Future Generations

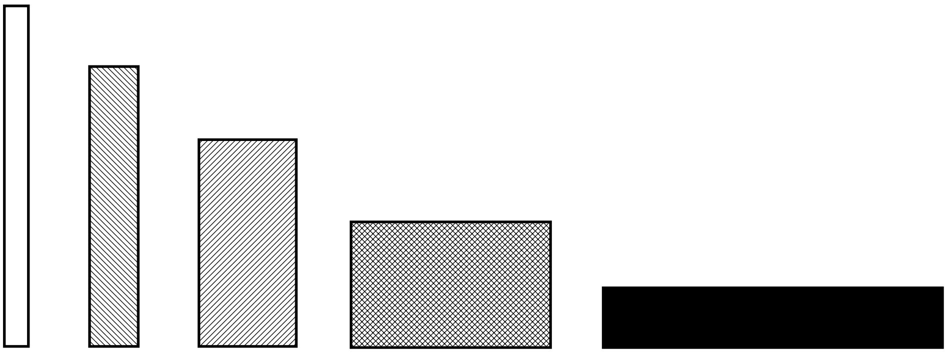
and

Interpersonal Compensations

Moral Aspects of Energy Use

by

Gustaf Arrhenius and Krister Bykvist



Acknowledgements

Several people have helped us to write this essay. Our greatest debt is to Wlodek Rabinowicz, who has been an excellent supervisor of the project. He spent a lot of time and energy reading drafts of the essay. Without his painstaking criticism and helpful comments this essay would lack in precision, relevance, and logical correctness.

Earlier drafts of the essay were discussed in Sven Danielsson and Wlodek Rabinowicz's seminar at the Department of Philosophy, University of Uppsala. The participants of the seminar contributed with helpful criticisms. Apart from Sven and Wlodek, we would like to thank Thomas Anderberg, Erik Carlson, Tomasz Pol, Peter Ryman, Rysiek Sliwinski, and Jan Österberg. We are especially grateful to Erik Carlson. His critical eye detected many flaws in earlier versions of our theory.

A summary of this essay was presented in a seminar at Hässelby Castle, June 1994, organised by NUTEK (Näringsutvecklingsverket), where Bengt Hansson and Lars Ingelstam made interesting comments.

Parts of Chapter 3 were presented at Australian National University, Canberra; Monash University, Melbourne; and Queensland University, Brisbane, February 1994. We would especially like to thank Peter Singer and Yew-Kwang Ng for their helpful suggestions.

Parts of Chapter 4 were presented at University of Toronto, Department of Philosophy, Toronto, May 1994 and 1995, and at the Learned Society, Canadian Philosophical Association, Montreal, June 1995. We are especially grateful to Danny Goldstick, Tom Hurka, Andrew Latus, Howard Sobel and Wayne Sumner for their valuable comments.

Bengt Bykvist, Rebecca Fodden, John Gibson, Wlodek Rabinowicz and Pura Sanchez checked and improved our English. Their advice has been of invaluable help to us.

This essay was sponsored by NUTEK, as part of the research project "Future Generations and Interpersonal Compensations: Moral Aspects of Energy Use." We are very grateful to NUTEK for their generous assistance.

Gustaf Arrhenius and Krister Bykvist
Öland, August 1995.

CONTENTS

INTRODUCTION

PRESUPPOSITIONS AND DEFINITIONS

THE WEIGHT OF EVIL

MORAL DUTIES TO FUTURE GENERATIONS

SUMMARY AND ENERGY APPLICATIONS

Chapter 1

INTRODUCTION

> The long sweep of human history has involved a continuing interaction between peoples' efforts to improve their well-being and the environment's stability to sustain those efforts. Throughout most of that history, the interactions between human development and the environment have been relatively simple and local affairs. But the complexity and scale of those interactions are increasing. What were once local incidents of pollution shared throughout a common watershed or air basin now involve multile nations - witness the concerns for acid desposition in Europe and North America. What were once acute episodes of relatively reversible damage now affect multiple generations - witness the debates over disposal of chemical and radioactive wastes.[1]

It is a truism that choices of energy policies are morally relevant, since almost any choice has morally relevant features. Yet, we might ask whether there are some features that are more disturbing and problematic than others. The overall aim with this study is to identify and clarify these features, features that any *rational* decision maker ought to know. What are these problematic features? A convenient way of presenting them is to let the reader reflect upon the following stories.

1) The Uranium Mining

> Radiation from uranium causes sickness to people living near the uranium mines. They get cancer and suffer a lot. Moreover, other people living near this mine fear that they themselves will become sick. The number of the affected people is small as compared to the number of people benefited by the nuclear power. The consumers of the nuclear power are marginally benefited by the energy produced by the uranium. That is, they would be well-off without it, but they are slightly better-off with it. Due to the consumers' great number their total happiness outweighs the total unhappiness of the minority living near the mine. Does the value of the gains compensate the value of the great losses?

2) The Power Station

[1] Report from Swedish Council for Planning and Coordination of Research (1987), p. 10.

A government plan to build a new power station to produce the energy needed for the society as a whole. For this to be done some people have to move from the area where the station is planned to be situated. These people are highly attached to the area. They and their ancestors have lived here for a long time and many traditions and customs are tied to the geographic area. The only way to get them to move is by force. Thus, if they move, they will be very frustrated. On the other hand, if the station is built, then a lot of other people in other areas will each gain some marginal welfare. These winners are of such a great number that their total happiness will exceed the losers' total unhappiness. Are we justified in building this power station?

3) The Energy Consumption of Present People

If our high energy consumption is maintained, then what we leave to our successors could be environmental pollution, overpopulation, depletion of natural resources, global warming, and nuclear waste dumped on land and at sea. Naturally, these factors are disadvantageous for the people living in the future. On the other hand, our high energy consumption creates welfare here and now. Are we then forbidden to continue to consume energy at the cost of sufferings for future people? Or are the happiness and sufferings in the remote future of less worth than the happiness and sufferings now and in the near future? That is, are we justified in discounting welfare effects in the distant future, at some rate *n* per cent per year?

4) Different Energy Systems - Different People

Our choice of energy systems will affect not just the welfare of future people but also the identity of these people. If not the pair of cells, the ovum and the spermatozoon, that a particular person in fact grew from, had been joined in a conception, then this particular person would never have existed. The choice of energy systems affects, perhaps in a purely accidental way, who has intercourse with whom and when. For instance, a new energy system might create new means of communications. These new means will create new opportunities for people to meet and have intercourse. Assume now that we have to choose between two energy systems that would affect the identity of future people and the welfare of present and future people. More exactly, assume that we have a choice between *A*, the future in which the first system is chosen, and *B*, the future in which the second system is chosen. *A* has better welfare effects in the sense that the future people in A would be much better-off than the future people in B. The future people in *A* are, however, not identical with the future people in *B*. Finally, assume that if *B* is chosen then we, the present people, will be marginally benefited. Do we run the risk of doing anything wrong if we choose *B*? The present people would be benefited and no one would be harmed since for each future person in *B* it holds that she would not exist if we had chosen alternative *A*.

5) The Overcrowded Earth

> We are profiting on the earth's resources at the expense of our successors. This in combination with a steadily increasing population could give us a future where the earth is crowded with people each having a life barely worth living. Assume that we have an opportunity to avoid this overpopulation and to create a world with a much smaller number of people each living a happy life. Although the future lives in the overpopulated world will have low quality, they will be of such a great number that the total happiness in this alternative exceeds the total happiness in the other alternative. Therefore, might it not be claimed that it is better to overpopulate the earth?

One important aim of this study is to formulate an acceptable *principle of beneficence* applicable to the cases above. In Chapter 2 we present the concepts needed for this task. The focus is on questions such as "What kind of evaluation are we after?" and "What do we mean by welfare?"

In Chapter 3, the problems illustrated by examples 1 and 2 are examined and a theory applicable to cases such as these is formulated. We look at cases where we can affect the welfare of presently existing people, concentrating on the problem of compensation. Can the happiness of some people compensate the sufferings of others? Can the happiness of one part of a person's life compensate the sufferings within another part of the same life?

In Chapter 4, the problems illustrated by examples 3, 4 and 5 are examined and a theory applicable to cases such as these is formulated. We look at cases where we can affect the welfare of future people, as well as cases where we can affect the number and the identities of people. These cases raise questions such as: "Should the welfare of future people be discounted?", "How should we evaluate populations with different number of people?", "How should we evaluate populations with the same number but with different persons?"

Finally, in Chapter 5, we summarise the results from chapters 3 and 4 and formulate a general theory of beneficence applicable to all the cases above. With this general theory at hand, we comment on each case 1 to 5.

Chapter 2

PRESUPPOSITIONS AND DEFINITIONS

In this chapter we want to clarify what we mean by axiological expressions such as "x is intrinsically better than y", and how these expressions can be said to have bearing on the question of the normative status of actions. Section 1 deals with the bearers of intrinsic value. In section 2, we explore the nature of the compared entities, give definitions of the comparative value concepts and show the link between axiology and the normative status of actions. Furthermore, section 2 contains a list of the kinds of evaluations we are after.

1. The Concept of Welfare

1.1. A Rational Reconstruction

As was hinted at in chapter 1, we are interested in evaluating alternatives where peoples' welfare varies. Hence, we hold the uncontroversial belief that positive welfare is something good and negative welfare is something bad. But we want to go further and say that the good and the bad are to be *exhaustively* identified with positive and negative welfare, respectively. This is of course controversial, but we shall not defend this position here. Does it follow, then, that the results reached in this essay are irrelevant for a pluralist who believes that other things besides welfare have value? Not necessarily, for if you hold the first uncontroversial belief, then you need some principles stating how to compare alternatives as regards welfare, since in some situations welfare is the only axiologically relevant factor that varies. In addition, one could argue that in order to evaluate alternatives as regards every axiologically relevant factor, you must begin evaluating each factor in turn, and after that make an overall evaluation where each factor is given its proper weight.

Now, we have to explain exactly what we mean and should mean by this axiologically relevant concept of welfare. The aim is here to give a rational reconstruction of the concept of subjective welfare commonly used in our moral practice, but also frequently used by classical utilitarians. We refer to this concept by terms such as "suffering", "displeasure", "frustration" and "unhappiness", on the negative side, and by terms such as "pleasure", "satisfaction" and "happiness", on the positive side. So the explicandum in our reconstruction is this dual concept of welfare. The reason why we want to reconstruct this ordinary concept instead of using a highly technical concept from the beginning is that the intuitions behind the problem with future generations and interpersonal compensations are tied to and therefore easily expressed in terms of this common notion.

To reconstruct a concept is to transform it into a more exact concept, the explicatum, in a way that makes it possible to use this new exact concept in most of the cases in which the explicandum is used. In other words, it must be close to the ordinary usage. This new concept must also be simple and fruitful, the latter taking priority when these demands conflict.[1] The demand for fruitfulness does not say anything without a clear formulation of the problem and an idea of the role we want to give the concept in the problem-solving. Let us therefore present some considerations concerning the role of this concept of welfare.

First of all, we want to make not just *comparative* welfare statements such as "*P* is better-off in one state than in another" but also *categorical* statements such as "*P* has a happy life" or "*P* has an unhappy life." Since we are not intending to exclude sentient animals from the evaluations, "*P*" can stand for a particular sentient animal as well as a particular human being. But this is not enough, because our intuitions concerning the problems in this essay can also be applied to situations in which no one's *life* is being ruined or made worth living. Instead it may be some moments of happiness or unhappiness that are at stake. So we need a categorical concept tied to moments as well as to whole lives.

Before we start our own reconstruction we have to check whether there is any reconstructed concept ready to use. According to a somewhat simplified picture, the discussions about welfare and its value can be localised in three main areas: social choice, economics and moral philosophy.[2] Regarding the first two areas, we have a vast literature on problems concerning the measurement of welfare, and a common standpoint in this discussion is that it is possible to measure welfare in a way that makes it meaningful to say things like "I am happier now then I was before" and "The difference in happiness between the state where I am eating fish and the state where I am eating meat is smaller than the difference in happiness between the state where I am sick in cancer and the state where I am eating fish".[3] Is it possible to meet the demands stated above from this viewpoint? Clearly, we can give meaning to comparative welfare statements such as the statement that a person *P* is worse off in *x* than in *y*. One serious drawback, however, is that on this approach it is difficult to see how we could say that he is happy or unhappy in x, but this is just what we need in order to make an adequate moral judgement about the change from *y* to *x*. For instance, if this change resulted in a lot of small improvements for other people, we need to know how badly off *P* is in *x*, before we can judge it to be a change for the better (all things considered).

Another weakness is that the concept of social state is often used in an "atemporal" sense, with the consequence that an explanation of how the values of the temporal parts in a social state influence the value of the whole is missing. Moreover, the social states are often seen as a part of the total outcome of a social policy, and not as the totality of the future consequences. For a consequentialist this atemporal and restricted use of social state is unsatisfactory. He needs a concept of welfare which

[1] For a brief description of the method of rational reconstruction see for example Alchourrón (1971) pp. 8-9.

[2] We here omit the psychological and sociological studies of the subject. One reason for this is that these disciplines are not interested in the *moral* value of welfare.

[3] If the former kind of statement is meaningful, we can measure welfare on an ordinal scale. If the latter kind of statement is meaningful, we can measure welfare on an interval scale. For more comments on this see section 1.4.

helps him to evaluate the whole outcome, and therefore an analysis of the part-whole relation of welfare is desired.

The traditional moral philosopher often discusses welfare with a more optimistic view regarding its measurability. In a rather naive manner, he talks about "amounts" of suffering and happiness, and sees no dangers in summing up these amounts. Here we get the impression that we can meaningfully make both comparative and categorical statements about welfare, but after a closer look you recognise that you are left without guidance concerning the understanding of this quantitative treatment of welfare.

In the sequel we want to draw an outline of a welfare analysis capable of distinguishing the categorical welfare from the comparative welfare as well as distinguishing the welfare of a life from the welfare of a life period. However, it must be noted that the following is nothing more than a sketchy explanation of the measurability of welfare; a full treatment of this intricate problem requires a much more detailed discussion.

1.2. The Interpretations and Structures of Welfare

In normal conversation, we say things like "I am happy now", and "On the whole, he had a miserable life". Our usage of these expressions seems to show that we think it is meaningful to tie happiness and unhappines to moments as well as to lives. By using the latter expression we do not mean to say that the person was unhappy at every moment in his life; he might very well have been happy at some moments. We want to say that if he had some happiness, this was outweighed by his unhappiness. So, it seems that we think that the welfare value of a life is dependent upon the welfare-values of some parts of the life. In our interpretation of welfare we want to make room for these intuitions, and explain the meaningfulness of these expressions.

Let us call the moments whose values determine the welfare value of the life *welfare moments.*[4] For short, a moment with positive welfare, i.e., a happy moment, is called a *positive moment*, and similarly, we have *negative moments* and *indifferent moments*. Both the positive and the negative moments can be compared within their own category, as is expressed by "I am happier now than I was before", and "I am unhappier now than I was before". An important assumption underlying all argumentation in this essay is that these facts about welfare values are empirical , like, for example, the facts about the length of some objects.

In our dealing with welfare we want, *in principle,* to be open to the following three traditional interpretations:

> (1) *Hedonist welfare*. A welfare moment is a pleasurable, unpleasurable or indifferent experience.[5]

[4] Notice that the welfare value is an empirical property, and hence not identical with intrinsic value.

[5] Here pleasurable experiences are not to be identified with bodily pleasures. And the same holds for unpleasurable experiences and bodily pains. Our interpretation of hedonistic welfare is so wide as to regard both bodily pleasures and intellectual pleasures as genuine examples of this welfare.

(2) *Preferentialist welfare*. A welfare moment is a moment of preference satisfaction or frustration. The indifferent moment may be seen either as a moment of both satisfaction and frustration, the former exactly balancing the latter, or a moment of neither satisfaction nor frustration.

(3) *Objective welfare*.This interpretation of welfare is very different from the previous two, since on this account the welfare of a person is not defined in terms of subjective items such as mental states or satisfactions and frustrations. Instead, certain things are good or bad for us, and this holds even if we at certain times would not want to have the good things or avoid the bad things. These things are thought of as those valued or disapproved of by every person who rationally reflects upon what would make his total life well-lived. The good things might include the development of one's abilities, knowledge, friendship, good health, freedom, dignity; the bad things could be losing liberty or dignity, bad health, sadistic pleasure, being deceived and so forth.[6] A positive moment would on this interpretation be a moment of a person's life in which he is in possession of more good things than bad and a negative moment would be the other way around. For instance, think of a moment of a person's life in which she is well educated, has a stimulating and improving work, many friends, good health and a loving family. Compare that to a moment in which she is unemployed, socially isolated, dependent and poor in health.

We think that our talk about positive moments with different welfare values is meaningful irrespective of the choice of the mentioned interpretations. But the talk of positive, negative and indifferent moments is much easier to make clear given the hedonist or preferentialist account. Just think of the problem of choosing the things that constitute objective welfare. So, we think that it is more convenient to avoid the objective account when discussing our problems in this essay.

Left with the hedonist and the preferentialist alternative we think that the former ought to be chosen. The reason for this is that the preferentialist account is more clearly a family of theories than the hedonist account. Depending on how one answers the following questions, we get different members of the family. Should we count every preference, or should we divide the preferences in different types? In, for example actual versus ideal, personal versus external, malevolent versus benevolent, local (whose objects are small parts of a world) versus global (whose objects are large parts of a world, perhaps the whole world itself), those existing at the time of choice versus those existing in the future (possibly contingent on the choice), first-order versus second order (whose objects are first-order preferences)? Should we exclude from our counting or give lesser weight to satisfactions based on some of these types? A formulation of the most acceptable account of preferentialist welfare would obviously require an essay of its own.

It is important to note that the choice of the hedonist interpretation does not mean that the problems described in this essay are dependent on a hedonist axiology, nor that we think that this interpretation of welfare is the best one. It is for its relative

[6] Aristotle's Nicomachean Ethics is the classical formulation of this view. This concept of welfare also assumes a fundamental role in John Rawls' theory of justice. See Rawls (1972) p. 62.

simplicity that we have chosen the hedonist interpretation. To make this clear we shall, when confronted with a particular problem where one could suspect that a different interpretation could yield a different result, make some comments on where a preferentialist or objective interpretation would lead us.

Let us continue with the analysis of welfare, seen now as a purely hedonist analysis. The welfare moments are here seen as experiences. We have said that welfare moments can be grouped into three mutually exclusive types: positive, negative and indifferent. According to the hedonist approach, this means that each experience is either pleasurable (positive), unpleasurable (negative), or neither pleasurable or unpleasurable (indifferent). These experiences are ordered by the relation "__ is at least as pleasurable as __", where each blank is to be filled in with a name of an experience. Just as we assume that we can numerically represent the relation of weight, "__ at least as heavy as __", that holds between material bodies, we assume that we can numerically represent the welfare relation that holds between experiences. But, of course, we do not think that it is as easy to measure welfare as it is to measure weight. The numbers representing the welfare relation are called utility values, or utility for short. (Sometimes we use "utility" in a more narrow way letting it stand for a positive number assigned a pleasurable experience, while "disutility" stands for a negative number assigned an unpleasurable experience. The context will make this clear.) If an experience is assigned a number greater than zero, then it is a positive moment. The greater positive number an experience is assigned, the more positive, i.e., more pleasurable, the experience is. If the assigned number is less than zero, then it is a negative moment. The greater negative number an experience is assigned, the more negative, i.e., more unpleasurable, the experience is. Finally, if the assigned number is zero, then it is an indifferent moment. Some comments on the meaningfulness of these assignments will be given in section 1.4.

When it comes to individuating the welfare moments, we assume that each moment necessarily belongs to a certain person and a certain time. Furthermore, each moment necessarily has a certain utility. So, *P*'s welfare moment with utility 5 is not the same moment as his moment with utility 6. Finally, we state that one and the same experience can occur in different possible states of affairs. This means that it might have been the case that my actual experience occured in a situation differing from the actual one. For instance, the headache I am suffering from right now when I am wearing my brown jacket might still have occurred in the situation where I left my brown jacket at home. Following the mainstream in philosophy, we talk about these "mights" as "possible worlds", a notion that is similar to common language expressions such as "possible scenarios" and "possible states of affairs".[7]The welfare moments are basic in the sense that they are the smallest utility-carrying units. That is, no fraction of these experiences can be assigned utility. To make things easy, we assume somewhat unrealistically that every moment has the same duration, both intra- and interpersonally. Doing this enables us to say that if two possible experience-streams have the same number of moments, then they have the same duration, and if they have different number of moments, then they have different durations. Thus we need not add anything about the duration of the moments and

[7] More exactly stated, a possible world is a maximal consistent set of states of affairs.

the streams when we evaluate alternatives. Intuitively, the duration of a welfare moment must be rather short, say, no longer than a few seconds.

1.3. Mixed and Pure (Dis)Utility

We have said that the welfare-value of a life is a function of the welfare values of the moments occuring in the life. More exactly, we want to say that the utility of an individual's life is the sum total of the utility or disutility of each welfare moment occuring in the life, i.e., each utility and disutility is given the weight 1 and then summed up. In general, we say that the utility of a compound experience is the sum total of the utilities of all moments occuring in this whole. Note that nothing is here said about the *intrinsic* value of a set of welfare moments. We are still assuming that we are in the realm of empirical facts.

Now, let us reconstruct some commonly used welfare concepts. If the utility of an individual's life is greater than zero, we say that he has *a satisfactory life*. If the utility is zero, we say that he has *an indifferent life*. And finally, if the utility is smaller than zero, we say that he has *an unsatisfactory life*. Similarly, we can say that if the utility of some successive moments in his life is greater than zero, we say that he has *a satisfactory period*. In an analogous way, we can define *an unsatisfactory period* and *an indifferent period* in an individual's life. Note that these concepts are here meant to be purely descriptive and non-evaluative.

Furthermore, an important distinction which bears on the problems raised in this essay, is that between mixed and pure disutility. A life or a period of a life, has *mixed disutility* if it has a total utility smaller than zero, and contains some moment with a utility greater than zero. It has *pure disutility* if the total utility is smaller than zero, and there is no moment in it that has positive utility. In an analogous way, we can define the positive counterparts, *mixed utility*, and *pure utility*.

For simplicity and clarity, we shall restrict the use of "mixed disutility" to *intertemporal* cases. Maybe it is true that we sometimes, at the very same time, have a pleasurable *and* an unpleasurable experience, but apart from the fact that the existence of this schizophrenic experience is something one might be sceptical about, our approach in this essay can, in these special cases, be said to rely on an *overall* judgement. Concerning these mixed experiences we ask whether the experience is more pleasurable than unpleasurable, more unpleasurable than pleasurable, or exactly balanced in the pleasurable and unpleasurable aspects.

1.4. Measurement

It is now time to make explicit some of the assumptions concerning measurement, which underly our characterisation of utility and disutility.

We agree with the prevalent opinion on measurement. That is, measurement is seen as a process of assigning numbers or other mathematical entities to the objects we want to measure, in a way that makes it possible to represent the *qualitative* relations between those objects with *quantitative*, i.e., mathematical, relations between the assigned numbers or entities.[8] In our case the qualitative relation is of course the *welfare relation* interpreted as "__ is as least as pleasurable as __", where each blank is to be filled in with a name of an hedonist moment, i.e., the name of an experience. These experiences are the measured objects.

In the literature on welfare there are three scales that are commonly referred to: the ordinal, interval and ratio scales.[9] On what scale must we measure welfare to be able to make the welfare comparisons that are relevant for our problems? The ordinal scale will not do, because using this scale will only yield a measure which for two experiences *x* and *y* assigns the qualitative relation that *x* has greater welfare, lesser welfare or the same welfare as *y*.[10] And we are then unable to make comparisons of welfare differences. That is, we cannot say that the gap in welfare between the experiences *x* and *y* is greater than the gap between *z* and *w*. To see the importance of this comparison, imagine a situation where *x* has higher utility than *y* and *w* has higher utility than *z*, and you want to know which of the pairs (*x, z)* and (*y, w)* has higher total utility.[11]

We think that the appropriate scale must be the ratio scale. And the argument is as follows. The qualitative structure we want to give a quantitative representation of is best characterised as an *extensive system.* Roughly speaking, the difference between an extensive and an intensive system is that the in the former you have some mode of combination that corresponds to the arithmetical operation of addition.[12] An example of this is entities having length. For instance, combining the elements by placing them end to end on a straight line yields an entity longer than each of the parts. In the case of welfare, an analogy with length will not do, because in the structure of welfare we have objects with zero-value and also objects with negative value. The best analogy here is perhaps a system of weights where you have some peculiar objects with negative weight. Some are without weight, which means that if they are joined to a whole the weight of this whole remains the same, and some objects have negative weight, which means that if they are joined to some whole they make the whole lighter.[13]

[8] For presentations of this view see Coombs (1970) and Roberts (1979).

[9] An *ordinal* scale is a scale which is unique up to an order-preserving transformation, i.e., any transformation of the scale that preserves the order of the scale values yields another admissible scale. So, the admissible transformations are all functions *f* satisfying the condition that x > y iff *f*(x) > *f*(y). Here we can meaningfully compare the order of scale values. An *interval* scale is a scale which is unique up to a positive linear transformation, which means that not only the order of the scale values is preserved but also the order and ratios of differences between scale values . The admissible transformations are all functions of the form $f(x) = \alpha x + \beta$, $\alpha > 0$. Here we can meaningfully compare differences between scale values. A *ratio* scale is a scale which is unique up to a similarity transformation, which means that the ratios of the scale values are preserved. The admissible transformations are all functions of the form $f(x) = \alpha x$, $\alpha > 0$. Here we can meaningfully compare ratios of scale values.

[10] We leave it to the reader to interpret "greater, lesser and the same welfare" in hedonistic terms.

