherwise face. On this view, we should strive to increase population indefinitely, and as we do so, we and our descendants will be better and better off.

 © Huemer 2008

Perhaps the most plausible view is something in between: at low population levels, increases in population improve average welfare due to such factors as economies of scale and fruitful interactions among diverse people; this is before the population has become large enough to put a strain on resources or the environment. But at very high population levels, further increases in population decrease average welfare for the traditional reasons. The correct population-utility graph probably looks something like figure 8. Again, the optimum point is P, and Q represents a low-average-utility alternative that clearly has lower total utility.

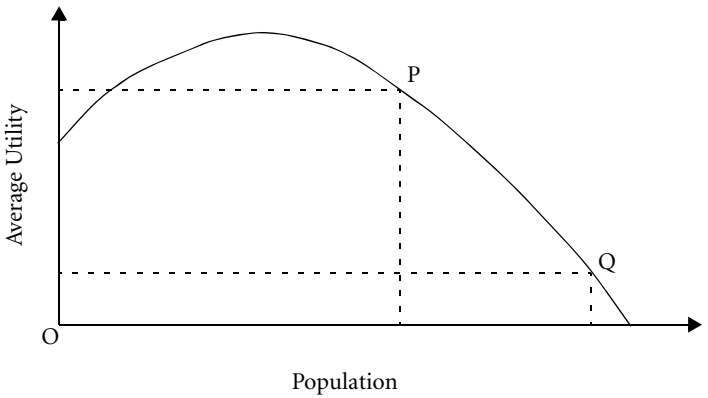

Fig. 8

These graphs are only qualitatively correct. To determine the population-utility curve with any precision would require detailed empirical research. Nevertheless, the graphs suffice to make the point that the Total Utility Principle does not enjoin us, in reality, to pursue the world of cramped apartments and daily gruel. Perhaps its critics will therefore look upon the principle with less revulsion than has hitherto been customary.[36]

Philosophy Department
University of Colorado
Boulder, CO 80309
USA

MICHAEL HUEMER

[36] I owe thanks to the editor of this journal for helpful comments on this paper, as well as Stuart Rachels, whose comments were great both in number and in quality.

 © Huemer 2008

References

Andreou, Chrisoula 2006: 'Environmental Damage and the Paradox of the Self-Torturer'. *Philosophy and Public Affairs*, 34, pp. 95–108.

Anglin, Bill 1977: 'The Repugnant Conclusion'. *Canadian Journal of Philosophy*, 7, pp. 745–54.

Arrhenius, Gustaf 2000: 'An Impossibility Theorem for Welfarist Axiologies'. *Economics and Philosophy*, 16, pp. 247–66.

Attfield, Robin 1991: *The Ethics of Environmental Concern*, 2nd edition. Athens, GA: University of Georgia Press.

Blackorby, Charles, Walter Bossert, and David Donaldson 2004: 'Critical-Level Population Principles and the Repugnant Conclusion'. In Ryberg and Tännsjö 2004, pp. 45–59.

Boonin-Vail, David 1996: 'Don't Stop Thinking About Tomorrow: Two Paradoxes About Duties to Future Generations'. *Philosophy and Public Affairs*, 25, pp. 267–307.

Broome, John 1996: 'The Welfare Economics of Population'. *Oxford Economic Papers*, 48, pp. 177–93.

——2004: *Weighing Lives*. Oxford: Oxford University Press.

Cowen, Tyler 2004: 'Resolving the Repugnant Conclusion'. In Ryberg and Tännsjö 2004, pp. 81–97.

Davidson, Donald, J. C. C. McKinsey, and Patrick Suppes 1955: 'Outlines of a Formal Theory of Value'. *Philosophy of Science*, 22, pp. 140–60.

Ehrlich, Paul and Anne Ehrlich 1990: *The Population Explosion*. New York: Simon and Schuster.

Fotion, Nick 1997: 'Repugnant Thoughts About the Repugnant Conclusion Argument'. In Fotion and Heller 1997, pp. 85–97.

Fotion, Nick and Jan C. Heller (eds) 1997: *Contingent Future Persons*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Geach, Peter 1956: 'Good and Evil'. *Analysis*, 17, pp. 33–42.

Huemer, Michael 2003: 'Non-Egalitarianism'. *Philosophical Studies*, 114, pp. 147–71.

——2005: *Ethical Intuitionism*. New York: Palgrave Macmillan.

——2008: 'Revisionary Intuitionism'. *Social Philosophy and Policy*, 25, pp. 368–92.