[11] There is a special case where it is possible to compare differences by only using ordinal information. Suppose the utilities of the states could be ranked, from greater to lesser, in the following order: x, w, z, y. Here the difference between x and y must be greater than the difference between w and z.

[12] For a more exact characterisation see Krantz (1971) p. 73.

[13] This analogy is mentioned in Danielsson (1986), p. 53, fn. 6.

In the welfare structure the new entities yielded by the combination are wholes of particular experiences[14], the objects are the experiences, the relation is the welfare relation, the positive value is the positive welfare (pleasure), the negative value is the negative welfare (displeasure), the zero-value is the indifferent welfare, and the combining operation is the construction of possible compound experience-streams such as lives or periods of lives.

To be an extensive system means more specifically that the qualitative relation and the combining operation must satisfy some axioms that are analogous to certain axioms satisfied by the arithmetical operation of addition. For instance, the relation must be commutative, i.e., combining *a* with *b* must have the same value as combining *b* with *a*. It can be shown that if the axioms are satisfied, then we can represent the qualitative structure in such a way that the objects are measured on a ratio scale, and the value of a whole is the sum of the values of its parts.[15] Applied to welfare structures, this means that *if* welfare behaves in accordance with the axioms, then we can meaningfully compare ratios of utility values, and moreover, the utility of a compound experience is equal to the sum of the utilities of the moments occuring in this experience, which was precisely what we assumed in section 3.3. This is a big "if", we admit, but we do not have the space here to argue for the antecedent in this conditional. Though, we find this result well in accordance with the common usage of the concept of hedonist welfare.

So far we have only dealt with the intrapersonal case, i.e., the possibility to measure a person's welfare, and to make ratio comparisons of the utility of different moments belonging to one person. When it comes to the evaluation of different lives, or periods of different lives we have to have some interpersonal standard. More specifically, we want to be able to judge that the utility of *i*'s moment is *n* times as great as the utility of *j*'s moment, where *i* and *j* are different persons. The vast literature on the problem of interpersonal comparisons shows the complexity of this problem, and here we want only to make some brief remarks on the possibility of interpersonal comparisons.

First, interpersonal comparisons is not an all or nothing affair. We can have *partial* ratio comparability, meaning roughly that the interpersonal ratios is not a number, but an interval of numbers.[16] So, we can for example say that, the ratio of *i*'s utility and *j*'s utility is *between n* and *m*. The numbers in this interval are generated from different interpersonal normalisations. The consequence of this is of course that if we have an axiological rule which judges states of affairs according to the aggregated utility or disutility, then with this partial comparability, we run the risk that the rule gives different results depending on which of the numbers between *n* and *m* we pick.

One way out is to postulate the axiological rule that if a world A has at least as much aggregated utility as a world B under *every* acceptable interpersonal normalisation, then A is at least as good as B. This means accepting some lacunas in

[14] These wholes are not identical with individual experience streams, since when we measure the welfare of a group of people we combine experiences from *different* persons.

[15] A presentation of the theorem is found in Krantz (1971) p. 74.

[16] A formal account of partial comparability is given in Sen (1979) pp. 106-111.

the intrinsic value-ordering, but on the other hand we then do not have to lean upon a very unrealistic view on interpersonal comparisons. When, later on, we consider different welfare axiologies, it must be remembered that the axiology in question can always be formulated in this manner.

Second, it is not possible to make a sharp distinction between intra- and interpersonal comparisons. If you think that the possibility of intrapersonal comparisons is due to some similarity between the person-segments, that the states are connected by some psychological relations as for instance memory traits, this seems to make it reasonable to accept interpersonal comparisons. For consider the situation in which you have to judge one future experience as more or less pleasurable than another future experience, and these future person-segments of yours would be very different from the person you are now. If you think that this is still possible, then why not also think it is possible to judge *another* person's experience, if this person is very similar to you? The method in both cases must be some form of empathy and identification, and the relation of similarity seems to hold *between* persons as well as within the same person at different times.

1.5. Welfare Representations

Throughout this essay we invite the reader to employ his or her intuitions with respect to different welfare distributions. That is, we represent the welfare of each person's life or part of life, and ask which distribution is the best. Sometimes this is done by giving numbers, and sometimes we make use of geometrical figures. Now, often it is important to interpret the numbers or the sides of the figures as "small" or "large", and not just relative to the other numbers in the example, but in an absolute sense. What has previously been said about the measurability of welfare does not permit us to make an inference from the proposition that a certain moment has an utility of 100, to the proposition that this moment has a great utility in an absolute sense, for recall that welfare is measured on an ratio scale. Hence any similarity transformation of the number 100 is permitted (See footnote 9). Thus we could equally well have represented the utility of this moment with 0.1, 1, or 10. An analogous case here is measurement of lengths. Lengths are also measurable on a ratio scale, and the length of a certain object can be measured in metres, centimetres and so forth, where each measure is a similarity transformation of the others. Consequently, if we say that an object has the utility 100, this says nothing about how great this utility is. Analogously, if we say that a certain object has the length 100, this says nothing about how "great" this length is, i.e., whether this object is short or long.

How, then, are we to give meaning to the concept of absolute utility size? We think that this can be done in a similar way as is done with the concept of absolute length size. When we say that a particular object is very long, this is always meant in relation to a certain context. For instance, if we say that a tree is tall, this is done having in mind the context of trees or certain kinds of trees. Thus, when a tree is said to be tall, it is said to be tall *compared to other trees*, and not compared to matches. Furthermore, the comparison set of trees is not a set of trees of any conceivable lengths, but it is a set of trees with lengths that have been reported in our world. Hence, the fact that there is a logically possible birch of 100 metres doesn't render the actually existing birch of 10 metres a small birch.

We think that this reasoning can also be applied to absolute utility sizes. Here the comparison set consists of moments or lifes with utility. When we say that a moment has great (small) utility, this is meant to imply that this moment has a utility that is greater (smaller) than most of the other reported moment utilities in our world. And when we say that a life is very satisfactory (very unsatisfactory), this is meant to imply that this life has a total utility (disutility) that is greater than most of the other reported lifetime utilities in our world. The same holds when we speak about the utility of parts of a life, whole lives, populations, and so forth. In an analogous way, we interpret great losses and gains of utility (disutility).

Notice, then, that when we represent the welfare distributions with numbers or geometrical figures, the absolute sizes of the numbers or of the figure sides, are quite arbitrarily chosen. To interpret the examples correctly you must observe whether the chosen number representing the welfare is regarded as a great or a small utility (disutility).

2. Axiological comparisons

2.1. Value Concepts

The comparative value-concepts used in this essay are defined with the relation "at least as intrinsically good as" as primitive, where "intrinsically good" should be read as "good in and of itself" (henceforth the attribute "intrinsically" is omitted.) So, *x* is *better* than *y* if and only if *x* is at least as good as *y* and it is not the case that *y* is at least as good as *x*. *x* is *equally as good as y* if and only if *x* is at least as good as *y* and *y* is at least as good as *x*. With these definitions at hand we are not committed to full comparability as we would have been if we defined "*x* is indifferent to *y*" as "it is not the case that *x* is better than *y* and it is not the case that *y* is better than *x*".

We assume that "at least as good as" satisfies the usual conditions. Thus, for all *x*, *x* is as least as good as *x* (reflexivity), and for all *x*, *y*, and *z*, if *x* is as least as good as *y*, and *y* is as least as good as *z*, then *x* is at least as good as *z* (transitivity).

The best element in a comparison set is that which is better than every other element in the set. *A best element* in a set is an element at least as good as every other element in the set. Finally, *a maximal element* is an element such that there is no element in the set which is better than this element. So, even if we lack full comparability we can sometimes pick out maximal elements. If, for example, the value structure is: x and *y* are incomparable but both are better than *z*, we can pick out the maximal elements *x* and *y* from the set (*x*, *y*, *z*)[17]

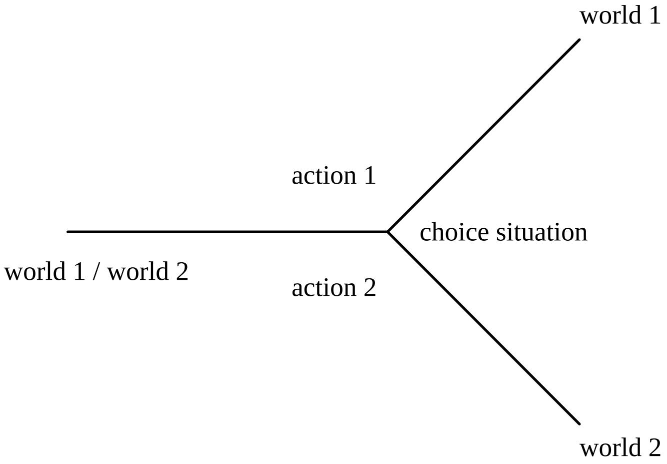
2.2. Compared States

Generally speaking, we see the compared elements as states of affairs. These states are either *whole* possible worlds, as when we say that this world is better than that world, or *parts* of possible worlds, as when we say that this year was better than the one before, or that this day is better than the one before.

Because of our consequentialist inclinations, we think it is convenient to see the whole possible worlds as consequences of actions. Depending on which action we choose, different worlds will be realised. This typical choice situation can be represented as a set of branching possible worlds unified by a common node in a world tree. The following picture represents a choice situation with two alternative actions.

[17] These definitions are found in Sen (1979) p.11.

A branching world tree



The expression "world1/world2" refers to the fact that before the time of choice world 1 and world 2 are identical. The picture is supposed to capture the choice situation where if action 1 would be performed then world 1 would be realised, and if action 2 would be performed then world 2 would be realised.

We simplify the discussion by assuming that the performance of an action has a *unique* possible world tied to it. Furthermore, it is assumed that when we judge an action, we know which unique world is tied to the action, thereby avoiding every probabilistic consideration, and restricting our axiological comparisons to "sure" consequences. When it comes to application to energy problems, something will be said about risks and chances.

The axiological comparisons between the possible worlds could be called *inter-world comparisons,* in contrast to comparisons between parts of one and the same world, which we could call *intra-world comparisons.* A second possible case of inter-world comparison is one between a part in one world and a part in another world. When we in the following compare one *alternative* with another, it is presupposed that we are making an inter-world comparison. When we compare one *state* with another, it is on the other hand indeterminate what comparison we are making, but if nothing explicit is said the comparison is assumed to be valid in both inter- and intra-world cases.

When treating the problems raised in this essay it is important to distinguish the compared alternatives according to the identity and the number of persons. The following three kinds of choice are especially important: *same people choices*, *same number choices* and *different number choices*.[18] The first kind of choice does not affect the number of persons, nor the identity of persons. Major social decisions that affect the welfare of future generations are also likely to affect who will exist. Same number and different number choices are of this type, where the former affect the identities but not the number of future people, while the latter affect the number, and hence also the identities, of future people.

[18] See Parfit (1984), p. 356, and Chapter 4, section 1.

Similar distinctions could be made as regards *moments*. We could then for each person say whether the alternatives contain the same number of moments or not. If for every person the alternatives have the same number we say that we have a *same number of moments case*, and if not we say that we have a *different number of moments case*.

2.3. What Kind of Examples are Relevant?

As we proceed with our investigation of different moral principles, we shall come across different cases that we use as a "test" for these principles. These are cases that we have firm beliefs about and we can test a principle by checking whether it complies with our considered beliefs in these cases. We shall collect such cases and these will be our "conditions of acceptability" with which every principle of beneficence must comply.[19]

Most often, this kind of testing has to rely on more or less hypothetical cases rather than actual ones. One could object that such examples are unreal or artificial and therefore should not have any implications for our moral beliefs. This kind of objection can take two forms:

> *The Impossibility of Imagination Objection:* There are (hypothetical) cases that are so far apart from our daily experience that we cannot imagine what such cases would involve, or we are bound to be very unsure of what such cases would involve. Therefore, our intuitions will be unreliable.
>
> *The Actual World Objection:* Moral principles only need to solve problems that can occur in our world with its natural laws and with the kinds of beings that actually inhabit it.[20]

Take the example of Nozick's "utility monster"[21] who has a quality of life that is millions of times higher than anybody in this world. Can we imagine what it would be like to be such a monster? Most probably not, and that is, in line with the Impossibility of Imagination Objection above, a reason for not testing our axiological and normative principles on cases involving this kind of being.

A case that we are going to discuss in Chapter 4, the Repugnant Conclusion, consists in two alternative outcomes. In one of them we have a population of five billion people with good quality of life. In the other we have a huge population of many hundred billion of people with very low quality of life. Could one not argue

[19] The use of hypothetical cases in philosophy in general and ethics in particular is probably as old as philosophy itself. Consider the following famous passage from Plato's *Republic* (p. 331cd): "'What you say is very fine indeed, Cephalus,' I said. 'But as to this very thing, justice, shall we so simply assert that it is the truth and giving back what a man has taken from another, or is to do these very things sometimes just and sometimes unjust? Take this case as an example of what I mean: everyone would surely say that if a man takes weapons from a friend when the latter is of sound mind, and the friend demands them back when he is mad, one shouldn't give back such things, and the man who gave them back would not be just, and moreover, one should not be willing to tell someone in this state the whole truth.' / 'What you say is right,' he said. / 'Then this isn't the definition of justice, speaking the truth and giving back what one takes.'" Bloom (1968).

[20] See Hare (1981), pp. 5, 113-116, 194-96, 47-49.

[21] Nozick, (1974), p. 41.

that we cannot assess the virtues and vices of this latter alternative? No, we can imagine how the lives of these people would be. It could, for example, be like the lives led by unemployed people in Europe.

Another problem with the "utility monster" case is that its existence could be said to involve a factual impossibility, that is, it requires a major change in the laws of nature, including the laws of human nature.[22] A common way of defending the use of hypothetical cases in ethics is to say that the artificiality involved is of the same kind that has been fruitful in other fields of inquiry, such as, natural science. Scientific experiments are set up so that the influence of all factors not being studied is, as much as possible, "artificially" eliminated. But natural sciences need not account for factual impossibilities in their theories and it could be said to be quite peculiar that moral theories have to stand up to higher demands than natural science. So, in line with the Actual World Objection above, we should restrict the domain of test cases to possible worlds where there are no changes in natural laws, laws of human nature, and so forth. This criterion will not create a sharp and indisputable border, e.g., there is much dispute about how plastic the human nature is. It would, however, dismiss examples such as Nozick's "utility monster."

In practice we cannot, because of the finite stock of resources, produce an enormous population of people whose lives are of very low quality. This is a *technical impossibility*. By adding some assumptions about the availability of resources or about new inventions that make it possible to use material of no worth today, we could suppose it to be possible. What is technically impossible today may not be so tomorrow. Perhaps one could argue that principles should only be accurate in technically possible worlds, but one would then be on a very shaky ground. We can be more sure that the laws of nature will not change drastically over the next centuries than we can be about what is technically impossible in the future. Indeed, the border between technical and factual impossibilities is not sharp - is a population of one thousand billion people on the earth only technically impossible? The answer to such questions depends of what faith we put in different sciences. That the distinction between two concepts has a vague boundary is not an argument against its usefulness. We do not want to grant Sextus Empirikus the argument that incest is not immoral, on the ground that touching your mother's big toe with your little finger is not immoral, and all the rest differs only by degree. Almost all predicates in natural language are vague but they are usable provided they have clear cases and clear counter-cases. Nozick's "utility" monster is a clear case of a factual impossibility; a population of a hundred billions is a clear technical impossibility today, but no factual impossibility.

We could rank these different kinds of impossibilities. It does not matter much whether a principle has strange implications in cases involving factual impossibilities. It should handle cases involving technical impossibilities, but we have to be careful with the more extreme technical impossibilities, which might involve factual impossibilities. It is better if a principle can handle these extreme technical impossibilities, but if the opposite is true, this does not always mean that

[22] Cf. Parfit (1984), pp. 388-389. Parfit makes the distinction between "deep" and "technical impossibilities". What he calls a "deep impossibility" is similar to what we call a" factual impossibility".

we should reject it. Furthermore, we should be careful with examples that are so artificial that we may have problems with knowing what such cases really involves.

Call a possible world that is compatible with the laws of logic a *logically possible world;* such a world may involve factual and technical impossibilities. A world that is compatible with both the laws of logic and natural science we call a *factually possible world;* such a world may involve technical impossibilities.

The way of representing problematic cases with numbers and diagrams in the following needs a comment. The numbers and the diagrams give information about the utilities and disutilities of whole lives or parts of lives. This information describes an abstract case. When we picture a special choice situation by giving a short story about how the alternative worlds come to differ in welfare, we present a *specification* of an abstract case. For the same abstract case many different specifications can be given. If we want to have something for our axiological principles to work upon we must present an abstract case, but not necessary a specification. The relevance of a specification partly lies in the fact that one sometimes may wonder whether the case represents more than just a merely *logical* possibility. By giving a specification we show that the situation is not just logically possible but also factually possible. One might also argue that by making things more concrete it is easier to see what practical implications can be drawn from the discussion. It is easier to see the similarities between a constructed case and real cases if we give a not too unrealistic illustration of the constructed case.

Although all the cases presented in this study are factually possible many fail to be likely cases, i.e., cases that we *often* find in actual choice situations. This failure is unavoidable if we want to purify the cases so that our intuitions can give clear answers. Besides, if we instead were to give very realistic examples, i.e., real policy choice situations, they would probably involve so many people in so many different welfare positions that all simplicity and transparency would be lost.

To sum up: The examples in this essay are not just merely conceivable, but satisfy the stronger requirement that our present knowledge does not show them to be factually impossible.[23] We shall formulate the conditions of acceptability in terms of cases which can be given factually possible specifications. Consequently, we avoid both the Impossibility of Imagination Objection and the Actual World Objection.

2.4. Welfarism

To determine the intrinsic value of a world or a part of a world we must focus on the welfare content. But nothing has so far been said about how the values of the worlds are dependent on these welfare moments. Welfarism gives a hint how this dependence is to be understood, and constitutes the point of departure for the axiological considerations in this essay. A rough formulation of welfarism could be rendered as follows.

> If two worlds differ with respect to their intrinsic values, then these worlds differ with respect to welfare moments.

[23] See Glover (1977), pp. 33-35 for a similar view.

So, if two worlds have different intrinsic value then there is at least one welfare moment in one world that does not obtain in the other. This is of course a very weak assumption, which constitutes more of an axiological skeleton than a full-fledged axiology. But welfarism gives us an axiological frame for our study. *How* the intrinsic value of a welfare set is computed is still undetermined. This is as it should be because how the value is to be determined is exactly the problem we want to solve.

2.5. Normative Implications

We have already admitted that we are consequentialist in spirit. More specifically, we are in this essay restricted to the act-oriented version, according to which the normative status of an act is entirely determined by the consequences of this single act. This can be contrasted with indirect consequentialisms, where the normative status of an act is (partly) determined by consequences of other acts, as in rule-consequentialism where an act is obligatory if it can be subsumed under a rule whose general acceptance gives the best result.

Our choice of act-consequentialism is partly due to its theoretical simplicity and clarity. We do not think that the solution to our problem depends on choosing a specific consequentialism.

To make things clear we shall give the act-consequentialistic definitions of some common normative concepts. An action is right if and only if there is no alternative action with better consequences. An action is wrong if and only if it is not right. An action is obligatory or ought to be done if and only if it has better consequences than every alternative action. In other words, the consequences of the action constitute the best element.

It must be noticed that whenever we talk about the normative implications of a specific axiology, as when we say that from the axiology in question it follows that you ought to perform an action, the gap between axiology and normativity is supposed to be filled by act consequentialism.

Chapter 3

THE WEIGHT OF EVIL

1. Introduction

> The worst in life, the lot of the completely unhappy people, the ceaseless infernal suffering, the hopeless degradation, a child slowly dying in pain - I cannot see that all beauty in the world or even the most extraordinary thoughts can 'outweigh' this, and neither can other people's happiness and culture.[1]

> Imagine that you are creating a fabric of human destiny with the object of making men happy in the end, giving them peace and rest at last, but that it was essential and inevitable to torture to death only one tiny creature - that baby beating its breast with its fist, for instance - and to found that edifice on its unavenged tears, would you consent to be the architect of those conditions? Tell me, and tell the truth.
> No, I wouldn't consent, said Alyosha softly.[2]

> We should surely not want to subject one individual to unspeakable suffering to give some insignificantly small benefit to many others (even an innumerable myriad of them).[3]

Many of us believe that something reasonable and important is said in these quotations. Perhaps we are not willing to go as far as to claim that some suffering can never be compensated, but we believe that unhappiness and suffering have greater weight than happiness. That is, expressed in our terminology, we believe that disutility has greater weight than utility.

The overall aim with this part of our essay is to give an account of this weight, which means that we shall try to formulate a welfarist act-consequentialism that takes seriously the weight of disutility. In other words, we are looking for an acceptable negativist utilitarianism. With this theory at hand we hope to find guidance when it comes to evaluating the energy problems presented in Chapter 1, especially examples 1 and 2. For surely, if we want to know whether we are justified in forcing people to move from an area, or letting people become ill for the sake of the welfare for others, we must provide a general analysis of the weight of disutility.

In order to reach the most acceptable form of negativism we shall evaluate some significant variants of negativism. As the analysis proceeds, we shall list some important disadvantages with the presented negativisms. These disadvantages can be used when stating conditions of acceptability. That is, every acceptable negativism

[1] Ingemar Hedenius in Bergström (1984) p. 125. Our own translation from Swedish.

[2] Dostojevsky's *The Brothers Karamazov*, bk V, Ch. IV, p. 258, (1923).

[3] Rescher (1966) p. 29.

must lack these disadvantages. Finally, in section 8, we shall ask the key question whether it is possible to formulate a negativism satisfying all the conditions.

The disadvantages are cases where the analysed negativism gives the intuitively wrong answer. In other words, we presuppose that we have some pretheoretical considered judgement about these cases, and when the negativism and our intuition differ the fault is presumed to lie with the negativism. One can say that we assume that the judgements about these cases are the "facts" that the negativist principles should match. Since so much depends on these cases it is important to formulate them as clearly as possible. We do not want to have cases so vaguely and ambiguously drawn that our considered intuitions have nothing to work upon. Viewed from one side perhaps one alternative is definitely better than another, but viewed from another side the judgement is either no longer sure or completely altered. The possible interpretations of a case and its alternatives must be restricted. In order to comply with this demand the following points must be made.

Firstly, it must be remembered that in this chapter the focus is entirely on *same people choices*, especially on cases in which it is only the welfare of presently existing persons that is affected.[4] The problems about how to evaluate same number of people choices and different number of people choices are left to Chapter 4. Furthermore, we shall not, here in Chapter 3, ask whether the affected people fall into different morally relevant populations. This problem is also left to Chapter 4, (see especially sections 3.1 and 6.5). In the present chapter, we shall for simplicity assume that each world has only one morally relevant population: the set of all inhabitants of the world.

Secondly, to clarify in what welfare respects two compared worlds differ we shall introduce some concepts of invariance tied to negative, positive and indifferent moments, respectively. Two worlds w_1 and w_2 are *negatively invariant* for an individual i (who exists in both worlds) if and only if (iff), for any negative moment m belonging to i, m obtains in w_1 iff m obtains in w_2. The worlds are *positively invariant* for i if, for any positive moment m belonging to i, m obtains in w_1 iff m obtains in w_2. Finally, the worlds are *indifferently invariant* for i if, for any indifferent moment belonging to i, m obtains in w_1 iff m obtains in w_2. The concepts negative, positive and indifferent variance are yielded by negating the corresponding invariances. Obviously, if two worlds are positively, negatively and indifferently invariant for i, then i's sets of moments are identical. Notice that if a person lacks positive (negative) (indifferent) moments in two worlds, then the worlds are positively (negatively) (indifferently) invariant for her.

Thirdly, we shall specify whether the discussed case is a *different number* or a *same number of moments case,* (for the definitions of these concepts see Chapter 2, section 2.2.) Information about moment variance and invariance is insufficient when it comes to evaluation. For suppose that, for a certain individual i, the worlds A and B are negatively and indifferently invariant, but positively variant, and i's lifetime utility in A is higher than i's lifetime utility in B. Is this information sufficient for a proper evaluation of A and B? Is it evident that A is better than B? Not at all, for it

[4] There is one exception to this. When we discuss the Elimination Argument in section 4.2 we consider a version of this argument that is formulated in terms of different number cases.

might be the case that the life in *A* only consists of dull moments (whose utility is positive but close to zero), whereas the life in *B* consists of intense pleasures. Due to the greater number of positive moments in *A* the life in *A* has greater utility than the life in *B*. Hence, information about the number of moments is important.

Finally, to avoid confusions about moral weight some comments are needed. It is important not to confuse the concept of *intrinsic* weight used in this essay with other morally related weight concepts. For example, this essay does not deal with the epistemic or deliberative weight of disutility. To say that disutility has epistemic weight means roughly that we can with greater epistemic security decide what is evil than what is good. Is there any doubt that suffering is bad (but not necessarily the only bad thing)? In contrast, when it comes to the question whether happiness is good, many of us are more sceptical. Distinguish this from the view that disutility has greater deliberative weight, here meaning that when confronted with an actual choice situation we often can make judgements about what would harm people with greater confidence than about what would make them happy. This is not the same as the former concept because we can have the situation where we with equal degree of confidence decide that disutility is evil and that utility is good. But at the same time we may in many choice situations lack the information needed for deciding what would make people happy. That disutility has this deliberative weight motivates some total utilitarians to interpret the principle of minimising disutility as a rule of thumb, compliance with which is believed to maximise total utility in most cases. That is, when we do not have the time to do the full utility calculus we ought to follow the minimising rule. And this obligation holds not on a negativist ground but on a pure classical utilitarian one.[5]

2. The Drawbacks of Total Utilitarianism

Example 1 in Chapter 1, "The Uranium Mining, clearly shows why total utilitarianism does not give an adequate weight to disutility, and therefore why a search for a negativist alternative is justified. It is a paradigm of a problematic case, and consequently it should be used to formulate one important condition of acceptability. But to see this more clearly we have to spell out the story in example 1.

As described in example 1 people living nearby the mine "suffered a lot" due to the diseases they got from the radiation. Assume that these sufferings were so great that they made each life unsatisfactory on the whole. Further assume, as hinted in the example, that each of those benefited by the nuclear power would have had good lives anyhow. So, the question is whether to make many people's lives, which would be good anyhow, slightly more well-off at the cost of ruining some persons' lives.[6]

For the classical utilitarian the value of the benefits compensates the value of the burdens just in case the benefits *factually* outweigh the burdens, i.e., just in case

[5] A classical utilitarian arguing for this interpretation is J.C.C. Smart. See Smart (1973) p. 29.

[6] The wrongness in making people with satisfactory lives even more satisfied at the cost of one person's severe suffering is the spirit in the principle of unacceptable trade-offs proposed by Ragnar Ohlsson. For a concise formulation see Ohlsson (1979) p. 76.

the sum of the utility differences is greater than zero.[7] If those benefited are of a great number then, although each benefit is marginal, the sum of these small benefits will factually outweigh the great sufferings. This utilitarian conclusion could be generalised and equally well applied to cases where the losers are undergoing the most infernal sufferings, while the winners' gains are just noticeable. If the number of winners is sufficiently high to make the sum of their small gains outweigh the great losses, then the value of the gains compensates the value of the losses. (Notice also that, provided that the compared alternatives have the same number of persons, this holds even for the average utilitarianism, where the aim is to maximise total utility per capita.)

Now, this consequence of utilitarianism may be stated roughly as follows. Irrespective of how many persons that are worse off in one alternative *A* as compared to another *B*, each having an unsatisfactory life in *A*, and irrespective of how unsatisfactory each such life is in *A*, their losses can always be compensated by making persons that are well-off in both alternatives better-off in *A*. But this statement is not clear. Some comments are needed.

First of all, to make things easier to grasp we assume that we have a same number of moment case for each person.

Secondly, gains and benefits come in different types. Gains can be mixed or pure. Mixed gains consist of gains and losses where the gains factually outweighs the losses. Pure gains consist of gains only. In a similar way we have mixed and pure losses. To avoid begging questions concerning *intra*personal compensations, we shall formulate the condition of acceptability tied to *inter*personal compensation in terms of pure gains and losses for each person.

Although we focus on pure gains and losses, we still have different interpretations of these gains and losses. For depending on what type of welfare variance that holds for a person we get different types of pure gains and losses. For instance, take the pure gain or benefit. If a person is benefited in *A* compared to *B*, then the total utility of his life in *A* is greater than the total utility of his life in *B*. But this does not tell us whether the alternatives are positively variant, (i.e., he has more pure positive utility in *A*), negatively variant, (i.e., he has less pure disutility in *A*), or both. We think that the absurdity of the utilitarian view on interpersonal compensation in the example above holds irrespective of how we interpret pure gains and losses.

Finally, "can" has different meanings in different contexts. The sense of "can" we are intending is best seen by stating the utilitarian consequence more precisely. The conclusion that every acceptable negativist theory must avoid is this:

The Interpersonal Absurdity

[7] Note the difference between value compensation and factual outweighing. The relata of the latter relation are utilities, while the relata of the former relation are intrinsic values.

> For any number of pure losses, each making a life unsatisfactory or more unsatisfactory, and for any size and type of these losses, and for any size of pure gains for persons that would have good lives anyhow, there is a number n and a type of pure gain such that if n is the number of gains of this type and size, then the value of the gains strongly compensates the value of the losses.

That the value of some gains strongly compensates the value of some losses just means that the gains have positive value and the losses have negative value, and the whole constituted by the losses and the gains has positive value.[8] The axiological relevance of this conclusion is obvious. For if we compare two alternative worlds and find that the value of the losses is strongly compensated by the value of the gains, then we can also say that one of the worlds is better than the other. For instance, if *G* is the gain from world *A* to world *B*, and *L* is the loss from *A* to *B*, then if the value of *G* strongly compensates the value of *L*, *A* is better than *B*.

We use the terms "good" and " well-off" to indicate not just that the lives are satisfactory, i.e., that utility plus disutility yields a positive sum, but also that the lives do not contain a lot of negative moments with great disutility.[9] This gives us a more clear-cut condition of acceptability, since the most problematic cases of compensation are cases where uniformly happy lives stand against unsatisfactory ones.

One could argue against using this condition as a condition of acceptability by saying that there is nothing strange with a principle that implies that, for any size of *negative* benefit, (i.e., a change from more to less disutility), the value of this type of benefit can always compensate the value of losses. But in response to this, imagine that the disutility in question is just noticeable and that each winner lives in heavenly delight except for this insignificant disutility. Would we then be prepared to say that for any size of losses, if the number of the negative benefits is sufficiently great then the value of the benefits compensates the value of the losses? Our answer is No, and therefore we formulated the condition of acceptability without paying any special attention to negative benefits.

Hitherto we have been discussing the absurdity of the utilitarian view on *inter*personal compensation. But could one not argue that the utilitarian view on *intra*personal compensation is also problematic, although arguably trade-offs between lives are more problematic than trade-offs within lives? We want to claim that irrespective of whether we are comparing alternatives where the gains and the losses come in one and the same person or alternatives where the gains and the losses come in different persons, the utilitarian view on compensation is absurd.

Consider the following case. Assume that there is some medicine that if taken daily in the childhood would make the adult part of our life happier than it would have been without the medicine. The medicine would make us just noticeably better-off at each moment of our adult part of life. The problem is that this medicine causes great sufferings when taken, and this sufferings are so great that they would totally

[8] *Strong* compensation, should be distinguished from *weak* compensation, which is the case when the value of the gains exactly balances the value of the losses .

[9]For a more precise statement of a "good" life see Ch. 3, section 8.2.

ruin our childhood and make it wholly unsatisfactory. Assume furthermore that our adult part of life would be good anyhow, (again, "good" is not synonymous with "satisfactory", see above.) Is it better to take the medicine if the gains for us as adults outweigh the losses for us as children? The utilitarian answer is Yes, since this situation mirrors in all relevant aspects the former interpersonal one. We agree with the utilitarians in that this situation and the former one are similar in the relevant aspects. But for us this is precisely the reason why we should reject taking the medicine. Hence we state the following Intrapersonal Absurdity.

The Intrapersonal Absurdity

> For any number of pure losses, each making some of a person's stages unsatisfactory or more unsatisfactory, and for any size and type of these losses, and for any size of pure gains for other stages of the same person that would be good anyhow, there is a number n and a type of gain such that if n is the number of gains of this type and size, then the value of the gains strongly compensates the value of the losses.

The Intrapersonal Absurdity is related to cases where the pure gain for one person-stage stands against the pure loss of another person-stage in the same person. Unfortunately, the concept of a person-stage is loose. But we think that irrespective of how we make this concept precise the Intrapersonal Absurdity should be avoided. We will say somewhat more about this in section 8.1.

Some authors would argue that in taking this view on intrapersonal compensations we are presupposing a particular concept of person. They claim that the concept of person and personal identity gives, or at least makes more plausible, a specific answer to the question about the intrapersonal compensation. Without getting too deep into the perplexing problems of personal identity, we think that the different views on the concept of person can roughly be divided into two distinct families: one viewing persons as *endurers* and the other viewing persons as *perdurers*.[10] To view persons as endurers means that one view them as beings who *move* in time, and about whom it consequently makes sense to say that they are wholly present at each time. And two beings are identical if they have the same substance. So, when we say that someone suffers at a time we do not mean that it is a temporal part of the being that suffers, but the whole being at that time.[11] To view persons as perdurers is to view them as four-dimensional objects *stretching* over time. And two temporal parts are parts of the same being if some special psychological or physical relations hold between them. So, when we say that someone suffers at a time this is here to be understood as saying that the being is just partly present at that time, and it is a temporal part of that being who suffers. (Notice that in defining welfare moments we have not taken a stand in this dispute. We are open for defining '*P*' and 'person' either as referring to a certain temporal part of a perduring person or as referring to a whole enduring person present on a certain time.)

With this distinction at hand the writers claim that if we view persons as perdurers then it is more plausible to consider experiences at a time rather than whole persons who have them as the proper moral units.[12] And this implies that intrapersonal compensation is very similar to interpersonal compensation, since the temporal parts of one and the same person are so weakly connected. Perhaps some of *my* temporal parts bear more resemblance to *another* person's part than a future part of mine.

[10] This distinction can be found in writings of Quine and D. Lewis. See also Noonan (1989)

[11] Of course, on this view it is not denied that we can properly say that the suffering obtains in a part of the being's *life*. But, again, it is the *whole* being living his life day after day.

[12] See Parfit (1984) pp. 336-347, and Haksar (1991) p 246.

We think that it is a mistake to hold that we first must decide how to view persons before we state conditions of acceptability. Even if you think that it is the suffering of temporal parts that matters and we think it is the suffering of whole persons that matters, we may despite this agree on how alternative worlds ought to be evaluated. The reason is simply that it is far from clear what is meant by caring about whole persons. Why should this care preclude that we care about the welfare a whole person has at different times, and give different weight to these welfares? Could not our concern about whole persons forbid us to sacrifice the welfare of this person at a time for the welfare of this same person at another time, as is presupposed in the discussion about the Intrapersonal Absurdity? The part-caring perdurer-theorist may perfectly agree with us in this constraint on intrapersonal compensation but give it a somewhat different description: we are not allowed to sacrifice the welfare of one temporal part for the welfare of another temporal part of the same person.

To sum up. One of the greatest drawbacks of total utilitarianism is its implication concerning compensation. Utilitarianism implies both the interpersonal and the Intrapersonal Absurdity. (To save words we will in the following call the conjunction of these conclusions *The Absurdity*.) The problem with the total utilitarianism lies in the sum-ranking assumption according to which a world with higher total utility is better than a world with a lesser total. Ruled by this assumption total utilitarians are prevented to pay any attention to how an aggregate of utility is made up from individual utilities and disutilities. Or, as Parfit describes utilitarianism:

> When we choose between social policies, we need to be concerned only with how great the benefits and burdens will be. Where they come, whether in space, or in time, or as between people, has in itself no importance.[13]

3. Alternative Approaches

As Chapter 2 shows our approach can be summarised as a welfarist act consequentialist position. To this we added the assumption that our moral information is restricted to the description of welfare moments (for simplicity interpreted as hedonist moments). The relevant information is restricted to facts about the owners, the timings, and the intensities of different welfare moments. No information about the persons' motivations or the sources of their pleasures and displeasures is given. An important question to answer before we start to deal with the problems is whether we miss the optimal solution by restricting ourselves to this special approach. That is: are we forced to deny either this neutral welfarism or consequentialism in order to reach the optimal solution? Let us first consider the welfarist component.

A welfarist need not be neutral. He might, for instance, give less or no weight to pleasures with bad sources. One possibility here is to disqualify utility stemming from malevolent desires like sadism in order to forbid some problematic

[13] Parfit (1984) p. 340.

interpersonal compensations. A common problem with total utilitarianism is that according to this theory it seems possible to judge the Roman gladiator games as a good institution. If the utilities of the laughing sadistic public outweigh the disutilities of the suffering gladiators, and other things are equal, then the institution is good. This is of course a horrendous implication of total utilitarianism, but disqualifying malevolent utilities gives us just a partial solution to the compensation problem. We can easily imagine a situation where the utilities and the disutilities all are non-malevolent. For instance, imagine a scientist who has produced a medicine that will cure some widespread but minor inconvenient disease like cold. The only way to produce this medicine was to make the subject of experiment suffer a lot. Hence the scientist will make a lot of people slightly better-off at the cost of ruining the subject's life. Suppose that the scientist was motivated by a desire to benefit a lot of people, and not by a desire to torture the subject of experiment. Then there were no malevolent desires to discount in this case. But still we think that the compensation is problematic, especially if those who suffered from cold would have been well-off anyhow.

Some writers seem to hold that the compensation problem can only be fully captured in a non-welfarist and non-consequentialist framework where the focus is on a special category of *wrong-doings*, instead of states of suffering. Sometimes this category of forbidden actions is supposed to comprise actions in which the agent treats other persons merely as means, but sometimes the category is just a list of bad actions such as torture, killing without reason, lying and so forth. We could perhaps in line with Nozick try to formulate a constraint-based morality where these wrong-doings constitute constraints on action. That is, these wrong-doings and their consequences are ruled out as alternatives to consider in the choice situation.[14] But the problem here is, as Nozick himself seems to realise, that in some tragic situations the only way to prevent a catastrophe is to perform one of these forbidden actions.[15] Imagine, for example, that the only way to have prevented the killing and torturing of the Jews in the second world war would have been to torture a couple of Nazis. This deontological approach seems then to be prevented to give a satisfactory weight to disutility, when we have to choose between different horrendous sufferings.

One might object here that a constraint-based morality can formulate constraints that are context dependent. An action type that violates the constraints under normal conditions may not violate the constraints under catastrophic conditions. Thus, for example, under normal conditions, torturing someone violates the constraints, but torturing someone to avoid the torturing of millions may not violate the constraint. But here the constraint-based morality and the welfare-based morality seem to meet similar problems, since one reason why it is sometimes right to torture one to save others from torture is that we thereby avoid great sufferings. Then, both theories must decide what kind of sufferings can be compensated by what kind of benefits. We see no reason why it would be easier for the constraint-based morality to answer this question.

[14] Scheffler (1988) pp. 134-141.

[15] Ibid., p. 137 footnote 5.

Among the non-consequentialist writers it is common to criticise total utilitarianism on the ground that this theory does not take seriously the "distinction between persons".[16] Sometimes this distinction is meant to imply that *interpersonal* comparisons of welfare are morally impossible: one person's disutility cannot be outweighed by another person's positive utility. In contrast, these writers see nothing strange in intrapersonal compensation. Might it not be possible to solve the problem with the weight of disutility just by forbidding interpersonal compensation? The problematic cases listed in this study are often cases where one person's disutility stands against other persons' utility or disutility. So, why not just forbid these trade-offs between different persons?

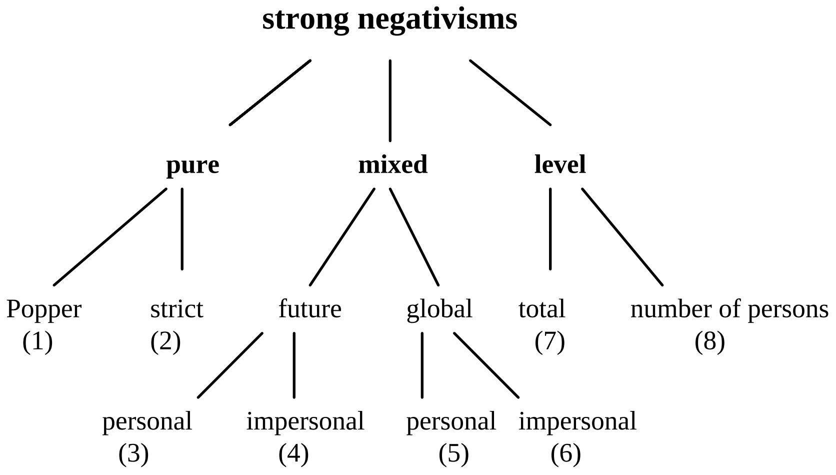
This approach is problematic in two respects. First, banning every interpersonal compensation is too strong. It bans every unproblematic interpersonal compensation together with the problematic cases. To take an intra-world example, would we not think that if one person suffers from an insect bite in a world, this does not make the world bad on the whole if other people in this world live in paradisiac enjoyments? Or to take an inter-world example, imagine that the choice of alternative *A* over *B* means that one person will have slightly more disutility in *A*, while a lot of persons will escape horrendous sufferings in *B*. Surely, the value of the gains is strongly compensated by the value of the loss, and therefore the choice is justified. Second, as shown in section 2, *intra*personal compensations are not at all unproblematic.

4. Strong Negativism

The fundamental idea behind the intuition about the weight of evil is that evil has a greater weight than good, which in our welfarist framework reduces to disutility having greater weight than utility. This assumption is the mark of the family of negativist utilitarianisms, and by specifying this assumption in different directions we get different members in this family. The first division to make, if we want to examine these specific members more closely, is to distinguish negativisms which give all weight to disutility from those which give some weight to positive utility, but more weight to disutility. We can call these groups *strong negativism* and *weak negativism*, respectively. Our investigation starts with the former, more specifically, with pure negativism, where one is exclusively concerned with minimising pure disutility. To make it easy to follow the presentation of strong negativisms, we here give a taxonomy. Each branch corresponds to a particular strong negativism, roughly described by the labels in the branch. We will use the numbers when referring to a particular negativism.

Taxonomy of strong negativisms

[16] See Rawls (1971) pp. 26-27, Ohlsson (1979) pp. 28-30.



4.1. Popper

If we are interested in analysing negative utilitarianism, a very natural point of departure is Karl Popper's sketchy revision of total utilitarianism, presented in his *The Open Society and Its Enemies.*[17] Unfortunately, Popper's exposition is muddled, in part due to his dual view on revising the total utilitarianism. Some parts in his text seem to suggest that the revision is done by substituting the maximising principle with a minimising principle, but other parts in the text imply that, in addition to the substitution, the minimising principle should be incorporated in a pluralistic deontology. In the following we will, staying within a consequentialist framework, not try to give an exegesis of Popper's own view, but rather expound his view on substituting the maximising principle with a minimising ditto.

When discussing the minimising principle Popper distinguishes between *avoidable* and *unavoidable* suffering. He says that we ought to minimise avoidable suffering, but when the suffering is unavoidable the aim is to distribute suffering as equally as possible. So, the substitute for the maximising principle is this two-headed principle. Unfortunately, we are left without guidance as how to interpret these different kinds of suffering, and we must therefore *give* a reasonable meaning to these expressions.

We propose that "avoidable suffering" should be interpreted as "avoidable amount of total suffering", which mean that a situation with this kind of suffering has alternatives with *different* amounts of total suffering. We understand "unavoidable suffering" in an analogous way as "unavoidable amount of total suffering", which means that the alternatives in a situation with this kind of suffering have the same amount of total suffering. Since Popper is a pure negativist, an amount of suffering is here seen as the sum of the *pure* disutilities contained in an alternative. If this negativism should be complete it must be able to handle situations where some alternatives have the same amount of suffering and other do not. The following completion seems reasonable. First, pick out the alternatives that minimise the total amount of suffering, i.e., the alternatives each of which has at least as little disutility

[17] See Popper (1962:1) pp. 284-285 and Popper (1962:2) p. 387.

as every alternative. Then, if we have a number of minimising alternatives, the best alternatives are the alternatives with the most equally distributed total suffering, but if we just have one alternative that minimises disutility, then this alternative is the best one. This means that we here have a *lexical ordering* regarding minimising and distributing suffering.

The strict version of strong negativism (2) assigns no special weight to equal distribution of suffering. The aim here is just to minimise pure disutility.

With Popper's principle (1) or the strict negativism (2) at hand, we have no problem with avoiding the conclusion that any loss can be compensated by enough of "positive" gains (more pure positive utility). It is only unhappiness that counts, and even though one world contains more positive utility than another this does not make it better. But when it comes to negative gains (more pure disutility) Popper's principle gives far too *small* weight to suffering. As described in section 2, making already well-off person's better-off might mean marginally minimising pure disutility. To put some flesh on it, take the example from section 3 and add some details. The scientist has the opportunity to cure happy people's colds and thereby save them from the minor displeasures caused by this harmless disease. The only problem is that this can be done only by performing some experiments on human beings, and these experiments will give the subjects horrendous sufferings. Irrespective of how great these sufferings will be, we can imagine a number of "negative" gains that would compensate the losses, i.e., make the total sum of pure disutility smaller. So, if we are strict negativists or popperian negativists then the value of theses losses could always be compensated by the value of the negative gains. And this hold irrespective of how small the negative gains are. Hence, these negativisms do not avoid the Absurdity.

4.2. Right to Eliminate?

The most famous and discussed argument against Popper's strong negativism is the elimination argument, first proposed by R. N. Smart.[18] Assume that we have a technical possibility to painlessly and instantaneously eliminate humanity. If it is unavoidable that there will be some suffering in the future if the existence of humanity is continued, then we ought, according to the minimising principle, eliminate humanity. The choice is between no future suffering at all and some future suffering, and minimising means choosing the former alternative. Smart emphasises that this obligation holds irrespective of the amount of future happiness that eventually will obtain if humanity continues to exist. For Smart, the conclusion is therefore that Popper's negativism is not acceptable, and that the classical variant is to be preferred.

Is it really true that Popper's principle and the strict principle (2) give this absurd result if we assume that our aim is to minimise pure disutility? Worth considering is whether this argument is dependent on the choice of a hedonist axiology. That the hedonist is in trouble cannot be doubted, but might not the preferentialist avoid these absurdities? A preferentialist could, for example, claim

[18] Smart (1958) pp. 542-543.

that most people now living prefer to live, and that these preferences must be counted when elimination is at stake. So, the elimination results in a lot of frustrated preferences, and we must balance the evil of this against the evil of the unhappiness in the future of humanity.

There are at least two ways of criticising this approach. First, it could be claimed that the concept of a frustration is such that a frustration exists only if somebody prefers a state and it is not the case that this state obtains. If we eliminate people with preferences to live then the preferred state will not obtain, but the preferee will not exist either. Hence, one necessary conditions of frustration is missing, and we have no frustration to count.

Second, even if the above analysis is mistaken, with the consequence that in the case of frustration we actually have frustrations to count, we must balance these evils against the evil of the future unhappiness. It is without doubt so that in a likely case every future person will have *some* disutility, and even if this suffering for each person is negligible the size of the future humanity will make the total pure disutility rather great. Comparing this total disutility with the disutility of the frustrated preferences to continue to live, will, assuming the doomsday prophets are wrong, give the result that the elimination alternative has less total disutility than the non-elimination alternative.

A more general attack against the elimination argument would be to claim that the intuitive appeal of this argument is based on highly controversial evaluations concerning the happiness and suffering of future persons contingent on our choice. The contingency is constituted by the fact that we have the power to decide whether humanity will continue to exist and reproduce, which means that we here have a different persons, different number comparison. And it could be claimed that the controversial intuition underlying the elimination argument is that we ought to prevent suffering even in the special sense of choosing an alternative with the result that suffering in *new and different* people is prevented.[19]

This is a mistaken view, because we could, at least when it concerns the hedonist approach, easily formulate the elimination argument focusing entirely on the happiness and unhappiness of the people existing in the choice situation. That is, the talk about utility and disutility can be restricted to *now* existing persons. In contrast to this, it seems more difficult to show that a preferentialist pure negativist must in some absurd way proclaim elimination when we restrict our attention to now existing persons. The second argument against the preferentialist approach presupposed that we counted the whole future humanity and not just the future of now existing persons. But alter the example so that each future life of the present people is long and contains a lot of small disutilities. Could it not be the case that each life would be very satisfactory, but at the same time the total pure disutility of the future lives would be greater than the disutility of the frustrated preferences to live? Due to the length of each life the small disutilities would add up to a rather great pure total. And surely it would be absurd to proclaim elimination when we have the opportunity to let people have very satisfactory futures.

[19] The discussion on this subject is found in Ch. 4 in this essay.

Irrespective of the interpretation of welfare, the pure negativist expresses an evaluation of suffering that sounds similar to Schopenhauer's:

> (...) were the evil in the world even a hundred times less than it is, its mere existence would be sufficient to establish a truth that may be expressed in various ways (...) that we have not to be pleased but rather sorry about the existence of the world; that its non-existence would be preferable to its existence (...)[20]

This pessimism is close to the pure negativist's view. For surely, even if in some situations we ought not to eliminate mankind, since we will thereby frustrate preferences, it would have been better, on the pure negativist's view, if we never existed in the first place. For after all, to live as a human always brings some pain. But is not this a kind of perverted pessimism rather than an acceptable negativism?

4.3. Mixed Approach

So far, the principles of strong negativism have been spelled out as concerned only with pure disutility. One could argue that this approach is too extreme in the sense that it is never possible for utility to compensate disutility. Would it not be more reasonable to argue that some disutility can be compensated by utility? The mixed negativist thinks so. Besides, another reason for considering the mixed approach is that one can wonder if not the intuitive appeal of the problems so far presented is dependent on interpreting the strong negativist as a pure disutility minimiser. To give the mixed approach a fair chance to stand the tests, we must first explicate it in more detail and formulate the most reasonable version.

According to this approach, the aim is to minimise mixed disutility but when some alternative just contains pure disutility, the pure disutility in this alternative is counted as well as the mixed disutility in the other alternatives. When the utility outweighs the disutility or when we have no disutility at all, we could for convenience say that we have a zero mixed disutility. (Otherwise the mixed approach would suffer from serious incomparability; a state with mixed disutility could not be compared with a state where the utility exactly outweighs the disutility, or where disutility is missing.)

The minimising can take one of the two forms: a *future oriented* form where the object is to minimise mixed disutility in the future, and a *global* form where the object is to minimise the mixed disutility of the whole world-history. Furthermore, this mixed disutility can be constructed in two ways. Thereby we get two different versions of the mixed approach. One way is, for each person, to sum the pure disutility with the pure utility in his life, and then sum these mixed individual disutilities. If the sum for one person is negative we have a case of intrapersonal mixed disutility, but if one person just suffers pure disutility then this should also be added to the sum. According to this version the disutility to minimise is the sum of these personal disutilities. Let us therefore call it the *personal view*. Another way is to sum the disutility of every negative moment with the utility of every positive

[20] Schopenhauer (1969: 2) p. 576.

moment. The sum to minimise is this impersonal disutility. Let us therefore call it the *impersonal view*. The essential difference between the personal and impersonal approach could be brought out by saying that the former denies but the latter accepts that one person's utility can compensate *another* person's disutility.[21] The following diagram illustrates the different kinds of disutilities.