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

Locke, Don 1987: 'The Parfit Population Problem'. *Philosophy*, 62, pp. 131–57.

Mackie, John 1977: *Ethics: Inventing Right and Wrong*. New York: Penguin.

 © Huemer 2008

Moore, George Edward 1903: *Principia Ethica*. Cambridge: Cambridge University Press 1960.
Mulgan, Tim 2002: 'The Reverse Repugnant Conclusion'. *Utilitas*, 14, pp. 360–4.
Narveson, Jan 1967: 'Utilitarianism and New Generations'. *Mind*, 76, pp. 62–72.
——1973: 'Moral Problems of Population'. *The Monist*, 57, pp. 62–86.
Ng, Yew-Kwang 1989: 'What Should We Do About Future Generations? The Impossibility of Parfit's Theory X'. *Economics and Philosophy*, 5, pp. 135–253.
——1990: 'Welfarism and Utilitarianism: A Rehabilitation'. *Utilitas*, 2, pp. 171–93.
Norcross, Alastair 1997: 'Comparing Harms: Headaches versus Human Lives'. *Philosophy and Public Affairs*, 26, pp. 135–67.
Nozick, Robert 1993: *The Nature of Rationality*. Princeton, NJ: Princeton University Press.
Parfit, Derek 1984: *Reasons and Persons*. Oxford: Clarendon.
——1986: 'Overpopulation and the Quality of Life'. In Ryberg and Tännsjö 2004, pp. 7–22. Originally published in Singer 1986.
Quinn, Warren 1990: 'The Puzzle of the Self-Torturer'. *Philosophical Studies*, 59, pp. 79–90.
Rachels, Stuart 1998a: 'Counterexamples to the Transitivity of *Better Than*'. *Australasian Journal of Philosophy*, 76, pp. 71–83.
——1998b: 'Is it Good to Make Happy People?' *Bioethics*, 12, pp. 93–110.
——2001: 'A Set of Solutions to Parfit's Problems'. *Noûs*, 35, pp. 214–38.
——2004: 'Repugnance or Intransitivity: A Repugnant but Forced Choice'. In Ryberg and Tännsjö 2004, pp. 163–86.
Ryberg, Jesper 1996: 'Is the Repugnant Conclusion Repugnant?' *Philosophical Papers*, 25, pp. 161–77.
Ryberg, Jesper and Torbjorn Tännsjö (eds) 2004: *The Repugnant Conclusion: Essays on Population Ethics*. Dordrecht, The Netherlands: Kluwer Academic Publishers.
Sider, Ted 1991: 'Might Theory X Be a Theory of Diminishing Marginal Value?' *Analysis*, 51, pp. 265–71.
Sidgwick, Henry 1907: *The Methods of Ethics*, 7th edition. Indianapolis: Hackett. Reprinted 1981.
Sikora, Richard I. 1978: 'Is It Wrong to Prevent the Existence of Future Generations?' In Sikora and Barry 1978, pp. 112–66.
——1981: 'Classical Utilitarianism and Parfit's Repugnant Conclusion: A Reply to McMahan'. *Ethics*, 92, pp. 128–33.

 © Huemer 2008

Sikora, Richard I. and Brian Barry (eds) 1978: *Obligations to Future Generations*. Philadelphia: Temple University Press.
Simon, Julian L. 1996: *The Ultimate Resource*. Princeton, NJ: Princeton University Press.
Singer, Peter (ed.) 1986: *Applied Ethics*. Oxford: Oxford University Press, 1986.
Sinnott-Armstrong, Walter 2006: *Moral Skepticisms*. Oxford: Oxford University Press.
Slovic, P., B. Fischhoff, and S. Lichtenstein 1978: 'Accident Probabilities and Seat Belt Usage: A Psychological Perspective'. *Accident Analysis and Prevention*, 10, pp. 281–5.
Tännsjö, Torbjörn 2002: 'Why We Ought to Accept the Repugnant Conclusion'. *Utilitas*, 14, pp. 339–59.
Temkin, Larry 1987: 'Intransitivity and the Mere Addition Paradox'. *Philosophy and Public Affairs*, 16, pp. 138–87.
——1996: 'A Continuum Argument for Intransitivity'. *Philosophy and Public Affairs*, 25, pp. 175–210.
Thomson, Judith Jarvis 2001: *Goodness and Advice*. Princeton, NJ: Princeton University Press.

 © Huemer 2008