Mixed disutilities

	past utility	future utility	personal sums
person A	-9	5	-4
person B	7	-6	1
impersonal sum	-2	-1	-3

Here -1 is the impersonal future mixed disutility, -2 is the impersonal past mixed disutility, and -3 is the impersonal global mixed disutility. The personal global mixed disutility is the sum -4 and the personal future mixed disutility is -6.

Is there an reasonable version of mixed negativism? We can drop the global versions, i.e., the principles (5) and (6), because they do not give adequate weight to future disutility. Consider a case where one alternative has the future pure disutility -18, and the other the future pure disutility - 1. The alternatives have the same past utility 20. These numbers are taken to represent the past and future utilities of one and the same person. According to both versions of the global mixed approach these alternatives are indifferent, because the total utility in both cases is positive (2 and 19) and hence we have zero disutility in both cases. (Notice that in this special case the sum to minimise for the impersonalist is this personal sum.) Furthermore, assume that the two possible futures have the same number of moments, and that each moment in the -1-future has less disutility than each moment in the -18-future. This details do not alter the evaluation. The alternatives are still indifferent. But then it is shown that (5) and (6) do not satisfy the very reasonable *Negative Pareto Principle* that states that

> If two worlds A and B
> (i) contain the same set of individuals,
> (ii) for each individual it is the case that A and B have the same number of moments, A and B are positively invariant, and A and B are indifferently invariant,
> (iii) for some individual some negative moment in A has less disutility than the corresponding negative moment in B, but for the rest A and B are negatively invariant for each person,
> then A is better than B.

[21] The impersonal mixed approach is, of course, a hideous negativism in the eyes of those who are convinced that the solution to the puzzle with disutility is to forbid every interpersonal compensation. See section 3 in the current chapter.

Not even the future oriented variants, i.e., principles (3) and (4), satisfy this compelling Negative Pareto Principle. For assume a person has to choose between two satisfactory futures, which have the same number of moments, are positively and indifferently invariant, and each negative moment in one future has less disutility than each negative moment in the other. Then, absurdly enough, the principles judge the futures as indifferent, since they have the same amount of personal (and impersonal) mixed disutility.

In sum, none of the mixed negativisms satisfies the reasonable Negative Pareto Principle, and hence we can without doubt dismiss all of them.

4.4. Level Negativism

There is an alternative interpretation of strong negativism according to which not any pure or mixed disutility should be counted but just the disutility which is below a certain level. On this view, only disutility below some threshold is morally significant. The disutility above the level is maybe bad for the sufferer, but it cannot make an alternative worse.

A rough formulation of this negativism, principle (7), is:

> a state T1 is at least as good as a state T2 iff the total disutility below the level in T1 is at least as small as the total disutility below the level in T2.[22]

This 'total disutility' can be interpreted in many ways. It can be a sum of disutilities tied to welfare moments, individual lives or sets of individual lives with that in common that the mixed or pure disutility of these units is below a certain level. Irrespective of how we interpret this total, one problem is how we should determine this important level. Furthermore, this principle runs the risk of not satisfying the Negative Pareto Principle. For consider two alternatives which are positively and indifferently invariant, and each negative moment in A has less disutility than each negative moment in B, but the total disutility in each world is above the level. Principle (7) seems to be forced to judge these alternatives as indifferent.

An alternative way to compare states according to the special disutility is, instead of minimising sums of disutility, to minimise the *number of persons* whose disutility is below a certain level, i.e., principle (8). Rescher formulates this approach so

> The number of individuals whose share of utility falls below the 'minimal' level is to be made as small as possible.[23]

[22] Some authors have formulated a similar principle, but one where the level is higher. They claim that there is a certain level of *happiness* such that if individuals fall beneath it, morality requires that we push them as close to the level as possible. But once above this point there is no particular obligation to improve the lot of others. Since the happiness below the level is morally significant, this principle is a variant of *weak* negativism (section 5). If strong level negativism has problems with Negative Pareto, weak level negativism will have problems with Positive Pareto (section 4.5). For a defender of this view see Locke (1987). pp. 144 - 157.

[23] Rescher (1966) pp. 96-97.

Rescher's approach does not distinguish between utility and disutility. To repair this, a reasonable modification of the principle above could be:

> a state T1 is at least as good as a state T2 iff the number of individuals with a disutility below the level is at least as small in T1 as in T2.

But once again we get problems with Negative Pareto. For assume that we have two alternatives A and B, which have the same number of moments, are positively and indifferently invariant, and each negative moment in A has less disutility than each negative moment in B. Moreover, assume that the total (mixed) disutility for each person in each alternative is below the level. Then A is indifferent to B, despite the fact that each person in this latter state would suffer much more than in the former.[24]

A more general attack against both forms of level negativism is that they do not pay sufficiently attention to the unhappiness of the persons whose disutility is above the level. Consider a state where only one person suffers below the level and the rest is happy, and compare it to a state where everyone suffers but not below the level. The former state must, according to level negativism, be better no matter how great the number of people suffering in the latter state is.

4.5. The Weakness of All Strong Negativisms

Is every strong negativism disqualified? Level negativism and pure negativism seem to be impossible to save, but maybe a revision of the mixed negativism would handle the problems with Negative Pareto?

Whatever revision you make of the mixed approach, it is still a strong negativism. And all the strong negativisms are per definition incapable of judging situations where the disutility is invariant but positive utility varies. We can vary this positive utility either by varying only the degree of positive moments and holding the number of positive moments constant, or by varying the number of positive moments. In both cases strong negativism gives counterintuitive results.

For example, none of these strong negativisms is compatible with the reasonable *Positive Pareto Principle* that states that

> If two worlds A and B
> (i) contain the same set of individuals,
> (ii) for each individual it is the case that A and B have the same number of moments, A and B are negatively invariant, and A and B are indifferently invariant,
> (iii) for some individual some positive moment in A has more utility than the corresponding positive moment in B, but for the rest A and B are positively invariant for each person,

[24] Rescher claims that this principle should be applied in an economy of scarcity. (Ibid., p. 96) That is, it should be applied in a situation such that, if everyone is given a share proportional to his claims and desert then someone or everyone is pressed beneath the level. This restriction doesn't influence the example above, for we can without difficulties imagine that the counterexample described is an economy of scarcity.

> then A is better than B.

If we can make at least someone more happy at some moment without giving anyone more disutility at some moment, i.e., if the extra utility is for free, naturally the best thing to do is to give all of them this extra utility.

A similar reasoning can be applied to cases where the number of positive moments varies. Consider a case where we have the choice to prolong a person's life with a long and very satisfactory future. She could, for example, have a weak heart, and by giving her a new heart her life would be prolonged. Assume that her past life was satisfactory, and the happy future would be free from disutility, and each of the extra positive moments would have greater utility than each of the foregoing. According to every *hedonist* strong negativism, the choice between prolongation and non-prolongation is a matter of indifference no matter how great the utility of the extra moments would be. But surely, if every extra moment has positive utility higher than any past positive moment then it must be better to prolong the life. Or to be more exact, we hold the following principle, *The Limited Positive Mere Addition Principle*[25], as correct:

> If two worlds A and B are (i) negatively invariant , (ii) indifferently invariant , (iii) positively invariant except that A contains at least one more positive moment than B and each of these extra moments has a utility greater than the utility of any of the moments common to A and B, then A is better than B.

Notice that the hedonist judgement would not be altered if the non-prolongation in fact had been a painless killing of the person whose life otherwise would be long and very satisfactory. On the other hand, the preferentialist could insist that we must consider the preference to live that the man might have. And if non-prolongation frustrates a preference we have no longer two negatively invariant states.

In sum, it is one thing that strong negativism gives all weight to disutility when disutility stands against utility, i.e., when one state contains more disutility and the other contains more utility. But in all the examples above except the last one this is not the case, and to give everything to disutility in cases like these is surely not reasonable, and should not be so even for a convinced negativist.

These disadvantages might be removed if we abandoned the stubborn and exclusive focusing on disutility, so typical for strong negativism. Let us not give *all* weight to disutility, but just *more* weight. This is the essential feature of weak negativism, which we now turn our attention to.

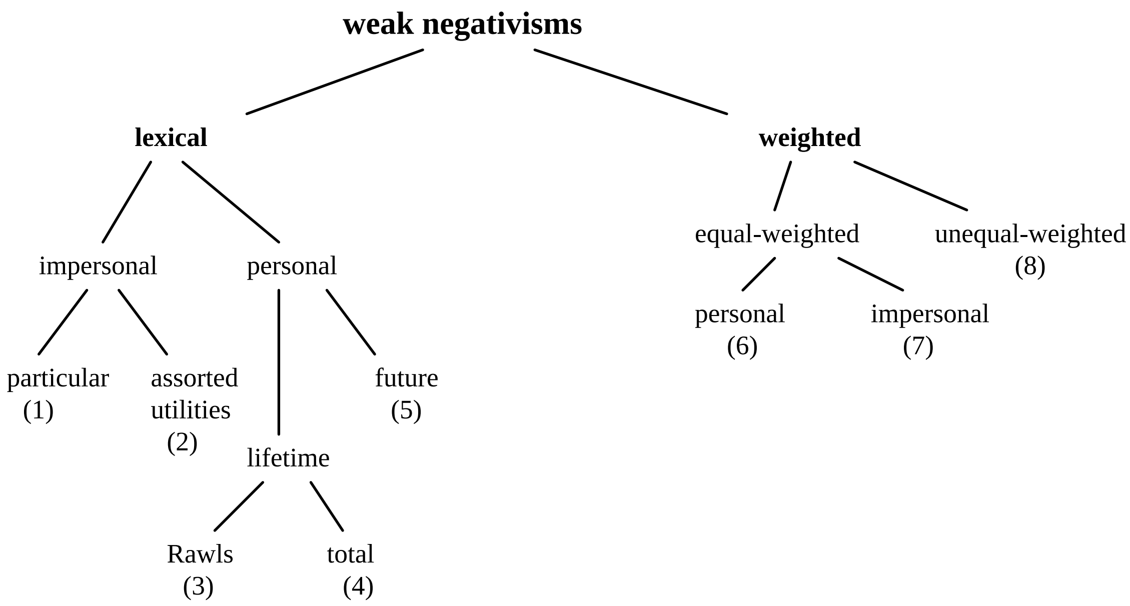
5. Weak Negativism

We will consider two ways of interpreting the claim that disutility has more weight but not all the weight. We could call the negativisms generated by these interpretations *lexical negativism* and *weighted negativism*. The latter can be divided

[25] For a more general principle concerning positive additions see section 7 in the current chapter.

into *equal-* and *unequal-weighted negativism*. To make it more easy to follow the discussion we once again give a taxonomy.

Taxonomy of weak negativisms



5.1. Lexical Negativism

To use lexical weight when analysing the relation between different values is not nothing new or unusual in philosophy. Ross, for instance, applies a lexical structure when comparing pleasure with virtue, and he writes:

> With respect to pleasure and virtue, it seems to me much more likely to be the truth that *no* amount of pleasure is equal to any amount of virtue, that in fact virtue belongs to a higher order of value, beginning at a point higher on the scale of value than pleasure ever reaches; in other words, that while pleasure is comparable in value with virtue (i.e., can be said to be less valuable than the virtue) it is not commensurable with it, as a finite duration is not commensurable with an infinite duration.[26]

This intuitive explanation of a lexical structure can be given a more precise meaning using the following general formulation of *lexical betterness*:

> a vector of values $(a_1,..., a_n)$ is better than a vector $(b_1,..., b_n)$ iff there is some i such that $a_i > b_i$ and for all $j < i$ it is the case that $a_j = b_j$.

So, if we want to express Ross' claim we ought to place the measure of virtue before the measure of pleasure in the vector. We could then say that virtue has greater *lexical weight* than pleasure.

[26] Ross (1930) p. 150.

If we want to apply this value structure to weak negativisms we have to represent a world W with a utility vector of the disutilities and utilities belonging to W ordered from the most important to the least important. The claim that disutility has greater weight can now be expressed by letting the disutilities have greater lexical weight. But still the utility has some weight in the sense that if the disutilities are the same in the alternatives, and hence we cannot minimise the disutility any further, then we ought to maximise the utility. Depending on what kinds of disutilities we choose in establishing this order, we get different lexical negativisms. Is it the disutility of particular welfare moments, of some individual sets of moments, or of some collective sets of moments that should have the greatest lexical weight? Is it the disutility of the worst-off welfare moment, individual, or group of individuals that has the most importance? As we answer these questions the different lexical negativisms will be stated.

Let us start with considering a lexicalism that takes seriously the disutility of particular welfare moments, i.e., principle (1). The following is a very concise formulation of this negativism.

> Given two worlds which contain the same number of moments of phenomenal experience, that is better which has the best (more pleasurable or less painful) worst-off (most painful or least pleasurable) phenomenal experience, (or, in case of a tie, the best second worst-off, etc.) Any world is equal in value to another which has the same number of phenomenal experiences as the first at each level of value plus any number of null-valenced moments, so if we rank worlds containing the same number of moments we can rank also those which do not.[27]

This formulation does not say a word too much, but to make it easily understood some comments are not out of place. The referred moments can be identified with the moments in a world, and it is the disutilities and utilities of these moments that constitute the vector, ordered from the most unpleasurable (or least pleasurable), a_1 , to the least unpleasurable (or most pleasurable), a_n. So, less formally, if you compare two vectors with the same number of moments, you look for the worst-off moment in each vector, and if one is worse than the other the world with the worse one is worse than the other world. If they have the same disutility you go on to the second worst-off and the procedure is repeated.

The last sentence in the quotation means that if you add some indifferent moments to a world this does not change the value. The indifferent moments do not make any difference. Therefore, if you want to compare two worlds w_1 and w_2 that have different number of moments you transform the world with the lesser number, w_2, to a world with same number by adding indifferent moments. Then, the value relation (better, worse, or indifferent) that the transformed world stands into w_1 is also the relation that w_2 stands into w_1.

The measurability assumption required for this negativism is very weak. It suffices if we can measure welfare ordinally. Unfortunately, this advantage is, as far as we can see, the only one. Firstly, the assumption that indifferent moments make

[27] Mendola (1990) p. 79. Parfit seems to describe a similar lexicalism in Parfit (1984) pp. 344-345.

no difference is problematic.[28] Compare two lives with the same number of moments, the first composed of moments with splendid utility, and the second composed of moments with a utility near zero. According to principle (1), the first life is better than the second. This is uncontroversial and implied by the Positive Pareto Principle. But modify the example so that one more moment with small utility is added to the second life, while other things are equal. Then principle (1) evaluates the splendid life as worse than the second life. For when we compare these lives we have to add an indifferent moment to the splendid life, and then this worst moment in the splendid life is worse than the worst moment in the second life.

Secondly, look once again at the elimination argument. According to this negativism we ought to eliminate humanity if the prolongation of humanity will give us *one* welfare moment with a disutility greater than any in the alternative world. This would even hold in a situation where the disutility of this state was negligible (and hence all the other negative moments were negligible), and obtained in a very satisfactory life. Furthermore, if we had to choose between a world where one person suffers one short moment in his life, but otherwise lives in heaven, and a world where each moment in his life gives disutility, but each moment to a lesser degree than the negative moment in the first world, we must according to this approach choose the latter. This would not be altered if the second life were much longer than the first one.

In sum, this negativism does not allow any intrapersonal compensation, but surely some compensations are uncontroversial. Moreover, the impersonal way of focusing on the disutility of the worst-off moments does make this approach insensitive to the numbers of unpleasurable moments. For example, a long life where the person in each moment is in great pain, with one moment worse than the others, is better than a much shorter life where each moment has a low degree of pain except one which has somewhat greater disutility than the worst moment in the former life. For instance, the profile (-8,-7 ,-7 ,..., -7) is better than (-9, -1, -1), irrespective of how many moments, all with -7 disutility, you have in the former profile.

One way of abandoning the emphasis on particular moments, while still remaining in the impersonalist frame, is to give the greatest lexical weight to the total disutility of all the negative moments obtaining in the world, i.e., principle (2). If two worlds have the same total of this disutility, we should maximise the total utility of all the positive moments. That is, in the first place minimise total pure disutility, and in the second place maximise total pure utility. Here the number of negative moments is not irrelevant, because more moments means a greater total. But still every intrapersonal compensation is forbidden; more pleasure cannot compensate more displeasure. So, for example, when we have the option to prolong a life that will be very satisfactory both overall and in the future, this is forbidden if the future contains just a slight pain. Notice that this absurd result also holds for a more personal lexical negativism where the greatest lexical weight is given to the person who has the greatest total of pure disutility. This total disutility comes in the vector first followed by lesser totals of pure personal disutility, and when we come to positive utility we order them from the greatest personal total of pure utility to the

[28] This was pointed out for us by Wlodek Rabinowicz.

smallest. More alarming is that this approach does not avoid the Absurdity. It behaves here as the pure strong negativism, i.e., for any number of any size of losses if we have a sufficient number of negative gains then the value of the losses will be compensated by the value of the gains.

Maybe these drawbacks will disappear if we leave the impersonal frame and focus not on moments but on lives or periods of lives, thereby letting the utility intrapersonally compensate the disutility. One way here is to represent a world with a vector of (mixed) lifetime disutilities and utilities, giving the greatest lexical weight to the worst-off person and the least lexical weight to the best-off person, i.e., principle (3). We use this vector in the usual way. So, in the first place, maximise the lifetime utility for the worst-off, i.e., look at the worst-off in each alternative and choose the alternative where a worst off is best-off as compared to those worst-off in the other alternatives. If there is a tie make the second worst-off best-off and so forth. This comes very close to the welfare interpretation of Rawls' leximin.[29] This negativism avoids the Absurdity, since if we make the losers' lives unsatisfactory or more unsatisfactory then we will not make the worst-off best-off. Making the worst-off worse-off cannot be compensated by making better-off persons or person-stages happier. This holds irrespective of how many of these happy people or stages we have.

An important disadvantage here is that elimination is obligatory if someone in the non-elimination world is worse-off than all the others due to a slight pain, while the others in contrast live in heavenly pleasures. Imagine, for example, that the worse off person is very old and during the last minute he will feel some pain from his worn out heart. Assume that this can be captured in the following scheme.

	oldie's utility	the others' utility
elimination world	10	10
non-elimination world	9.9 (10 + -0.1)	30 (10 + 20)

Let the past utility in each world be 10 for each person. The last minute for oldie gives him the disutility of -0.1, and his lifetime utility is hence 9.9. For the others the future is twice as pleasurable as the past. In a situation like this the Rawlsian negativism must judge the elimination world as the best one, despite the facts that non-elimination just gives the worst-off person a slight pain but gives the others lots of pleasure. Notice that the absurdities with elimination cannot be avoided by modifying the Rawlsian negativism so that the lexical weights are given the *future* utility of each person and not the lifetime utility, i.e., principle (5). The futures to compare can be represented with (0,0,...,0) and (-0.1, 20,...,20), where each number

[29] We say close because in welfare economics the ranking of the lives is often thought of as an ordinal comparison, and not the result of for each life totting up utilities and disutilities to get an lifetime utility. It must also be noted that Rawls himself does not claim that we should distribute *subjective* welfare according to the maximin principle. Instead, he proposes an objective interpretation of welfare and argues that it is an index of "primary goods" that are the proper object of distribution.

represents the future utility of a person. And, clearly, the first vector representing the elimination must be judged as the best.

One could argue against the argument above that it presupposes a hedonist interpretation of welfare. If we were preferentialists we could claim that the important disutility of the frustrated life preferences is not represented in the elimination alternative. But irrespective of the interpretation of welfare the Rawlsian leximin has counter intuitive results when it is only positive utility that is at stake. The lexical character does not allow any interpersonal trade-offs at all. Compare the following vectors, (21, 21, 21, 21) and (20.9, 60, 60, 60), where each number represents the lifetime utility of a person, and in the first world the first person has 0.1 units more pure utility than in the second world while for the other persons the first world contains 39 units less pure utility. According to the Rawlsian negativism the first world or vector is the better one, despite that in the second all but one have a much more satisfactory life. The utility difference for the worst-off person is marginal and consists only in pure utility. Do we not want to say that the gains for the others compensate the loss for the worst-off, especially when this worst-off person has a satisfactory life anyhow?

To make room for some interpersonal compensation, we could give lexical weight to totals of lifetime utilities, giving more weight to the total disutility of unsatisfactory lives than to the total utility of satisfactory lives, i.e., principle (4). This approach gives the correct answer in both the elimination example and the trade-off example mentioned above. But if we modify the elimination example and state that person 1 has the lifetime disutility - 9.9 in the elimination alternative, and the lifetime disutility -10, (-9.9, -0.1), in the non-elimination alternative, then once again we have a counterintuitive result. Due to the slight pain in the last minute of the unhappy person's life in the second alternative it becomes obligatory to eliminate humanity, despite the fact that all the others would otherwise have lived satisfactory lives. Of course, this argument has no real bite for the preferentialist negativist. He would also count the frustrations of the happy people, all of them wanting to live. But surely one could imagine examples, not necessarily choices between elimination and non-elimination, where we can greatly benefit a lot of happy people at the cost of giving one unhappy person some marginal disutility. Should we not want to say that in cases like these compensation is permitted, since the cost for the unhappy person is very small?

5.2. Weighted Negativism

Another interpretation of the weight concept used when talking about the weight of disutility is the multiplicative one, where the disutility is weighted by a number α and utility is weighted by a number β smaller than α, and both α and β are greater than zero. That the disutility and the utility are weighted means here that they are *multiplied* by a certain number. These products give the intrinsic value of utility and disutility, and the aim is here to maximise the sum of these products. So the weighted negativism differs from the total utilitarianism not by substituting the maximising approach with a minimising, but by not giving equal weight to utility and disutility.

With these concepts at hand we can give meaning to the catch phrases "it takes a fairly great quantity of utility to outweigh a fairly small quantity of disutility", to take an axiological formulation, and "the infliction of pain on any person is justified only by the conferment not of an equal but of a substantially greater amount of pleasure on some else", to take a normative formulation. [30] If we, for instance, have the utility 2 and the disutility -2, one could say that the utility is too small to outweigh the disutility, or that we are not justified to inflict the disutility. How much more utility is needed depends on the weights. The catch phrases suggests that this amount of utility must be rather great, but this need not hold for every weighted negativism.

The weighting procedure can take different forms, and here we want to explore *equal weighted negativism* and *unequal weighted negativism*. The former gives the same weight to every disutility and the same weight to every utility. Though of course the disutility weight is not the same as the utility weight. The latter gives different weights to different disutilities and/or different weights to different utilities.

As with lexical negativisms nothing precise can be said about the weighted negativisms if we do not first specify the proper units of negativist concern. Two possibilities come naturally: one impersonal and one personal. This means, applying the distinction to the equal-weighted negativism, that either we give the weight to the utilities of welfare moments and the negative moments get the greatest weight as in principle (6), or we give it to the lifetime utilities and the unsatisfactory lives get the greatest weight as in principle (7). To give the weights to moments means that the sum to maximise is the sum constituted by the products α times the total disutility of all negative moments and β times the total utility of all positive moments, where α and β are characterised as above. To give weights to lifetime utilities means that the sum is constituted by the products α times the total disutility of all unsatisfactory lives and β times total utility of the satisfactory lives.[31]

No matter what weights we use and what negativist units we attach these weights to the equal weighted negativism cannot avoid the Absurdity. The reason is simply this. Since positive utility is given some weight each gain has some value,

[30] The axiological formulation is found in Griffin (1979) pp. 51-52. He calls this the basic sentiment version of weak negativisms. The normative formulation is found in Ross (1939) p. 75.

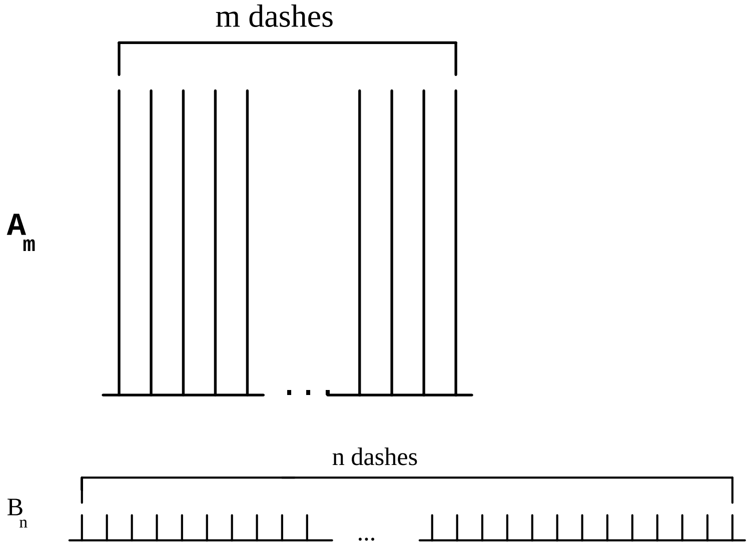
[31] The reason for this is simply the law of distributivity according to which $(\alpha a + \alpha b) = \alpha(a + b)$, where α can be any weight and a and b any number and hence any utility or disutility.

however small it may be. Furthermore each positive moment is given the same weight. So, although the value of each gain is relatively smaller now when utility has less weight than in total utilitarianism, we can without problems add a sufficient number of gains so that the loss is strongly compensated by the gains.

There is another weakness with every equal weighted negativism and, more generally, with every negativism that gives each positive moment equal weight. None of them can avoid *individual repugnant conclusions*.[32] These conclusions hold for different moment cases, as compared to the Intrapersonal Absurdity which only holds for same moment cases. One of these conclusion concerns distributions of positive utility, while the two others are about distributions of positive and negative utility. The *Positive Repugnant Conclusion* states that for any given life for a person with a given number of positive moments, all with a very high positive utility, there exists another possible life for him with a much larger number of positive moments, all with a very low positive utility (e.g., each with a utility just above zero), and this second life is better than the first.

To make things easier to grasp when contrasting this conclusion with the others we present these conclusions by using diagrams, where each dash represent the utility of a moment in a person's life.

The Positive Repugnant Conclusion



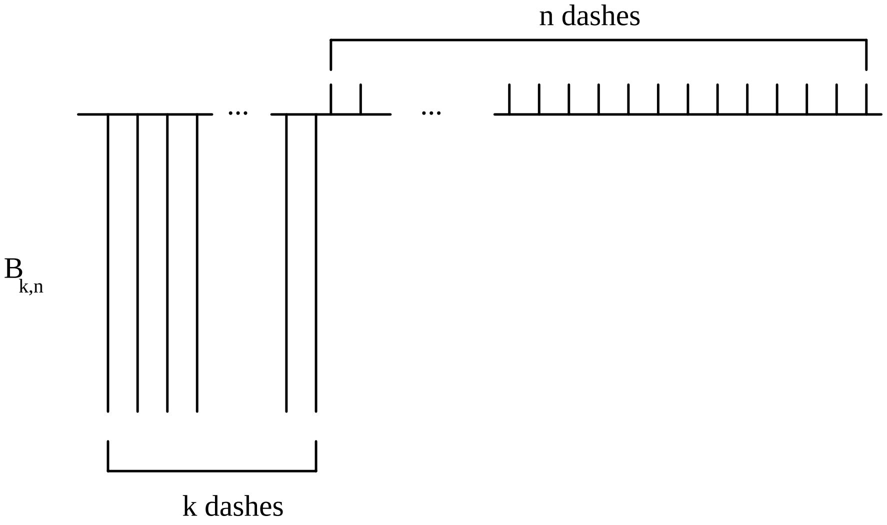
The conclusion could now be more exactly formulated thus. For any number *m* of very happy moments, there is a number *n* such that if this is the number of moments each with a utility near zero, then B_n is better than A_m.

Parfit gives a drastic illustration of the repugnancy inherent in this conclusion. Consider a life where the only good things would be muzak and potatoes. Are we

[32] Parfit (1984) p. 160. Repugnant conclusions can also be applied to *populations*. For a thorough discussion on this matter, see Ch. 4.

ready to say that for any splendid life there is always a much longer muzak-and-potato-life that is better than the splendid one?

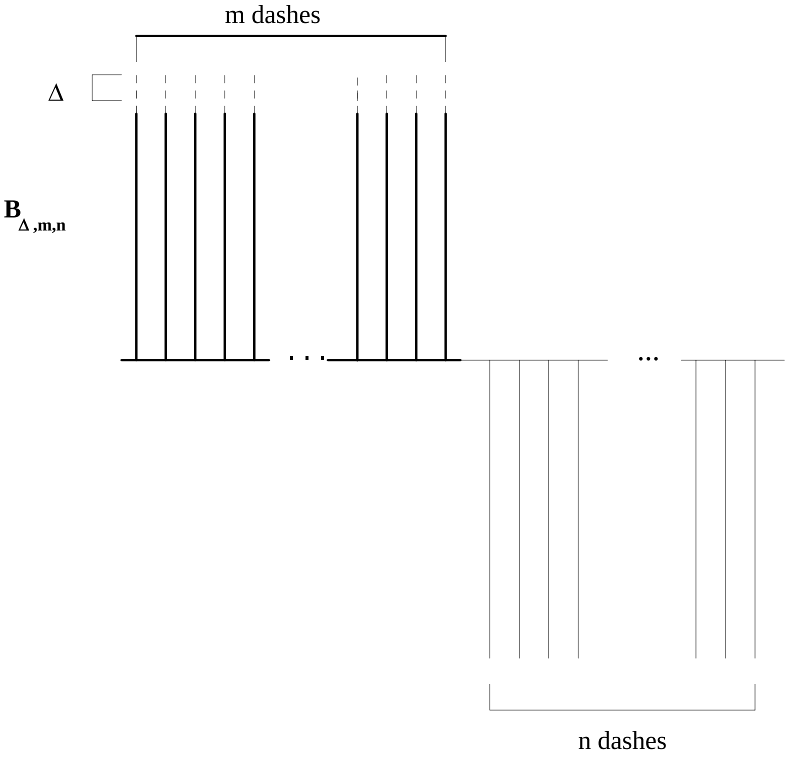
The Negative Repugnant Conclusion (1)



This conclusion states that for any number *m* of very happy moments, and for any number *k* of very unhappy moments there is a number *n* of moments each having a utility close to zero such that $B_{k,n}$ is better than A_m (see the figure above).

To illustrate consider a comparison between a splendid life and a life containing some intense sufferings except for some dull moments with muzak and potatoes. Is it better to take the life with the intense sufferings and potatoes if this life contains sufficiently many dull moments?

The last of this conclusion *The Negative Repugnant Conclusion (2)* is illustrated by the following diagram.



For any number n of very unhappy moments, and for any small difference , there is a number m of very happy moments such that $B_{,m,n}$ is better than A_m.

To illustrate consider a modification of the case presented in section 2. Assume that the medicine will make the first *x* years of our lives somewhat better-off at each moment but on the other hand this medicine have the strange result of prolonging our life. And this extra moments will all be of great disutility. Would we want to say that if *x* is sufficiently great then the prolonged life is the better one?

We answer No to all of these questions, but if you are a equal-weighted negativist then you must answer Yes to all of them. The reason is simply that if you give *some* weight to positive moments and moreover give the *same* weight to them, then for any life containing some sufferings the value of the life could be made how great as possible just by adding positive moments (or in the case of the third conclusion, just by adding positive gains).

So, our hope stands to the unequal-weighted negativism, i.e., principle (8). Before we explore this negativism we need to say something about the negativist aggregation method (section 6), and summarise the most important conditions of acceptability we have hitherto formulated (section 7).

6. Aggregation

So far we have not clearly distinguished between different aggregations of negativist units. When constructing weak and strong negativisms we have supposed that the aggregation was in terms of total sums of disutilities. In contrast to this approach we could equally well talk about average disutility, i.e., disutility divided with the number of moments or persons. In most discussed cases we have supposed that the number of moments was the same and hence that the number of persons was the same. And in these cases there is no difference between the total approach and the average approach. But of course, if we want to state the most acceptable negativism we are bound to make a choice between these approaches. Otherwise we would just have a partial theory which could only manage to judge cases with same number of moments.

We think that the average approach is not acceptable, no matter how we interpret the disutility sum which is to be divided. This sum could be interpreted to be the total pure disutility, the total impersonal mixed disutility, or the personal mixed disutility. The following example is a general attack on average negativisms.

World 1	$-n_1,\ldots, -n_n$
World 2	$-n_1,\ldots, -n_{n+1}$
World 3	$-n_1,\ldots, -n_{n+1+1}$
•	-
•	-
World n	$-n_1,..,-n_{n+m}$

For any *i* and *j*, $-n_i = -n_j$. Each $-n_i$ represent the disutility of a negativist unit. This list can be given different interpretations according to how we interpret these units. If we see them as moments of a person's life the list illustrates possible developments of a life, i.e., a person can have hells of different lengths with world one as the shortest. The subscripts are then pointing out a temporal position of a particular moment. Now, if we can prolong the life with moments of the same disutility as the average, then, on the average approach, it does not matter how many of these we add, since the average is not changed. So, since every world has the same average disutility they will be judged as equally good. But surely, a shorter hell must be better than a longer. To be more exact, we want to say that every acceptable negativism must satisfy the *Negative Mere Addition Principle.* It states that:

> If two worlds A and B are (i) positively invariant , (ii) indifferently invariant , (iii) negatively invariant except that A contains at least one more negative moment than B, then A is worse than B.

Notice that given these homogeneous utility profiles we get the same result whatever negativist average approach we choose. There is no difference between the pure and the mixed negativist, since the impersonal and personal mixed approaches must in a case like this focus on the personal pure disutility.

If we see the units as persons, and hence the subscripts are picking out persons, the situation is one where we can add different numbers of unsatisfactory lives. Once again we are as averagists bound to be indifferent between the worlds. But a crowded hell must be worse than a less crowded one.

Common for these approaches is that they take the disutility of each moment or person occurring in an alternative, sums these disutilities, and divide the sum with the number of moments or persons. An alternative averaging is to sum, for each time, the disutilities occurring at that time, dividing the sum with the number of moments or persons occurring at that time, and finally to sum across these different times. This approach is no improvement. Consider a case where the sufferings occur at one particular time in each of the worlds 1 to n. That is, in world 1 the negativist units $-n_1$,..., $-n_n$ occur at one and the same time, and similarly in world 2 the units $-n_1$,..., $-n_{n+1}$ occur at one time, and likewise for the other worlds. Again, the average negativist must judge these hellish worlds as equally good.

One might argue that these conclusions depend on taking a global perspective on the case. What we ought to compare are not the averages of the whole utility profiles but the averages of the future parts of the profiles. But assume that the choice situation is at the time where the unit $-n_1$ obtains. Then all possible futures relative to this time are equally bad which surely is repugnant. We conclude that the acceptable aggregation procedure cannot be averaging.

7. Conditions of Acceptability

So far we have ended up with a set of unacceptable negativisms each failing to satisfy at least some of the proposed conditions of acceptability. In order to prove the possibility or the impossibility of an acceptable negativism it is convenient to list the conditions we have proposed so far. (To trace the sources to these conditions see the references to the sections of the text.)

(1) *The Absurdity* (Section 2, 4.1, and 5.1) Every acceptable negativism must avoid the following conclusions:

> *The Interpersonal Absurdity*
>
> For any number of pure losses, each making a life unsatisfactory or more unsatisfactory, and for any size and type of these losses, and for any size of pure gains for persons that would have good lives anyhow, there is a number *n* and a type of gain such that if *n* is the number of gains of this type and size, then the value of the gains strongly compensates the value of the losses.
>
> *The Intrapersonal Absurdity*
>
> For any number of pure losses, each making some of a person's stages unsatisfactory or more unsatisfactory, and for any size and type of these losses, and for any size of pure gains for other stages of the same person that would be good anyhow, there is a number *n* and a type of gain such that if *n* is the number of gains of this type and size, then the value of the gains strongly compensates the value of the losses.

(2) *Elimination* (Section 4.2). Whether elimination is better than non-elimination cannot (always) be solely dependent on the negative moments in each alternative.

(3) *Negative Pareto* (Section 4.3 and 4.4).

> If two worlds *A* and *B*
> (i) contain the same set of individuals,
> (ii) for each individual it is the case that *A* and *B* have the same number of moments, *A* and *B* are positively invariant, and *A* and *B* are indifferently invariant,
> (iii) for some individual some negative moment in *A* has less disutility than the corresponding negative moment in *B*, but for the rest *A* and *B* are negatively invariant for each person,
> then A is better than B.

(4) *Positive Pareto* (Section 4.5).

> If two worlds *A* and *B*
> (i) contain the same set of individuals,
> (ii) for each individual it is the case that *A* and *B* have the same number of moments, *A* and *B* are negatively invariant, and *A* and *B* are indifferently invariant,

(iii) for some individual some positive moment in *A* has more utility than the corresponding positive moment in *B*, but for the rest *A* and *B* are positively invariant for each person,
then A is better than B.

(5) *General Positive Mere Addition* (Section 4.5)

If two worlds *A* and *B* are (i) negatively invariant, (ii) indifferently invariant , (iii) positively invariant except that A contains at least one more positive moment than *B*, then *A* is not worse than *B*.

Notice that this is a more general principle than the Limited Positive Mere Addition Principle stated in section 4.5. The *General Positive Mere Addition* is applicable to any addition of positive moments. But what could be wrong with saying that adding good things never makes the world worse?

(6) *Negative Mere Addition* (Section 6.2)

If two worlds *A* and *B* are (i) positively invariant, (ii) indifferently invariant, (iii) negatively invariant except that *A* contains at least one more negative moment than *B*, then *A* is worse than *B*.

(8) *Not every intra- and interpersonal compensation is impossible* (Section 3 and 5.1).

(9) *The individual repugnant conclusions* (Section 5.2).

8. Our Proposal

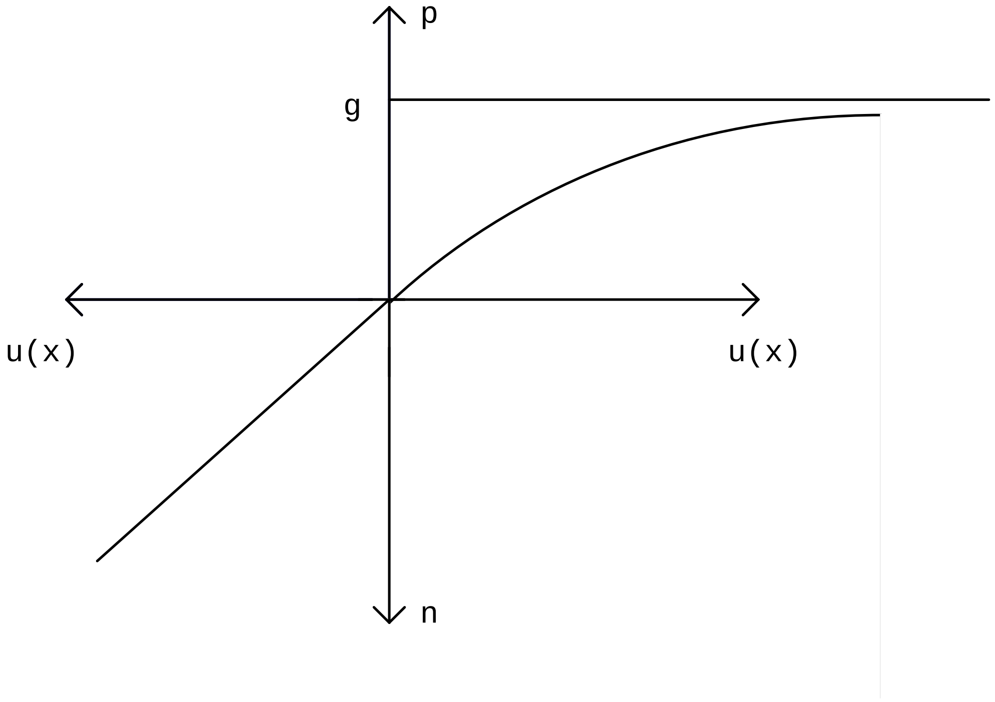
We ended section 5 by expressing our hope to the unequal-weighted negativism, (principle 8). To call it *a* principle is somewhat misleading since we have a whole family of differing theories each worthy the title. In this section we try to restrict the class of acceptable unequal-weighted negativisms.

8.1 Theory WUN

Theory WUN (Weak Unequal-weighted Negativism) can be divided into three steps. First the values of each moment in a life is calculated. These values are then used when we calculate the value of a life. The values of the lives are then used when we calculate the value of the population. Finally the value of the populations is used when we calculate the value of the world in which the population exists.

On the moment level we assign values to moments according to the following value graph.

Moment Value Function



$u(x)$ is a function that assigns utility to particular moments. p is the value function that takes the utility of positive moments as argument, so if $u(x) > 0$, then the value of x's utility equals $p(u(x))$, $p > 0$. As seen in the figure above p is a concave function with an upper limit g. This means that when u (> 0) approaches infinity then p approaches g without reaching g. The value of a moment of happiness has an upper limit that is asymptotically approached.

n is the value function that take the utility of negative moments as argument, so if $u(x) < 0$, then the value of x equals $n(u(x))$, $n < 0$. As seen n is a linear function, and the disvalue of a moment of unhappiness has no limit. (Alternatively, we could take n to be a convex function so that the weight of disutility increases with greater disutility.)

When it comes to indifferent moments we hold that indifferent moments have indifferent value, so if $u(x) = 0$, then the value of x's utility equals 0.

To calculate the value of a life, the *lifetime value,* group the moment values within the life into two sets: one for negative values (n-values) and one for positive values (p-values). Put the p-values in order of descending positive value (p_1,p_2, ..., p_n), where p_1 is consequently the greatest positive value and p_n the smallest positive value. In the case of ties any order of those tied will suffice. The lifetime value is then

> the sum of the n-values plus ($\alpha^0 p_1 + \alpha^2\, p_2 + \alpha^2\, p_3 +$,..., $+ \alpha^{k-1} p_k$) where k is the number of positive moments and $1 > \alpha > 0$.

This implies that there is a limit g' for the value of a satisfactory life, and

$$g' = g \text{ x } 1/(1-\alpha).$$

Irrespective of how many positive moments a happy life contains and how happy each of these moments is the value of the life cannot exceed g', but it approaches g' asymptotically. (Even if, *per impossibile*, each positive moment would have an infinite utility, and consequently each moment would have the value g, the value of the life could not exceed g' irrespective of the number of positive moments.) But, on the other hand, there is no limit to the disvalue of unhappy lives.

The value of a population is then calculated by subjecting the lifetime values to a similar dampening procedure as the one we used to calculate the lifetime value. Group the lifetime values into two sets, one for the negative ones and one for the positive. Put the positive lifetime values in order of descending lifetime value (P_1, P_2, ...,P_l), in case of ties, any order for those tied will suffice.[33] The *population value* is then

> the sum of the negative lifetime values plus ($\beta^0 P_1 + \beta^1 P_2 + \beta^2 P_3 +$,..., $+ \beta^{l-1} P_l$), where l is the number of positive lifetime values , and $1 > \beta > 0$.

And again we have a value limit. The value of lives with positive lifetime value has the limit

$$g'' = g' \text{ x } 1/(1-\beta).$$

Irrespective of how many such people we have, and how great positive lifetime value each life has, the value of the population cannot exceed g''. The value approaches g'' asymptotically. (If, *per impossibile*, each happy life consisted of an infinite number of positive moment each having an infinite positive utility then the value of each happy life would be g'. But according to the population function it holds that irrespective of how many of these heavenly lives we have the value of the population cannot exceed g''.) But, on the negative side, there is no limit to the disvalue of a set of unhappy lives.[34] Finally, the *value of a world* is simply the value of the population inhabiting the world.[35]

These conditions taken together do not describe a particular theory. Rather they define a set of theories, some of them differing a lot. By varying p, n, g, α and β we get different negativisms, ranging from the most extreme ones, where g is small and α and β are both close to 0, to the more moderate ones, where g is great α and β are both close to 1. We need not hold that α = β. For maybe we want to say that interpersonal compensations are more problematic than intrapersonal ones.

Moreover we may have different views on what should be counted as the ultimate axiological building blocks. Above we chose moments and their utility. But, of course, we have other alternatives. For example, we might choose the utility of

[33] Notice that if a life is satisfactory, i.e. has a total utility greater than zero, this does not imply that the lifetime value is positive. The reason is simply that WUN treats negative and positive moments *asymmetrical* both when it comes to assigning value to particular moments and when it comes to assigning value to aggregates of moments.

[34] Note that if a life is unsatisfactory, i.e. has a total utility less than zero, this implies that the life has a negative lifetime value.

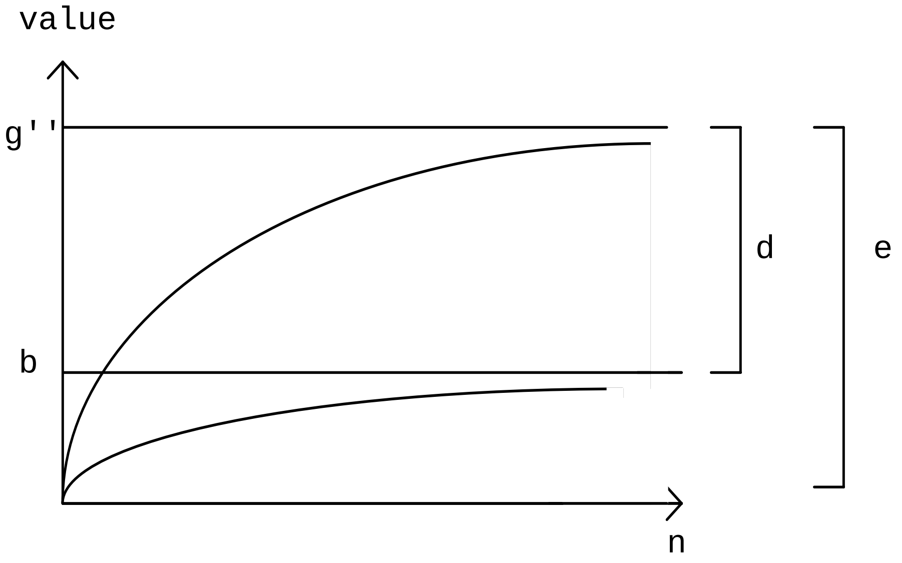
[35] Remember that in this chapter we, for simplicity, assume that each world has only one morally relevant population: the inhabitants of the world. But in fact, the concept of a relevant population is much more problematic, as will be discussed in Ch. 4, sections 3.2 and 6.6.

whole person stages as the argument for the value function on the first level. (The utility of a stage is then simply the total utility of the moment contained in the stage.) On this view there is nothing problematic with compensations *within* a person stage. One might go further and pick as the relevant utility the utility of *ages* within a person's life. The limiting case is, of course, when the utility of whole lives are seen as the relevant utility. But in that case, we have a theory that sees no problems in intrapersonal compensations. What partition of a life that is the relevant is a difficult problem. Here we have not the space to dig into this perplexing problem. But so much can be said, that there must be some partition. Otherwise, the intrapersonal compensation and the repugnant conclusions cannot be avoided. One might add that the problem of partition is nothing peculiar to WUN. For instance, any theory of equality that talks about equality between parts of people's lives meets the same problem.

Here it might be objected that since WUN incorporates so many unresolved issues it cannot be counted as a moral theory. Our response to this is that WUN cannot at the present stage be more precise. To make it more precise we have to make our intuitions on compensations more precise. Instead, one could argue that WUN's indeterminacy is an advantage, since a moral theory should not be more precise than the intuitions that the theory tries to reflect. Of course, we shall not be content with this, for to make moral progress we have to clarify our intuitions, and then, in connection with this, clarify our theory.

8.2 WUN and the Conditions of Acceptability

(1) Does WUN avoid the Absurdity? Let us start by focusing on the interpersonal case. Is it possible to find a certain number, size and type of pure losses, each making a life unsatisfactory or more unsatisfactory, and a certain size of pure gains for persons that would have good lives anyhow, so that there is no number and type of these gains that make the value of these gains strongly compensate the value of the losses? The answer is Yes. And this is best illustrated by using the following diagram.



Here n is the number of lives. b is the limit of the value of a population in which each life has the positive lifetime value k. The graph that approaches b asymptotically is consequently the value function of positive lifetime values when each value equals k. g'' is, as before, the upper limit of the value of a population in which each life has a positive lifetime value. The graph that approaches g'' asymptotically is the value function of positive lifetime values when (per impossibile) each lifetime value equals g'. And g' is, as before, the value of a life consisting of an infinite number of infinitely happy moments. d is the difference between these two value limits g'' and b. This difference d is the value difference between a population of n persons each with the lifetime value g', and a population of n persons each with the lifetime value k, where n equals infinity. [36] e is the difference between the limit g'' and the value limit of a population with the smallest possible average of positive lifetime values. A population with the smallest possible average is a population in which each person has the shortest possible life consisting of moments with the smallest possible positive utility.[37]

Assume now that we are comparing worlds where losers' losses are standing against winners' gains in the manner described in the Absurdity. Moreover, assume, which seems reasonably, that a *good life* is not just a satisfactory life but a life with positive lifetime value. If each winner has the positive lifetime value k in the world where loser is best-off and winner worst off (but still happy), then we immediately see that if the difference of total negative lifetime values between the worlds is equal to or greater than d then, irrespective of how many winners we add, each with the lifetime value k in their worst alternative, the value of the losses is not compensated by the value of the gains. Hence, theInterpersonal Absurdity is avoided.

Furthermore, if the difference of negative lifetime values is equal to or greater than e then, irrespective of how many lives we have, all with positive lifetime value, and irrespective of how great each gain is, the value of the losses cannot be compensated by the value of the gains. This is what we may call a *radical loss*.

If we, for simplicity, identify person-stages with moments, then we can easily show that the Intrapersonal Absurdity is also avoided. Let g'' now be the value limit of positive moments within a life. Let b then be the value limit of positive moments, where the value of each moment equals k. d and e are then the corresponding value differences. Now, if the difference of total negative moment value is equal to or greater than d then, irrespective of how many 'moment- winners' we have, each with the value k in their worst alternative, the value of the losses is not compensated by the value of the gains. And, of course, we have radical losses on the intrapersonal level too. If the difference in the value of negative moments is equal to or greater than e then, irrespective of how many moments we have, all with positive value, and

[36] Why is d the maximal difference? Well, the value difference between a g'-population and k-population of the same size equals $(\beta^0 + \beta^1 + \beta^2 +,...,+ \beta^{n-1})$ x (the average of the g'-population - the average of the k-population). The greater populations, the greater difference. The limit of this value difference equals (the average of the g'-population - the average of the k-population) / 1-β. This equals d, since (the average of the g'-population - the average of the k-population) / 1-β equals (the average of the g'-population / 1-β) - (the average of the k-population / 1-β) which equals g'' - b.

[37] The possibility here mentioned is a *factual* possibility. On this concept, see Ch. 2, section 2.3.

irrespective of how great each moment-gain is, the value of the moment-losses cannot be compensated by the value of the moment-gains. Hence the Intrapersonal Absurdity is also avoided.

We might add that besides the comparative concept of a radical loss, we also have the absolute concept of *radical suffering*. A set of lives constitutes a radical suffering if the sum of the negative lifetime values of these lives cannot be compensated by positive lifetime values. That is, the total of negative lifetime values is such that irrespective of how many and how happy the other persons are in the world, still the total of negative values is not compensated by the total of positive values. How do we decide if a suffering is radical? If a world contains unsatisfactory lives whose total negative lifetime value is -r and-r + g'' ≤ 0, where g'' is the limit of the value of a population of satisfactory lives, then this suffering cannot be compensated and the world must be considered as intrinsically bad. But notice that worlds that contain radical sufferings can nevertheless be compared. For instance, if the total of negative lifetime values is greater in one world than in another but the total of positive lifetime values is equal then the former world is worse.

(2) *Elimination* is no problem since happiness is always given some weight.

(3) *Negative Pareto* is fulfilled and the following is an outline of a proof.

First, we show that a negative Pareto-improvement for a person implies a higher *lifetime value* for that person. If a move from B to A means a negative Pareto-improvement for a person, then some negative moment in his life in A has less disutility than the corresponding negative moment in his life in B, and for the rest A and B are positively and negatively invariant for the person. This means that the person will have the same set of positive moments in A as in B. Hence the value of the positive moments will be the same. So, the only value difference between A and B is that the person has less *negative* moment value in A than in B. Now, according to WUN the lifetime value is the sum of the value of the negative moments and the value of the positive moments. Hence, the person will have a higher lifetime value in A.

Second, we show that a negative Pareto-improvement for at least one person implies a higher *population value*. If one state *A* is negatively Pareto-better than another *B* for a certain person, then this means that for this person we have either a transition from one *absolute* lifetime value to a higher one, or we have no transition but the lifetime value is greater in *A* than in *B*. A transition from one absolute lifetime value to a higher one is either a case where we go from negative to positive value, or a case where we go from zero to positive, or a case where we go from negative to zero.

Now, let us first look at the case where we have no transition. If each winner has negative lifetime value in both alternatives, then the sum of negative lifetime values will be less in *A*. If the sum of positive lifetime values is the same in *A* and *B* but the sum of negative lifetime values are greater in *A*, then *A* is better than *B* according to WUN.

If each winner has positive lifetime value in both alternatives, then *A* and *B* will have the same number of positive lifetime values. But for all *n,* the *n*th element in the ordered set of positive lifetime values tied to *A* (the *A*-set) will have at least as great value as the *n*th element in the set tied to *B* (the *B*-set), and for some *n*, the *n*th element in the *A*-set will have greater value than the *n*th element in the *B*-set. Hence, the value of the *A*-set will be higher than the value of the *B*-set. So, if each winner has a positive lifetime value in both *A* and *B*, then *A* is better than *B*.

Let us now look at the cases of transition. Take first the case where each winner has negative value in *B* and zero value in *A*. Since the only difference between *A* and *B* is that *A* contains less negative lifetime value, then *A* is better than *B*.

Obviously, any combination of the types of negative Pareto-improvements described so far, will make the world better according to WUN.

But what of cases in which some winners has either a transition from zero to positive or from negative to positive? In all of these cases the *A*-set will contain *more elements* than the *B*-set. Take the simple case where each winner has a positive lifetime value in A that is equal to or less than the smallest positive lifetime value in B, which we may call P_n. Then the weighted *B*-set (i.e., the set constituted by the weighted first element in the B-set, the weighted second element in the B-set and so on) and the weighted A-set (i.e., the set constituted by the weighted first element from the A-set, the weighted second element from the A-set and so on) will be identical up to $\beta^{n-1}P_n$. The only difference between the sets will be that in the weighted *A*-set we have some more terms, each having positive value. Hence the sum of the weighted *A*-set will be greater than the sum of the weighted *B*-set, and since we either have less negative lifetime value in *A* (i.e., when we have a transition from negative to positive lifetime value) or the same (i.e., when we have a transition from zero to positive lifetime value), *A* will according to WUN be better than *B*.

Obviously, any combination of the types of negative Pareto-improvements described so far, will make the world better according to WUN.

If some of the added positive lifetime values is greater than the smallest positive lifetime value in B, then the *A*-set and the *B*-set will be identical up to P_{i-1}, where *i* is the smallest number such that the greatest lifetime value of those added is greater than P_i. Now, from this follows that the sum of the weighted *A*-set up to the *i*th element will be greater than the sum of the weighted *B*-set up to the *i*th element. Since for any *i*, P_i in the *A*-set is at least as great as P_i in the *B*-set, we have that $\beta^{i-1} P_i$ in the weighted *A*-set is at least as great as $\beta^{i-1} P_i$ in the weighted *B*-set. Moreover, the A-set contains some extra lifetime values all of which should be given positive weight. Hence, the sum of the weighted *A*-set from its first to its last element is greater than the sum of the weighted *B*-set from its first to its last element. So, on the whole the sum of the weighted *A*-set is greater than the sum of the weighted *B*-set. And since *A* either contains a less or the same total of negative lifetime value, *A* is better than *B* according to WUN.

Finally, it is obvious that any combination of the possible negative Pareto-improvements will make the world better. Hence, WUN fulfils Negative Pareto.

(4) *Positive Pareto* is fulfilled. First, a positive Pareto-improvement for a person implies a higher *lifetime value* for that person. For, if a move from B to A means a positive Pareto-improvement for a person, then some positive moment in his life in A has more utility than the corresponding positive moment in his life in B, and for the

rest A and B are positively and negatively invariant for the person. This means that the person will have the same set of negative moments in A as in B. Hence the value of the negative moments will be the same. So, the only value difference between A and B is that the person has higher *positive* moment value in A than in B. Now, according to WUN the lifetime value is the sum of the value of the negative moments and the value of the positive moments. Hence, the person will have a higher lifetime value in A. Second, if one state *A* is positively Pareto better than another *B* then this means that for each winner we have either a transition from one absolute lifetime value in *B* to a higher one in *A*, or there is no transition but the lifetime value in A is higher. This is exactly analogous with the case of a negative Pareto improvement in (3). Hence, we can use the reasoning in (3) to establish the Positive Pareto Principle.

(5) *General Positive Mere Addition* is always fulfilled. First, if a person has at least one more positive moment in A than in B, then his ordered set of positive moments in A will have at least one more element than his ordered set of positive moments in B. This is analogous to the case (3) where we added lifetime values. So, we can apply the same reasoning to show that the moment addition makes the lifetime value higher. Second, if one state *A* contains at least one more positive moment than *B*, other things being equal, this means that for each winner we have either a transition from one absolute lifetime value in *B* to a higher one in *A*, or he has just a higher lifetime value in *A*. Again, the case is exactly analogous with (3).

(6) *Negative Mere Addition* is clearly fulfilled. First, if a person's life in A contains one more negative moment than his life in B, other things being equal, then his lifetime value in B will be less than his lifetime value in A. Second, if one state *A* contains at least one more negative moment than another state *B*, other things being equal, this means that for each loser we have either a transition from one absolute lifetime value in *B* to another and lower absolute value in *A*, or he has just lower lifetime value in *A*. But the losers might be described as winners if we imagine that the move was from A to B. But then, again, we have a case that is exactly analogous to (3).

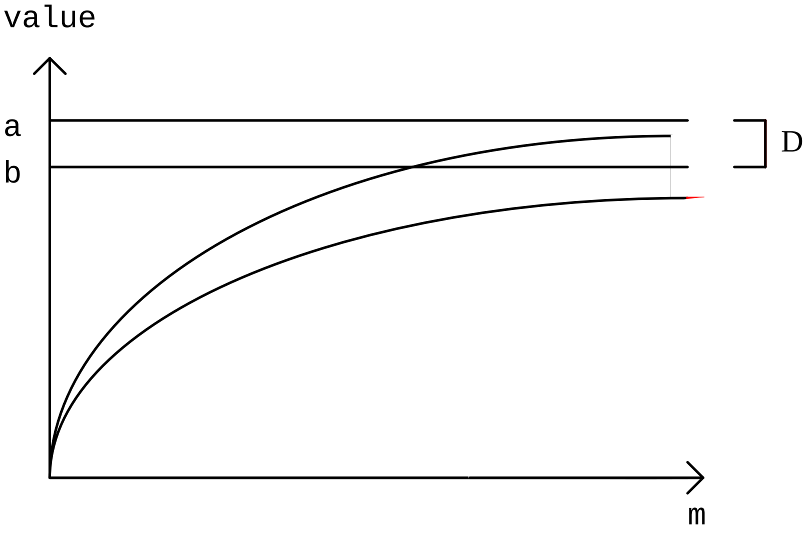
(7) *Not every inter- and intrapersonal compensation is impossible.* Obviously, this is implied by WUN.

(8) *The individual repugnant conclusions* are avoided. Take first the Positive Repugnant Conclusion. It states that for any number *m* of very happy moments, there is a number *n* such that if *n* is the number of moments each with a utility near zero, then the world with *n* marginally happy moments is better than the world with *m* very happy moments. Assume that we are comparing different possible lives for one person, and that there are no other welfare differences. Assume that one of these possible lives consists of *m* very happy moments, each with the moment value 100. Suppose that the highest possible value of a moment with utility close to zero is 10. Now, the value limit of a life consisting of moments with utility near zero equals 10 x (1/ 1-α). Then, if α = 0.9, then this value limit is 10 x (1/1-0.9) = 100. This means that irrespective of how great number of marginally happy moments we have the lifetime value of this long and dull life cannot exceed 100. And only if the dull life is infinitely long can its lifetime value equal 100. Since the lifetime value of the shorter life

consisting of very happy moments is greater than 100 (remember that WUN always assigns positive weight to the happy moment within a life), the dull life cannot have greater lifetime value even if it is infinitely long. Since there is no other welfare difference between the compared worlds, WUN does not rank the world in which the person leads a long but dull life higher than the world in which he leads a shorter but much happier life. Hence, the Positive Repugnant Conclusion is avoided.

Now, take the Negative Repugnant Conclusion (1). It states that for any number *m* of very happy moments, and for any number *k* of very unhappy moments there is a number *n* of moments each having a utility close to zero such that the $B_{k,n}$-world, (the world consisting of *k* very unhappy moments and *n* moments each having a utility close to zero), is better than the A_m-world (the world consisting of *m* very happy moments). Assume, as before, that the value of a very happy moment is 100 and the highest possible value of a moment with a utility close to zero is 10, and α = 0.9. Then for any *k*, *n*, and *m* there is no $B_{k,n}$ -world which is better than an A_m -world, since, as shown above, for any *n* and *m* there is no B_n-world, (a world consisting of *n* moments each having a utility close to zero), that is better than some A_m-world.

Finally, look at the Negative Repugnant Conclusion (2). I states that for any number *n* of very unhappy moments, and for any small difference , there is a number *m* of very happy moments such that the $B_{,m,n}$-world (the world consisting of *n* very unhappy moments and *m* very happy moments each marginally happier than the very happy moments in A_m), is better than the A_m-world (the world consisting of *m* very happy moments). That this conclusion is avoided is best shown by using a diagram.



The graph that is approaching the line *b* is the value function for very happy moments each having a utility of *x*. *b* is the value limit of these moments. Assume that an A_m-world consists of this type of moments.

The graph that is approaching the line *a* is the value function of very happy moment each having the utility *x* + which is marginally higher than the utility of the moments in the A_m -worlds. Assume that a $B_{,m,n}$-world consists of these *x* + -moments, and *a* is the value limit of these moments.

D is the difference between the value limits of the different sets of moments. Now, for some number *n**, the negative value of *n** unhappy moments outweighs *D*. So, there is no *m*, such that a $B_{,m,n^*}$-world is better than an A_m-world. But the conclusion under consideration says that there is such an *m*. Hence, the Negative Repugnant Conclusion (2) is avoided.

8.3 WUN's Drawbacks

WUN seems hitherto to be something of a success. It fulfils every condition of acceptability we have stated. But, not surprisingly, it has other drawbacks, of which the following are the most important.[38]

Although WUN avoids the Absurdity and thereby forbids that we make well-off people better-off by making other lives miserable, one might object that it does not take a proper interest in the welfare of *parts* of lives. For instance, WUN does not prevent that the welfare of one part of a life with negative lifetime value is sacrificed for the welfare of another part of another person who also has a negative lifetime value. As an illustration, consider a case where a doctor in a refugee camp is trapped in a dilemma. A group of refugees are ill and their sickness gives them periods of great pain. These periods are of equal length. The doctor knows that one of these periods is just about to begin. Suppose he has now two options. One, he gives the medicine to one person *P* and the rest is left without. *P*'s pains are hard to eliminate, so even if he is given the medicine he will still suffer. Further, assume that if this option is chosen then *n* persons will each have a period consisting of moments with the total value -6. Two, he gives medicine to all but *P* so that all but *P* are somewhat better-off as compared with option two, but still suffer. Furthermore, the unlucky person *P* suffers here horrendous pains. If this option is chosen then *n*-1 persons, the winners, will each have a period consisting of moments with the value -5, but at the same time *P*, the loser, will have a hellish period consisting of moments with the total value -100. Imagine that all the refugees have lives with negative value, but that the winners would have a better life than the loser irrespective of which option that would be chosen. Since all of these people have negative lifetime value WUN does not hinder that the loser's loss is compensated by the winners' gains. For any great loss for the loser, if the winners are of a sufficient great number than the compensation is a fact.

In this example a loss for a person who is worse off on the whole is compensated by gains for persons who are better-off on the whole but still unhappy. But maybe our intuitions on compensations are also such that we do not want to allow that a period in one life is totally ruined for the sake of making other periods in other lives just noticeable better-off, *irrespective* of the persons' lifetime values. Even if one person will lead a good life on the whole, we are not allowed to, for instance, make his childhood a hell, for the sake of making other periods in other lives slightly better-off. WUN cannot capture this intuition since uncompensated losses for a person are defined in terms of lifetime values.

[38] Another important drawback , an anti-egalitarian consequence, is spelled out in Ch. 4, section 6.4.

Admittedly, WUN is in both of these respects inadequate, but we think that a theory that gives some special concern to periods must nevertheless use a value function similar to the one used in WUN. That is, in the same way that WUN dampens positive lifetime values and gives greater weight to negative lifetime values, the period oriented theory must dampen positive *period values* and give greater weight to negative *period values*. This means that even if WUN is too lifetime oriented, it incorporates a general idea on how to weigh positive values against negative ones, which may be used in other contexts where the focus is on other units of a life.

Furthermore, even if we want to give more weight to periods of different persons' lives, it seems to be wrong to give *all* weight to periods and completely abandon the lifetime perspective. To give all weight to periods means that instead of moving from moment value to lifetime value, we move from moment value to period value. Each life is divided into periods and the value of a period is a function of the value of the moments within that period. The value of a world is then seen as a function of the value of the periods in that world. Hence, the theory does not discriminate between aggregation of values of different periods from the *same* person's life and aggregation of values of different periods from *different* persons lives. This means that *intra*personal interperiodical compensations are treated exactly in the same way as *inter*personal interperiodical compensations. The axiological importance of personal identity over time is restricted to the definition of periods, since a period is a set of successive moments within the life of one and the same person. We think that this impersonal feature makes this theory counter-intuitive.

If this reasoning is sound then we are left with the problem of finding a theory that combines the periodical perspective with the lifetime perspective. This is not a simple task since it is very hard to find out the proper way of defining the relevant periods. Should the relevant period be defined in temporal terms, for instance as one day, a week or a month? Or should it be defined in terms of some sort of nearness or connectedness that relates the set of successive moments that constitute a relevant period? We leave these problems unsolved.

Chapter 4

MORAL DUTIES TO FUTURE GENERATIONS

1. Introduction

The general problem of future generations is, evidently, the problem of our responsibility for those who come after us. There has been no generation in the history of humankind more able to affect subsequent generations for good or for ill than ours; a swift glance at the development of the last century makes it appear that what we might very well bequeath to our successors are environmental pollution, the destruction of habitats and species, overpopulation coupled with the depletion of natural resources, global warming, and nuclear waste dumped on land and at sea. We are profiting on the earth's resources at the expense of our successors.

On the other hand, it is tempting, when we look at human history from our present vantage point, to assume that generations to come will be better-off than we are, just as we are better-off than the generations before us. Therefore, it does not matter whether we do things now that will make the environment less hospitable for future generations, as they will be better-off than we are anyhow. It is only fair that we should have benefits at some cost to them rather than the other way around. Perhaps there will be no oil left in a hundred years but only a lot of drums with nuclear waste. By then, however, there might be technology to solve these problems, such as some safe way of reusing nuclear waste.

One way of approaching the problem of our responsibilities for future generations is by analogy with our responsibilities for our contemporaries in other parts of the world; that is, to treat distance in time as we treat distance in space.[1] Just as we, living in Sweden, are responsible for any (foreseeable) sufferings in Africa, which may result from our actions or inactions, and are under some obligation not only to prevent suffering but also to ensure that those other people at least have a decent living standard; so are we responsible for any future suffering that may result from our actions or inactions. Consequently, we are under some obligation to prevent that suffering from occurring and to ensure, to the best of our abilities, that subsequent people's lives are at a decent level of welfare. The questions pertaining to what extent we are responsible for such sufferings are pressing and familiar issues that different moral theories have different answers to. These questions will be discussed here although we shall focus more on problems that are specific to future generations, problems that do not occur when we discuss obligations to our contemporaries.

On one view, remoteness in time has, in itself, moral significance (cf. example 3, Chapter 1). A common concept in welfare economics and cost-benefit analysis is the

[1] This is a common analogy, see for example Locke (1987).

"social discount rate." On this view, we can discount effects of acts and policies at some rate *n* per cent per year. This view is discussed and subsequently refuted in section 2, "Social discount rates."

We can distinguish three kinds of social choice or policy options: *Same People Choices, Same Number Choices* and *Different Number Choices*.[2] The first kind of choice does not affect the number of persons, nor the identity of persons. Our moral problem is then restricted to how we should distribute utility and disutility among a given group of people. Most social theorists, unwarrantedly, presuppose this kind of choice.

Major social decisions that affect the welfare for future generations are also likely to affect who will exist. The pivotal but very plausible claims are that the identity of a person is dependent on the timing of his conception and that the implementation of a social policy must, by a number of perhaps minor but widespread and cumulative effects on people's lives, affect, possibly in a purely accidental way, who has intercourse with whom and when. Major social decisions can be Same Number or Different Number Choices; the former kind of decision affects the identities but not the number of future people, the latter kind affects the number, and hence also the identities, of future people.

It follows that if a social policy is put into effect, there will exist people who would not have existed if the policy had not been adopted; after several generations it is likely that there will be no one alive who would have existed otherwise. Suppose that a policy has as one of its effects that the lives which will be lived at a future date will be substantially less worth living than the lives which would have been lived, had the policy not been carried out. This is like example four in chapter 1, "Introduction," where we had to evaluate the implementation of an energy policy which would raise the welfare in the close future but lower it in the further future. The people whose lives will be of this relatively poor quality cannot complain that they have been harmed, or that the choice was against their interest, or that they are worse off than they might have been had another policy been chosen, for if another policy had been adopted they would not have existed at all. Thus, social policies which reduce the welfare of future people cannot be criticised on the ground that they violate rights, harm or are against the interest of the particular people who will live in the future since the people caused to be badly off will have lives worth living and would not have existed had the policy not been carried out.

This problem has been called the Non-Identity Problem and has many implications for normative theory. It seems to exclude all ethical theories that hold that we should maximise the good effects and minimise the bad effects that our actions have *on specific people,* i.e., the *Person Affecting Restriction*. This is of great import, since some writers have argued that with a Person Affecting Restriction one would avoid certain undesired implications of Total Utilitarianism.[3] According to Total Utilitarianism, what matters is the total amount of utility in the world. It follows that it could be better to expand a population even if everyone in the resulting population would be worse off than in the original population, and that it

[2] See Parfit (1984), p. 356 and Ch. 2, section 2.2.

[3] Most prominently Jan Narveson (1967, 1973, 1978) and, more recently, Partha Dasgupta (1993, 1994).

normally will be wrong for a person to remain childless, i.e., we have a moral obligation to create a lot of happy children. Indeed, a very huge population, say of a hundred billions, with a very low level of welfare, could be considered better than a population of say, five billions, with a considerably higher level of welfare. The latter implication of Total Utilitarianism has been called the "Repugnant Conclusion," a term coined by Derek Parfit. Jan Narveson has argued that this conclusion could be avoided if we impose a Person Affecting Restriction: An act or policy A is better than an act or policy B if and only if it is better for some specific persons, and we do not benefit a person by bringing her into existence. Cases involving the Non-Identity Problem, however, cannot be solved by this kind of comparative principles because no or only some persons will exist in both alternatives. This has led several authors to dismiss the Person Affecting Restriction altogether and to hold that only impersonal principles can be applied to the problems of future generation. In section 5, "The Person Affecting Restriction," we shall discuss this problem and contend that a Person Affecting Restriction creates more problems than it solves.

The bulk of this chapter will be devoted to finding an impersonal principle that can act as a guideline when evaluating acts and policies that affects our successors. As argued above, social policies that affect future generations' welfare will also affect their identities, and after several generations it is likely that there will be no one alive who would have existed, had not the policy been set up in the first place. In evaluating these policies we can only use impersonal principles. It can also be the case that only the identity of a part of the population is affected by a choice, i.e., *nonexclusive* Same Number and Different Number Choices. Even if one thinks that person affecting principles can evaluate the effects on people whose identity is not affected, we are going to need an impersonal principle to evaluate the effects on people whose identities change. The most desirable solution would be an impersonal principle that could evaluate all the effects of an act. When dealing with exclusive Different Number Choices, or when dealing with nonexclusive Different Number Choices and dismissing any use of the Person Affecting Restriction, alternatives can vary in three axiologically relevant aspects: population size, total quantity of welfare, and individual quantity of welfare. The questions we have to answer are: Does a population always get better when the quantity of welfare rises? Will that be the case even when the rise in the total quantity of welfare is accompanied by a decrease in individual quantity of welfare? Or is it the other way around: no amount of extra quantity of welfare can outweigh decrease in the individual quantity of welfare? Does the value of a population always rise when we add people with positive welfare, who will not affect other people's welfare? Even when the added people's welfare is far below the welfare of the original population? What is the optimal size of a population? Is it determinable? If so, how is it related to the average welfare of the population? Thus, generally put, our problem will be to find the best composition of population size, individual quantity and total quantity of welfare. These are very theoretical questions, but they have to be answered before we can give any reasonable answer to more concrete questions, such as "Is it ethically acceptable to leave nuclear waste to future generations to deal with?", "Can our destruction of certain environments be compensated by the technological capital we develop?" Our answers to concrete questions will depend on what kind of answer we give to the more theoretical and abstract questions on how to balance population size, total quantity and individual quantity of welfare.

In section 4, "Linear Values Theories," we investigate theories that attribute linear increasing value to total quantity of welfare and/or individual quantity of welfare. Section 6, "Variable Value Principles" deals with theories that assign linear increasing value to individual quantity of welfare and asymptotically increasing value to total quantity of welfare.

In section 3, "Representation and Specification of Alternatives," we shall discuss how we can best depict possible alternatives in a choice situation and the assumptions and definitions used in this chapter.

To sum up, when we have Same People Choices, the only way future people are different from our contemporaries is by virtue of their being remote in time. Social discount theories assert that this fact has moral importance. We shall criticise this view. When we have Same Number Choices, some or all future people differ from our contemporaries by having contingent identities, hence the Non-Identity Problem. If we do not think that this fact has moral importance, then we have to develop an impersonal theory of beneficence. Finally, in Different Number Choices, the population size of future people can vary, and this fact forces us to find a reasonable principle for how we should balance total quantity and individual quantity of welfare.

2. Social Discount Rates[4]

A common method applied in welfare economics and cost-benefit analysis is to discount the more remote effects of acts and social policies at some rate *n* per cent per year; two commonly employed rates are 5 per cent and 10 per cent. To justify discounting, we need to find some argument that can convince us that time, in itself, has moral significance or that another feature, which correlates with time, has moral significance.

To avoid confusion, we must distinguish between social discount rates that are applied to benefits and losses measured in monetary terms, on the assumption that there will be inflation, and social discount rates applied to the actual utility (welfare) that will be enjoyed by future people. We do not question the former kind of social discount rate, but the latter. For example, it has been seriously suggested that, when evaluating the risk of the disposal of nuclear wastes, we should apply a social discount rate to future *deaths* and *injuries*. Let us look at the six most intelligible arguments.

> *The argument from democracy*: Many people care less about the further future. If this is true of most adult citizens of a democratic country, then this country's government ought to employ a social discount rate. The government's decisions should "reflect only the preferences of present individuals,"[5] and failure to do so would be paternalistic, authoritarian or anti-democratic.

[4] For the arguments in this section we are indebted to Parfit (1984), appendix F.

[5] Marglin (1963) as quoted in Parfit (1984), p. 480.

This argument conflates two different questions:

(1) Are we morally justified in applying a social discount rate?

(2) If most people in a community answer yes to question (1), is a government justified in overriding this majority view?

The argument from democracy applies only to the latter question and is irrelevant to the former question, which is our concern. The only way a person's commitment to democracy can give him the answer to question (1) is that he assumes that what the present majority wants, or believes is right, *must be right*. But that is, of course, a bad argument: Even if it were the case that most Germans in the thirties thought it was right to exterminate Jews, it does not follow that they were morally justified in doing so.[6] Whatever most of us believe about social discount rates, the moral question remains open. Rather, this points to a disturbing problem for democracy. Most of us believe that social policies carried out by a government elected by informed adults would be an implementation in good democratic order. But when we reflect on the fact that most people affected by the policy, future people, were not able to have their voice heard in the election, then we may start doubt that the process was really democratic.[7]

> *The argument from probability:* More remote effects of acts or social policies are less likely to occur and that is a reason for discounting these effects.

As above, there are two different questions:

(1) Is a prediction about effects in the further future less likely to be correct?

(2) If a prediction is correct, may we give it less weight because it applies to the further future?

The answer to question (1) is often yes but that provides no reason for an affirmative answer to (2). Let us take the example of nuclear energy. When considering possible accidents, we must think far into the future, since some radioactive elements remain dangerous for many centuries. With a social discount rate of 5%, one death next year counts for more than a billion deaths in 400 years. Hence, it would be morally worse if an energy policy would cause one person's

[6] In the thirties, if all people in Germany, except the Jews, thought that the extermination of Jews was a good idea, or if it was the case that the majority's preferences to exterminate Jews were stronger than the Jews preferences not to be exterminated, then this act could be morally justified on some preferentialist theories. But such a a conclusion would be a good argument against such kinds of moral theories.

[7] The same problem occurs when social policies affect people in nations other than the nation carrying out the policy. This is often the case with environmental problems, and especially countries in the third world have considerably little to say about reforms that affect them in a substantial way.

death next year rather than a billion deaths in a distant future - quite an abominable conclusion one might say.

The argument from probability does not lead to this conclusion but we could argue that we ought to discount those predictions of effects that are more likely to be false. This view would not distinguish between effects that occur in the further future or in our time, so it is not a discount principle based on time. We could call this a *probabilistic discount rate,* as opposed to a *temporal discount rate* which is based on time. Predictions about the distant future are more likely to be false. We could therefore argue that the temporal and probabilistic discount rate correlate. But this is not really the case. Predictions about the further future are not less likely to be true at some rate of *n* percent per year. Whether predictions are likely to be true or not will differ from case to case and in many cases predictions about the further future are as likely, or more likely to be true than predictions we make about the near future. Moreover, using temporal discount rates misstates our moral view. It makes us claim that more remote bad consequences are less important rather than less likely to occur.

> *The argument from opportunity costs*: It is sometimes better to receive a benefit earlier, since this benefit can be used to produce further benefits. The delaying of some benefits thus involves *opportunity costs.*

If an investment yields a return next year, this will be worth more than the same return ten years later if the earlier return can be reinvested profitably over these ten years. The difference is the opportunity cost.

This argument fails in a similar way as the argument from probability. Certain opportunity costs do increase over time but whether opportunity costs will rise, and at what rate, will differ from case to case and does not correlate with time. For example, benefits which are consumed cannot be reinvested. Suppose we have to decide whether we should exploit a stretch of beautiful nature. When evaluating the benefit of enjoying this natural beauty according to a social discount rate, the benefits in later years count for much less than the benefit next year. This cannot possibly follow from the argument from opportunity costs, since the enjoyment of nature cannot be profitably reinvested.

> *The argument that our successors will be better-off*: Perhaps future generations will be better-off than we are now. We could then appeal to two arguments for discounting benefits and costs that we give to and impose on them. If we measure the benefits and costs in monetary terms, adjusted for future inflation, we can appeal to the diminishing marginal utility of money - the same increase in wealth generally brings a smaller benefit to those who are better-off. We may also appeal to some distributive principle - an equally great benefit given to those who are better-off may be claimed to be morally less important.

These are good arguments but not for a social discount rate. Our arguments for discounting the future benefits are not based on the fact that they appear further in the future, but that they will be enjoyed by people who are better-off than we are. Here, as above, there will not be a correlation with time. Even if we were justified in assuming that our successors are going to be better-off than we are, it is still unrealistic to believe that future generations' welfare will increase by a certain percentage every year. Indeed, as history shows us, later generations have sometimes been worse off than their predecessors.

> *The argument from excessive sacrifice*: If we do not apply a social discount rate to future benefits and costs, then any small increase in benefits that extend far into the future might demand any amount of sacrifice in the present. In time the benefits would outweigh the costs.

The same objections as above apply. This is not an argument for a social discount rate but an argument that no generation can be morally required to make more than a certain amount of sacrifices for the sake of future generations. If this is what we believe then this is what should influence our decisions. It does not follow that we ought to give less moral weight to people in the further future than to people in the closer future. Suppose that, at the same cost to ourselves now, not involving any excessive sacrifice, we could prevent a minor catastrophe in the nearer future or a major catastrophe in the further future. Adopting a social discount rate would imply that the greater catastrophe is less worth preventing.

> *The argument from special relations*: According to common-sense morality we may or should give more weight to people to whom we have special relations than to strangers. Thus we are morally permitted to give some kind of priority to our own interests, to our families, to our friends, to our patients, to fellow-citizens, and so forth.

This view cannot support discounting based on time, but a new kind of discounting, discounting on the basis of degrees of kinship. This new kind of social discount rate could perhaps correlate with time. We may think that we ought to give our children's welfare special weight, and the same could hold for our grandchildren, though to a reduced degree, and so on. But this is not so. A discount rate with respect to kinship should at some point cease to apply or at least reach a constant level - how could the argument from special relations give us reason to give less weight to people living in the year 2300 rather than 2200?

The argument above would also hold for people spatially remote from us. This seems to go along well with our moral inclinations; we may think that the U.S. Government is justified in giving more priority to the welfare of its own citizens. But this reasoning does not apply when it comes to grave harms. Suppose the U.S. Government decides to resume atmospheric nuclear tests and predicts that the fallout would cause several deaths. Would it then be morally better to perform the tests in the Indian Ocean rather than on American soil just because the people living in the Indian Ocean are strangers to the Americans?

We have discussed six arguments for the social discount rate. Remoteness in time correlates with a whole range of morally important facts, as does remoteness in space. None of these correlations are of such a nature that they can justify that we should care less about the effects our acts or social policies have in the future or at long-range, at some rate *n* percent per year or per meter.

When other arguments do not apply, we ought to be equally concerned about the predictable effects of our acts or social policies irrespective of when they occur.

This is of great moral importance. As noted in the introduction, there surely has been no generation in the history of humanity more able to affect subsequent generations than ours. Nuclear waste may be dangerous for thousands of years; global warming can radically change the conditions for life on earth, as can increase in insolation of ultraviolet radiation.

3. Representation and Specification of Alternatives

3.1. Schemes and Specifications

Parfit depicts the Repugnant Conclusion as shown below:

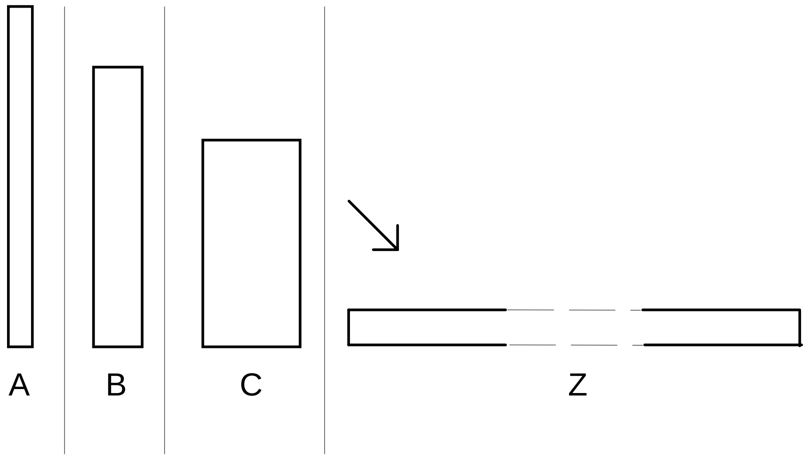


Figure 3-1. The Repugnant Conclusion (after Parfit 1984).

Parfit explains the diagram:

> The width of each block shows the number of people living, the height shows their quality of life. By this I mean their quality of life *throughout some period*. In such a period there would be *some change in the population*. But for simplicity, we can ignore this fact. For the same reason, we can assume that in these outcomes there is neither social nor natural inequality; *no one is worse off than anyone else*. This would never in fact be true. But it cannot distort our reasoning, on the questions I shall ask, if we imagine that it would be true. And this makes my questions take a clearer form (emphasis added).[8]
>
> *The Repugnant Conclusion*: For any possible population of at least ten billion people, all with a *very high quality* of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better, even though its members have lives that are *barely worth living* (emphasis added).[9]

This explanation leaves important questions unanswered: What is the relationship between these time slices? Do they belong to the same world or to different worlds, i.e., are they *real* alternatives to each other or not? Do we compare whole worlds, the futures of worlds, or do we only try to determine the intrinsic value of different time slices within worlds? What are the relative differences between people's individual quantity of welfare in these examples, how much worse is a "life barely worth living" than a "very high quality of life"? What do these concepts really mean, how can we specify a "life of very high quality"?

To properly understand the Repugnant Conclusion and other population problems we must be clear about what kind of parts, wholes and values we are

[8] Parfit (1984), p. 385.

[9] Parfit (1984), p. 388.

speaking about, what kind of comparisons we are working with, and the relation between the compared alternatives. We need a representation of the alternatives that includes all relevant information and, when we make simplifications, we must be clear about what kind of assumptions these rest on. Furthermore, we need some realistic specifications of what the alternatives consist in. In the following, we shall construct a way of representing these alternatives that will both represent the theoretical properties of the different problems, such as the relation between the different alternatives, the relative population size and the welfare level, and more concrete properties such as specific population sizes and absolute welfare levels. As noted in Chapter 2, "Presuppositions," the diagrams give abstract information about the utilities, disutilities and sizes of the populations. The diagram constitutes an abstract case or a scheme. When we picture a special choice situation by giving specific numbers and qualities of life, and sometimes a short story how these populations come to differ in welfare, we present a *specification* of an abstract case. It is important to remember that to the same scheme many different specifications can be given, none of the arguments below hinge on the particular specifications we have chosen. We could have used another specification, perhaps more in line with the reader's considered beliefs, without changing any of the reasoning below and the conclusions reached. The specifications make the examples more concrete and works as fixed points in the discussion.

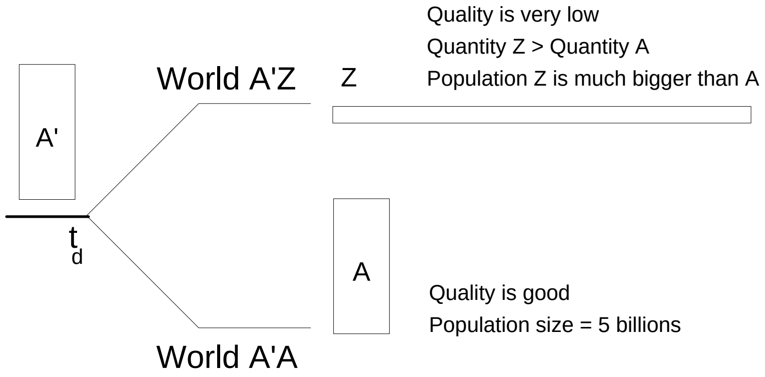


Figure 3-2 An inter-world representation of the Repugnant Conclusion

This representation is more comprehensive. A', A and Z are different time slices and t_d is the time of a decision or an event that makes the world branch. Here we can see that A and Z are exclusive alternatives to each other; it will either be the case that the time slice Z occurs or that the time slice A occurs. If A occurs, then we can say that the world A'A obtains, i.e., a world obtains which among its time slices will have the time slices A' and A. In the same manner, if Z occurs, then the world A'Z obtains. This means that we make *inter-world comparisons* in contrast with comparisons between slices of one and the same world, *intra-world comparisons*. The reason for only making inter-world comparisons is simple: Time slices from the same world will never be alternatives to each other, we are never going to have a choice between two slices from the same world. Consequently, one can doubt why the way we value time slices from the same world should have any impact on an ethical theory to be used for evaluating alternative actions or social policies.[10]

As discussed in Chapter 2, "Presuppositions," a typical choice situation can be represented as a set of branching possible worlds unified by a common node in a world tree. The choice situation at t_d can consist of individual choices, political decisions and so forth, but also natural events, e.g., environmental catastrophes, earthquakes or a plague. A choice does not necessarily need to have its axiologically relevant effects at the branching node; on the contrary, the effects can take place anytime in the future. A decision, by itself being a state of affairs, makes the world branch.

Following Parfit, the blocks represent the aggregated welfare of a time slice of a world. The width of these blocks represents the population size and the height represents the individual quantity of welfare in the time slice. In the cases we are going to consider in this chapter, differences between the welfare of specific individuals do not matter. Thus, talk about the individual quantity of welfare amounts to the same thing as talk about the average individual quantity of welfare. For short, we shall sometimes call this the *quality of welfare* or the *average welfare,* while the *quantity of welfare* or the *total welfare* refers to the total quantity of welfare. With "welfare," we mean the part of a person lifetime value that obtains in a time slice, i.e., her period value, as developed in Chapter 3.[11] Thus, the quantity of welfare is a mere sum of the period values that obtains in a time slice, and the quality of welfare is this average period value per person in a time slice.

Like Parfit, we assume that there are no egalitarian reasons to evaluate these alternatives differently. This can be done in the same way as Parfit did above, by presupposing that all people are equally well-off. Another way is to assume that when there are inequalities, these are of a magnitude that will not affect our evaluation of the alternatives or that the negative axiological values of the inequalities are the same in both alternatives and can consequently not be the reason for ranking these alternatives differently.

In the case above, people's welfare is positive, which means that this part of life is satisfactory for all persons, i.e., the intrapersonal aggregation of utility yields a positive result. This is not to say, as we mentioned in Chapter 2, "Presuppositions," that life during this period was worth living, subjectively or objectively. Whatever (reasonable) specific weight we give utility and disutility, we can construct cases such as the one represented in figure 3-2.[12] Indeed, to make things easy, one can simply assume that there is only positive utility in the case above. For example, the reason why a person has a very low quality of welfare could either be that there are only enough ecstasies to just outweigh the agonies, or that the good things in life are

[10] Furthermore, intransitive value orderings that occur with some axiological principles when making intra-world comparisons, do not arise when we transform these cases to similar types of inter-world comparisons. This is, for example, the case with Temkin's claim that all person affecting principles are intransitive. See Arrhenius (1992) and Temkin (1986).

[11] The arguments put forward in this chapter are not, however, dependent on the kind of intrapersonal aggregation we argued for in Ch. 3. These arguments hold equally well for theories that make use of a cruder intrapersonal aggregation function such as, for example, a mere totting up of utilities.

[12] For a more detailed analysis of these matters, see Ch. 3.

of uniformly poor quality, e.g., working at an assembly line, eating only potatoes and listening to Muzak.

Let us define and specify some welfare-concepts. When *quality is good*, life could be like the life of the average person in western societies; a life of *very high quality* could be like the life led by the best-off people living today. *Perfect quality* could be a condition where all our ideas of what constitutes welfare are fulfilled; there exists no way to raise the welfare of such people. A life of *very low quality* could be like the life led by unemployed people in Europe.[13] One can have different ideas about how much better "perfect quality" is than "good quality" and so forth . As we shall show below, in section 4, the answer one gives to that question will not make a difference to the solution to the problems discussed in this chapter.

These exemplifications are not indisputable, we can have different opinions about the quality of life of the average person in western societies or the unemployed European, but these descriptions are indeed only examples which works as fixed points in the discussion. Anyone who believes in other specifications of very low, good, very high and perfect quality can fill them in. All that we need to assume is that the quality of people's lives can vary - this is surely a reasonable assumption, one that both matches our intuitions and common language use.

3.2. The Demarcation of a Population

An intuition that underlies much of the discussion of future generations and population problems is that at a certain population size with a certain quality of welfare we cannot make a population better by considerably reducing the original people's quality of welfare but increasing the quantity of welfare by adding more people with positive welfare. Parfit, for example, states that this is the case in the world today with its five billion inhabitants.[14] Obviously, Parfit here thinks that a population is demarcated in time, that is, a number of people living during some period of time not all that long. Just think of the number of people that has lived during the 20:th century - that is definitely much more than five billions. One can argue that the concept "population" cannot be combined with all kinds of time slices because it is concerned with *people located in the world at the same time*. We can only speak about the size of a population at a specific time or a period of time that is not too long. For example, if we took a time slice of two centuries length, counted all the people and then concluded that the population size during this period was, say, thirty billion people, then that would surely not be in line with the common language use of the word "population." The intuitions we have in population question are tied to a concept of population where people have some kind of contemporaneity.

This may seem like a simple point, but it is of vital importance when using value functions where the population size is one of the arguments. Average Utilitarianism, consequently, yields different results with different definitions of

[13] See Ch. 2, section 1.5, for a more detailed discussion of representation of welfare. Recall that these concepts are defined in relation to a certain context, in our case the actual world. There could be other logically possible world where quality of life similar to the best-off people in our world would be looked upon as a very low quality of life.

[14] Parfit (1984), p. 402.

population size. Consider two outcomes, where in the first there would be one generation of people with very high quality succeeded by many generations with good quality. In the other outcome, only the first generation would exist. Average Utilitarianism seems to imply that it would be bad if all these generations with good quality came into existence. This conclusion only follows, however, if one makes use of a population concept which counts *all people* or *all future people that will ever live*. Moreover, there is nothing intrinsic in the Average Utilitarian Principle that forces us to adopt such a population concept; on the contrary, as we saw above, more in line with ordinary language use is a concept which stresses contemporaneity. If we want to use the average principle or other population sensitive principle as a population principle, then we have to make use of time slices of a length such that the population size is well defined like, for example, one year. Then, to decide whether the existence of the generations with good quality would make the world better, we can calculate the worlds average utility for every year and then sum these averages.[15] This would establish that the existence of these extra generations with good quality would be a good thing.

The reason that Average Utilitarianism yields different results with different population concepts is that the average depends on the size of the base upon which one calculates. This has led to much confusion about the average principle and most of the arguments launched against *the* Average Utilitarian Principle have in fact been arguments against different ways of establishing the denominator. These arguments can also be used against Variable Value Principles which also use population size as one of its variables when calculating the value of different outcomes. We shall therefore discuss these different arguments and other demarcation problems in further detail in the section dealing with Variable Value Principles, section 6. For the present purpose, it is enough to observe that our common language use and our intuitions are tied to a population size concept that leads us to restrict the length of time slices in order that the population will remain fixed.

How short should such a time slice be? It seems that the length should be as short as possible for it to be no population change in the time slice, i.e., as short as one welfare moment. However, that would be unworkable: The required information is simply unobtainable and our examples would be very hard to survey. We can use longer time slices as long as the changes in population size which take place are insignificant to the problem in question; we can then assume that there are no population changes during the given time slice or rather, that they are insignificantly small as compared to the great difference in population size between the compared slices. In this chapter, where we are going to discuss long-term effects of different policies, a ten-year time slice will be assumed in the specifications of different population problems.

When representing different time slices we would actually need to draw three-dimensional boxes, because we have three variables: population size, quality and time. By keeping the length of the time slices constant, we can stick to the two dimensional boxes.

[15] The way to sum up these averages need not necessarily be a mere totalling as we shall see in section 6.

4. Linear Value Theories

4.1. The Repugnant and the Reversed Repugnant Conclusion

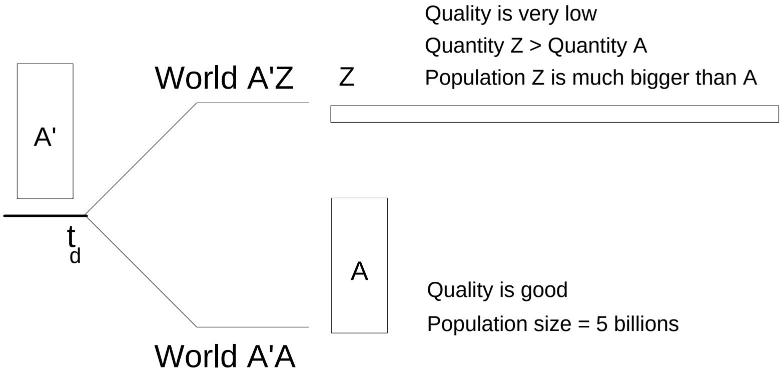


Figure 4.1-1. The Repugnant Conclusion

> *The Repugnant Conclusion*: For any possible population of at least five billion people, all with a good quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better, even though its members would have lives with a very low quality of life.[16]

The Repugnant Conclusion follows from the belief that a loss in quality of welfare always can be compensated by an increase in quantity of welfare. Total Utilitarianism is the paradigm principle that implies this because it only ascribes value to quantity. Derek Parfit has recently called attention to this implication, which is thoroughly discussed in his book *Reason and Person,* but it was already noted by Henry Sidgwick in 1907:

> Assuming, then, that the average happiness of human beings is a positive quantity, it seems clear that, supposing the average happiness enjoyed remains undiminished, Utilitarianism directs us to make the number enjoying it as great as possible. But if we foresee as possible that an increase in numbers will be accompanied by a decrease in average happiness or vice versa, a point arises which has not only never been formally noticed, but which seems to have been substantially overlooked by many Utilitarians. For if we take Utilitarianism to prescribe, as the ultimate end of action, happiness on the whole, and not any individual's happiness, unless considered as an element of the whole, it would follow that, if the additional population enjoy on the whole positive happiness, we ought to weigh the amount of happiness gained by the extra number against the amount lost by the remainder. So that, strictly conceived, the point up to which, on Utilitarian principles, population ought to be encouraged to increase, is not that at which average happiness is the greatest possible,--as appears to be often assumed by political economists of the school of Malthus--but that at which the product formed by multiplying the number of persons living into the amount of average happiness reaches its maximum.[17]

Indeed, this seems also to be the intent of William Whewell's argument from 1852, that if quantity of pleasure in the effects is the test of conduct, then Jeremy

[16] Compare Parfit's formulation in section 3.

[17] Sidgwick (1907) Bk. 4 Ch. 1 Sec. 2 Para. 4/6 p. 415.

Bentham's Greatest Happiness Principle should become the Greatest Animal Happiness Principle, and it would be our duty to sacrifice the happiness of human beings "provided we can in that way produce an overplus of pleasure for cats, dogs and hogs, not to say lice or fleas."[18] To avoid this, John Stuart Mill constructed a distinction between higher and lower pleasures, captured in the famous phrase "It is better to be a human being dissatisfied than a pig satisfied; better to be Socrates dissatisfied than a fool satisfied."[19] We could hold that lives or periods of lives with higher welfare contribute more to a populations value than the sheer quantity of welfare they contain, the quality of lives should also matter in our calculations.[20] We might treat higher-quality lives as ten times better than lower-quality lives even though the quantity the high-quality lives contain is, say, only five times as much welfare as the lower-quality lives.[21] But what is important to observe is that on this view it would always be possible for some number of lower-quality lives to have greater aggregated value than a given number of higher-quality lives. Consequently, every principle that ascribes linear increasing value to both quantity and quality and do not put an upper limit to these values implies the Repugnant Conclusion. This will be the case for every possible weighing of quantity and quality as far as the weighing constant is a finite positive real number. This is more easy to see if we put it in more technical terms. The general value function, *V*, that gives linear increasing value to quantity and/or quality, where *Q* stands for quality, *P* for population size, and *K* stands for quantity, would thus be

$$\begin{aligned} V &= v_1Q+v_2K \qquad\qquad v_1, v_2 \geq 0;\ \neg(v_1=v_2=0) \\ &= v_1K/P+v_2K \\ &= v_1Q+v_2PQ \end{aligned}$$

where v_1 and v_2 are the weighing constants. The function *V* is a linear increasing function of *K* and *Q*.[22] We shall call this way of weighing quality and quantity for a *linear weighing*.

If we assign $v_1=0$ and $v_2=1$ we get the ordinary Total Utilitarian Principle and with the reverse assignments we get the Average Utilitarian Principle. It is now easy to see that for any value of v_1 and v_2 greater than zero there always exist a value of *K* that would outweigh the value of *Q*. For example, with $v_1=2$ and $v_2=1$, a population with two persons with quality *q* would have the value $V=4q$ which equals the value of a population with six persons with quality $q/2$. The only consequence of giving more weight to quality is that there must be a larger population in the Z-slice.

[18] Whewell (1852), quoted in Mill (1852) and in Acton (1987). Mill, in quoting, omits Whewell's concluding phrase: "not to say lice or fleas".

[19] Mill (1865), p. 10.

[20] This is Parfit's view, see Parfit (1984), p. 402.

[21] This seems to be the idea of George Sher's interpretation of John Stuart Mill's distinction between the quality and quantity of pleasures (Sher 1979, pp. xii-xii), as Lemos (1993) points out. I agree with Lemos that this is a doubtful interpretation of Mill's thesis. Cf. also Feldman (1978, pp. 30-6) for a similar idea.

[22] *K* is a one dimensional linear increasing function of welfare; *Q* is a two dimensional function of welfare and population size, linearly increasing with welfare and linearly decreasing with population size. Neither *K* nor *Q* has any upper boundary in logically possible worlds and, consequently, the same holds for *V*.

Perhaps more surprising, we can see that the converse is also true, that is, for every weighing there always will be a value of *Q* that would outweigh *K*. In the example above, a population with one person with quality *4/3q* would have the value *4q*.

If we put very low value on quantity, then we get the following conclusion:

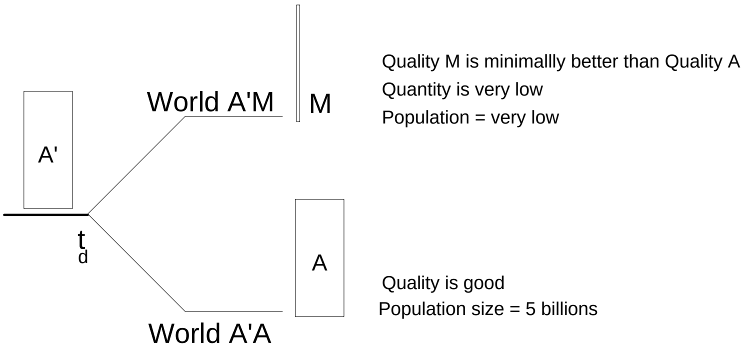


Figure 4.1-2 A Reversed Repugnant Conclusion - A minimal quality improvement case (quality can always outweigh quantity).

> *The Minimal Quality Improvement Reversed Repugnant Conclusion*: For any possible population of at least five billion people, all with a good quality of life, there must be some imaginable population with just slightly better quality, whose existence, if other things are equal, would be better, even though this population is very much smaller and, consequently, the total quantity is very much lower.

Here a very low population with just slightly better quality outweighs a much larger population. The extreme case is where just *one person* with minimally better quality outweighs an arbitrary large population. This is implied by the Average Utilitarian Principle, the most popular principle among economists.[23] According to this principle, it is worse if there is a lower average quality of life, per life lived, that is, all value is put on quality, no value at all is put on quantity. Consequently, an arbitrarily small increase in quality can outweigh an arbitrarily large decrease in quantity.[24]

The more value we put on quantity the better the quality must be for the single person in A or the larger must the population be in A. Even if we put very low value on quality (that is, very high value on quantity), we can end up with a Reversed Repugnant Conclusion.

[23] See, for example, Samuelson (1970), p. 551, who makes this principle true by definition.

[24] Surprisingly enough, even more radical principles than Average Utilitarianism has been proposed in population ethics. Fehige (1992) promotes a preferentialistic theory called "anti-frustrationism" where only preference frustrations counts. His theory implies that a world with one person is better than a world with billions of people with the *same* quality of life. The best world is a world with no people at all.

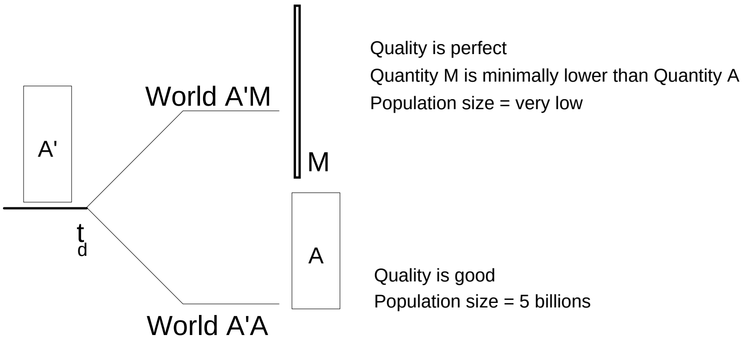

Quality is perfect
Quantity M is minimally lower than Quantity A
Population size = very low
World A'M
A'
M
t
d
A
Quality is good
Population size = 5 billions
World A'A

Figure 4.1-3 A Reversed Repugnant Conclusion - a great quality improvement, minimal quantity loss case.

> *The Quality Improvement Reversed Repugnant Conclusion*: For any possible population of at least five billion people, all with a good quality of life, there must be some imaginable population with much higher quality of life, whose existence, if other things are equal, would be better, even though this population is very much smaller and the total quantity is lower or the same.

If people's lives in M are so ecstatic that there is no loss in quantity, then even Total Utilitarianism could imply a single person Reversed Repugnant Conclusion. That would be tantamount to Nozick's "utility monster" and, as we argued in chapter 2, "Presuppositions," this is not a relevant test for moral principles. But putting an upper boundary to quality will not solve any problem. Let us take an example. In factually possible worlds there cannot be better quality then perfect quality. Assume that a good life is only twice as good as a life of very low quality and half as good as a perfect life. If we give quantity the weight one then quality has to have a weight greater than 10^{10} for A to be better than a Z-population of 20 billions. But with such a weight on quality a population of just one person with perfect quality would be better than A, that is, when avoiding the Repugnant Conclusion we got a Reversed Repugnant Conclusion instead. This is quite perplexing, especially for those who believe that arguing for small differences between different qualities would solve our problem.[25] Now, we could not plausibly argue that a population of twenty billions is a factual impossibility, but we could hold that there are a wider spectra of qualities of life. Assume then that a perfect life is a hundred times better than a good life which is a hundred times better than a life of low quality. This will make no difference as long as there is no upper boundary to the population size in Z - the extra weight on quality must be of the magnitude 10^8 for A to be better than a Z-population of 500

[25] This is perhaps the intent of Tännsjö's treatment of this problem, see Tännsjö (1991), pp. 40-45. He also launches the following argument against the repugnancy of the Repugnant Conclusion: "How we judge the Repugnant Conclusion will in the end and primarily depend upon where we think the level between a life worth living and a life not worth living more precisely should be drawn. Where are we situated in relation to this level? The Repugnant Conclusion will not be especially repugnant if we think that most people normally are quite close to this level and that they indeed often fall below this level." This argument rests on a misunderstanding of what is repugnant with the Repugnant Conclusion: It is not the low quality of life of the Z-people that constitutes the repugnancy, it is the conclusion that whatever good quality a population has, there is always another hypothetical population with lower quality that would be better.

billions, and again, with such a weight on quality a population of just one person with perfect quality would be better than A.

Perhaps we could argue that there is greater difference between a good life and a life with very low quality than between a life of perfect quality and a good life. Assume that a perfect life is ten times better than a good life which is a hundred times better than a life of low quality. For A to be better than a Z-population of 560 billions, the weight of quality must be more than $6*10^8$, and then a population of one person with perfect quality would be better than A.

We cannot escape from the Repugnant and Reversed Repugnant Conclusion by putting an upper boundary to quality and manipulating the differences between perfect, good, and low quality as long as there is no upper boundary to the population size. Could we find a good reason for an upper boundary to the size of a population and would that help us in any way? Observe that there is no factual impossibility to create enormous populations in the universe. Perhaps we could find a non-arbitrary reason for limiting the population in question to the population on earth.[26] But even if we side-step the problem of figuring out the maximum size of a factually possible population on earth, this will not be an attractive solution. We would not get Repugnant and Reversed Repugnant Conclusions, but similar ones. Assume, modestly, that the Z-population cannot be larger than 20 billions. Then we have to give quality a weight over 10^{10} for A to be better than a Z-population of twenty billion with half the quality of A; it follows that a population of just one person with the double quality would be better than B. Those who believe that when we have populations such as A above, we cannot make a population better by considerably reducing the original people's quality of welfare but increasing the quantity of welfare by adding more people with positive welfare, will have a hard time to accept an increase in quantity at the expense of halving the quality.

At any rate, the linear weighing of quality and quantity has another serious flaw, which we shall get back to in section 4.3.

4.2. Higher Goods, Lexical Orderings

The idea of higher goods has a long history in moral philosophy in discussions of the quality of life. The idea has often been that the lack of a certain kind of feature, like knowledge, intelligence, virtue and so forth, can not be outweighed by any amount of pleasure. The *locus classicus* is the famous passage in Philebus where Socrates convince Protarchus that a life of pleasure without memory, intelligence, knowledge, or true opinion is like the life of an oyster and hence not desirable.

> . . . if you had no memory you would necessarily, I imagine, not even remember that you had been enjoying yourself; of the pleasure you encountered at one moment not a vestige of memory would be left at the next. Once more, if you had no true judgement you couldn't even calculate that you would enjoy yourself later on. You would be living the life not of a human being but some sort of sea lung or one of those creatures of the ocean whose bodies are incased in shells.[27]

[26] We shall discuss criteria for spatial demarcations of populations in section 6.

[27] Philebus, p. 21b-c, Hamilton and Cairns, ed. (1985).

Aristotle holds that "No one would choose to live his entire life with the mentality of a child, even if we were to enjoy to the fullest possible extent what children enjoy".[28] Mill, as we saw above, can be interpreted as embracing the notion of higher goods. In what way do the idea of higher good differ from giving weight to both quality and quantity of utility? Franz Brentano puts it quite vividly when he claims that "[i]t is quite possible for there to be a class of goods which could be increased *ad indefinitum* but without exceeding a given finite good".[29] Still, this is open to different interpretations. We could hold that x is a higher good than y exactly if x is intrinsically better than any amount of y. This seems to be the view of W. D. Ross and Jonathan Glover:[30]

> With respect to pleasure and virtue, it seems to me much more likely to be the truth that *no* amount of pleasure is equal to any amount of virtue, that in fact virtue belongs to a higher order of value, beginning at a point higher on the scale of value than pleasure ever reaches. . . [31]

> For we may decide that we value people's lives having various qualities - - - and that the absence of these qualities cannot be compensated for by any numbers of extra worth-while lives without them. There is some analogy with common attitudes to what is valued within a single life. I enjoy eating fish and chips, but no number of extra hours eating fish and chips will compensate me for being deprived of the ability to read.[32]

In our schema of quality and quantity we have already dealt with this conception of higher goods. To assert that any increase in quality is better than any increase in quantity (without a concurrent raise in quality) is tantamount to asserting the Average Utilitarian Principle; to assert that any increase in quantity is better than any increase in quality (without a concurrent raise in quantity) is tantamount to asserting the Total Utilitarian Principle.[33]

Another conception of higher goods is that x is a higher good than y exactly if a greater number of x always are intrinsically better than a smaller number of x but a greater number of y is not always better than a smaller number of y. This would mean that there is a upper limit to the value of a set of y, but not an upper limit to the value of a set of x, and that sometimes a number of lower goods can outweigh a smaller number of higher goods. To take Glover's fish and chips example again, we could think that one hour of reading can be exchanged for eating fish and chips, but there is no amount of fish and chips that can compensate for the loss of all of one's periods of reading. A variant of this position is to say that x is a higher good than y

[28] Nichomachean Ethics, p. 1174a1-4.

[29] Brentano p. 158 (1907), quoted in Lemos (1993).

[30] This is also the way that Lemos interpret the notion of higher goods. As will be obvious when we proceed our investigation, we do not agree that this is the only way to interpret the idea of higher goods.

[31] Ross (1930), p. 150. Cf. Ch. 3, section 5.1.

[32] Glover (1977), p. 71.

[33] Another possible interpretation of this notion of higher goods is a principle that gives lexical priority to quality over quantity. Such a principle would always rank the alternative with the highest quality as the best irrespective of the amount of quantity. When two or more alternatives have the same quality, however, this principle would pick out the alternative with the greatest quantity. This principle would run into the same problems as principles that gives very little weight to quantity of welfare.

exactly if the upper limit for the value of a set x is higher than the upper limit for the value of a set y. Finally, x could be a higher good than y exactly if x always has value independent of the quality of x but y only has value if the quality is above a certain level. This would mean that one could substitute a certain amount of x for a greater amount of y only when y is above a certain quality. One perfect life could equally well be replaced by ten good lives but not by any amount of lives just worth living. For our purpose, we could formulate these views as follows:[34]

> The Valueless Level View: Quantity has no value in lives whose quality is below a certain level.
>
> The Lexical View: There is no limit to the positive value of quantity but no amount of quantity in lives below a certain quality level could be as good as the value of quantity in lives whose quality is above this level.
>
> The Limited Quantity View: The value of quantity has an upper limit.

The Valueless Level View can be interpreted in two ways:

> ***The Valueless Level View 1 (VLV-1)***: Quantity of positive welfare has value but only in lives with quality greater than x. Quantity of negative welfare always has negative value.

$$V = \sum z_i \qquad z_i = \begin{cases} u_i & u_i > x \text{ v } u_i < 0 \\ 0 & 0 \leq u \leq x \end{cases}$$

> ***The Valueless Level View 2 (VLV-2)***: Quality always has value. Quantity of positive welfare has value but only in lives with quality greater than x. Quantity of negative welfare always has negative value.

$$V = v_1 Q + v_2 \sum z_i \qquad z_i = \begin{cases} u_i & u_i > x \text{ v } u_i < 0 \\ 0 & 0 \leq u \leq x \end{cases}$$

Here, u_i stands for the welfare of the person with the index i; x is the limit for the valueless level; Q is the quality, and v_1 and v_2 are the weighing coefficients.

Version one implies a variant of the Repugnant Conclusion similar to the one of Total Utilitarianism with the exception that the level for the Z-people is moved from "barely worth living" to *x*. If the valueless level is low, then this would be unacceptable; on the other hand, if we raise the valueless level, and this raise must be quite high to avoid Repugnant Conclusions, then this level will be less intuitively

[34] See Parfit (1984), pp. 403-17.

acceptable: One person slightly above this level would outweigh an unlimited amount of people with quite good welfare - a Reversed Repugnant Conclusion.[35]

The second version has the same problem as the first one. In addition, this principle mimics Average Utilitarianism when all the people have a quality below the valueless level. Consequently, one person with slightly better welfare can outweigh an arbitrarily great number of people with slightly less welfare - another Reversed Repugnant Conclusion.

The Lexical View falls prey for at least two of the objections we launched against the Valueless Level.[36]

The Lexical View 1 (LW-1)[37]

$$\begin{array}{ll} z_i = u_i & u_i \geq x \\ w_i = u_i & 0 < u_i < x \end{array}$$

$$LW\text{-}1 = (f_1, f_2) \quad f_1 = \sum_{1}^{n} z_i \qquad f_2 = \sum_{1}^{m} w_i$$

(a_1,a_2) is better than (b_1,b_2) iff $(a_1>b_1)$ or $(a_1=b_1$ and $a_2>b_2)$

The Lexical View 2 (LW-2)

$$LW\text{-}2=(g_1,g_2) \qquad \begin{array}{l} g_1 = v_1 f_1/n + v_2 f_1 \\ g_2 = v_1 f_2/m + v_2 f_2 \end{array}$$

(a_1,a_2) is better than (b_1,b_2) iff $(a_1>b_1)$ or $(a_1=b_1$ and $a_2>b_2)$

One person with quality slightly above the lexical level can outweigh an arbitrarily great number of people with quality slightly below the level - a Reversed Repugnant Conclusion. Both above and below the lexical level, quantity can always outweigh quality, and hence we get Repugnant Conclusions. If we raise the lexical level, the less obnoxious Repugnant Conclusions above the level but the more obnoxious Reversed Repugnant Conclusions between the levels.[38]

The Limited Quantity View can be interpreted in two ways:

[35] Further, it implies a version of the *Absurd Conclusion*: It could be the case that it would have been better that a large population of people with quite good welfare and one person with slightly negative welfare, had not existed. See Parfit (1984) p. 410, 415.

[36] We could construct two more versions of the lexical view, LW-3=(f_1,g_2), LW-4=(g_1, f_2), but, as can easily be seen, both of these versions share the problems of LW-1 and LW-2.

[37] We are only defining the Lexical View for populations where everybody has positive welfare. A complete explication of this view would involve a way of aggregating negative welfare. Cf. fn. 38.

[38] Parfit holds that the Lexical View also implies the Absurd Conclusion. He writes: "---The existence of ten billion people below this level would have less value than that of a single person above the Blissful level. If the existence of these people would have less value than of only one such person, its value would be more than outweighed by the existence of one person who suffers, and has a life that is not worth living". This is not necessary, one could arrange people with negative welfare in two groups and weigh people with small sufferings against people with low welfare and so on.

The Limited Quantity View 1 (LQV-1)

$$V = \begin{cases} nQ & nQ \leq x \\ x & nQ > x \end{cases}$$

The Limited Quantity View 2 (LQV-2)

$$V = \begin{cases} v_1 Q + v_2 nQ & nQ \leq x \\ v_1 Q + v_2 x & nQ > x \end{cases}$$

The first interpretation yields the ridiculous result that when quantity has reached the level x, we cannot make a population better even if we raise the quality or add people with quality higher than average. The second interpretation avoids this conclusion, but there is another problem: Mere Additions.

4.3. Mere Additions

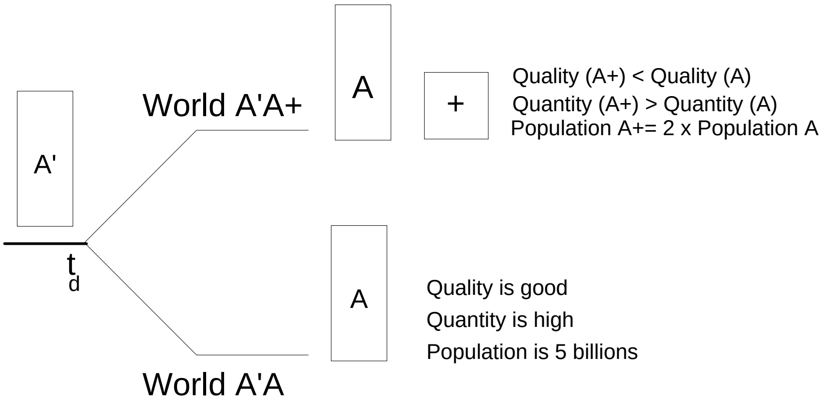

A'
t_d
World A'A+
A
+
Quality (A+) < Quality (A)
Quantity (A+) > Quantity (A)
Population A+= 2 x Population A
A
Quality is good
Quantity is high
Population is 5 billions
World A'A

Figure 4.3-1. A Mere Addition case.

Parfit describes a Mere Addition as follows:

> There is *Mere Addition* when, in one of two outcomes, there exist extra people (1) who have lives worth living, (2) who affects no one else, and (3) whose existence does not involve social injustice.[39]

The extra people, the +-people, are worse off than the people in the first group, the A-people. To avoid any involvement of social injustice, Parfit assumes "that the two groups in A+ are not aware of each other's existence, and could not communicate."[40] Parfit's exemplification is that "A+ is some possible state of the world before the Atlantic Ocean had been crossed. A is a different state in which the Americas are uninhabited."[41] This is a factual possibility, it is imaginable that it could have been the case that America was uninhabited before Columbus' "discovery". In our time we could imagine that there are perhaps still tribes in the Amazon that we do not know about, and ask ourselves whether it would be better if there were no undiscovered tribes in the Amazon.

Further on, Parfit claims:

[39] Parfit (1984), p. 420.

[40] Ibid.

[41] Ibid., p. 420.

> Whether inequality makes the outcome worse depends on how it comes about. It might be true either (3) that some existing people become worse off than others, or (4) that there are extra people living who, though their lives are worth living, are worse off than some existing people. Only (3) makes the outcome worse. - - - We cannot plausibly claim that the extra people should never have existed, *merely because, unknown to them, there are other people who are even better-off.*[42]

We can make a distinction between *social inequality,* inequality that is both known and removable, and *natural inequality,* inequality that is not known and not removable, people are worse off through no fault of anybody. In Parfit's Mere Addition case there is natural inequality, and he acknowledges that this is a bad feature of A+, but he denies that this feature makes A+ worse than A when it comes about by a Mere Addition. Observe that Parfit only claims that A+ is *not worse than* A, not that A+ is *better than* A. If one believes that the value of quantity has reached its limit in A, one could say that A+ is in no respect better than A, and in one respect, natural inequality, worse than A, but this bad feature does not make A+ worse than A when it comes about by a Mere Addition. We can formulate the following principle:

> *The Mere Addition Principle*: For any population, if by Mere Addition one adds any number of individuals with positive welfare to create a new population, then this new population is not worse than the original one.

This is a very compelling principle, one that is embraced by many (all?) contributors.[43] How could a population get worse by just adding people with positive welfare? Just think about adding one very happy person to an even happier population. Indeed, when an inequality is not removable, we think that even if the two groups know about each others existence, an addition does not make the outcome worse. We can add to this that the outcome gets even better for the original people, and formulate another compelling principle:

> *The Pareto Addition Principle*: For any population, if one increase the quality of the original people and adds any number of individuals with positive welfare to create a new population, then this new population is not worse than the original one.

Here, all the original people's quality is raised too, that is, the addition has affected them but in a positive way. The added people have positive but lower welfare than the original people but this inequality is not removable. This is easy to specify: The people in A could procreate and their children could have lower quality than their parents during the time slice. Most of us would probably say that such an addition makes the outcome better, but it is enough for our purposes to formulate the weaker claim that the outcome does not get worse.

[42] Ibid., p. 425.

[43] Among others, Parfit, Hudson, Sider, and Ng.

In figure 4.3-1, quantity has reached its upper limit. This implies, according to LQV-2, that the existence of the +-people make the outcome worse. The contribution that quantity gives to the total value will be the same in both cases because we have reached the limit to the value of quantity, but the contribution from quality will be less in population A+. Indeed, when this limit is reached the principle will behave like Average Utilitarianism, the extra quantity does not contribute any value at all. LQV-2 violates the Mere Addition Principle. Assume now that the A-people in population A+ gets higher welfare than the A-people in population A and call these people the α-people. As long as the difference in average utility between the α-people and the A-people is less than the difference between the α-people and the +-people, population α+ will be considered worse than A (the average welfare will be less in α+ than in A). Hence, LQV-2 also violates the Pareto Addition Principle.[44]

The following conclusion is probably the most repulsive one formulated in this essay:

> *The Sadistic Conclusion*: When adding people without affecting the original people's welfare, it sometimes can be better to add a number of people with negative welfare rather than a number of people with positive welfare.

As we saw above, when quantity has reached its limit, LQV-2 mimics Average Utilitarianism. It shares with this principle the objectionable feature that we can sometimes make a population better by adding people with negative welfare rather than positive. Suppose that a population of ten billion people has reached the level where quantity does not contribute with any value at all, and that the average welfare is 10 units. Then it would be better to add one million people all with the negative welfare -1 units rather than two million people all with the positive welfare 1 units (in the former case the average utility will be 4.5 units, in the latter case 4 units).

Let us return for a moment to the linear weighing principles we considered in section 4.1. How would they fare when it comes to the Mere Addition and the Pareto Addition Principle?

All principles that give linear increasing value to quality and quantity violates the Mere Addition Principle. Assume that we give quantity the weight 1, that is, the relative weight of quantity and quality is determined solely by the weight we give to quality. If we have a population A with quality $Q_A > 0$ and quantity $K_A > 0$, and an alternative population A+ consisting of A and one extra person with positive quality $u < Q_A$, then we get the following value functions:

$$V(A) = v_1 Q_A + K_A$$

$$V(A+) = v_1 Q_{A+} + K_A + u$$

[44] The second version of the Valueless Level View (VLV-2) also violates the Mere Addition Principle. Adding one person with positive quality but below the valueless level, when everybody else has a quality above the level, makes the outcome worse. If all the people's quality is below the valueless level, VLV-2 also violates the Pareto Addition Principle; VLV-2 then mimics Average Utilitarianism and, as we saw above, Average Utilitarianism violates both the Mere Addition and the Pareto Addition Principle.

$$V(A) - V(A+) = v_1(Q_A - Q_{A+}) - u$$

We know that $u<Q_A$, so it follows that $Q_A>Q_{A+}$. Now, for any $v_1>0$ there exists an u such that $0<u<v_1(Q_A-Q_{A+})$, i.e., such that V(A)>V(A+). In other words, for any linear weighing of quantity of quality, we can construct a case were a Mere Addition of one person with positive welfare makes the outcome worse.

This is easier to see if we look at linear weighing principles that at least give quality the same weight as quantity - recall, as we saw in section 4.1, that we need to give quality weights of the magnitude 10^8 to 10^{10} to avoid Repugnant Conclusions. Compare a population of one person with a welfare of 100 units with a population created by adding one person with 2 units; giving both quality and quantity the weight one yields the value 200 for the first population and 153 for the second. We can conclude that all principles that give linear increasing value to quality and quantity and that avoids the Repugnant Conclusions clearly violate the Mere Addition Principle. Similar reasoning can be done for the Pareto Addition Principle and the Sadistic Conclusion. Giving both quality and quantity the weight one, a population of one person with a welfare of 100 units yields the value 200, and a population of two people, one with a welfare of 102 units and another with a welfare of 2 units, yields the value 156 - a violation of the Pareto Addition Principle. Now consider a case where we can add either two persons with a welfare of 1 unit each, or one person with a negative welfare of -4 units, to a single person population with a welfare of 100 units. The former addition yields the value 136 and the latter 144 - a clear instance of the Sadistic Conclusion.

There is a vexatious relation between the Mere Addition Principle and the Sadistic Conclusion - when a theory violates the former then one can suspect that it is going to imply the latter. If an addition of many persons with positive welfare can decrease the value of a population, then one can construct cases where one adds one person with negative welfare who decreases the value of the population less than the addition of the many persons with positive welfare.[45] One might be tempted to accept that a population gets worse when one adds a person with low but positive welfare, as a way to solve the problems of population ethics.[46] We now know that we better resist that temptation.

4.4. Summary Linear Theories

We saw that *no* impersonal principle that assigns linear increasing value to quantity and/or quality of welfare could comply with our conditions of acceptability, avoidance of the Repugnant, the Reversed Repugnant and the Sadistic Conclusion and compatibility with the Mere Addition Principle and the Pareto Addition Principle. We came to the same conclusion concerning principles that assign linear

[45] For a more formal discussion, see section 6.4 in the current chapter.

[46] Parfit (1986) flirts with this solution in his most recent paper on this topic. He invokes perfectionist values to argue that A+ perhaps is worse than A, at least when the welfare of the +-people are very much lower than the welfare of the A-people (see figure 4.3-1).

increasing value to quality and quantity but made use of different kinds of limits to the value of quantity - lexical orderings and higher goods principles. In other words, *there exists no linear weighing function of quality and quantity that complies with these five conditions of acceptability.* In the Repugnant and Reversed Repugnant Conclusions the alternatives differ only in average welfare and population size. We cannot appeal to any other values, like equality, to justify our evaluation. In the Mere Addition and Pareto Addition cases there is natural inequality that perhaps could make the outcome worse, but we concluded that this inequality does not make an outcome worse when it is created by Mere Addition or Pareto Addition.[47] At any rate, if one were to accept that a Mere Addition or a Pareto Addition could make a population worse, then one would have to accept the Sadistic Conclusion.

5. The Person Affecting Restriction

5.1. Make People Happy, Not Happy People

An idea that underlies many arguments in moral philosophy and economics is that an action can only be good, or bad, if it has good effects, or bad effects for somebody. Narveson couches this in the slogan "We should make people happy, not happy people." He develops his view as follows:

> Morality has to do with how we treat whatever people there are. Utilitarianism, construed as a moral theory, says that we should aim at maximizing the happiness of people, the balance of their desirable over their undesirable experience. - - -. On this view, moral questions *presuppose* the existence of people. If we are contemplating an increase in the population, then we may consider how well or badly the new people would be likely to be in the circumstances they would occupy. But suppose we decide, in the end, not to bring them into existence. Then, even if they would have been perfectly happy, still, nobody misses anything; or anyway, it is only we, the people he would have desirable effects upon, who are missing something. But there *isn't any loss* of anything by the *contemplated party*, since he doesn't exist. On the other hand, if we actually do bring him into existence, then of course we must be concerned for his welfare as anyone else's. - - - The suggestion, in other words, is that we can have a moral reason, arising from *concern for the welfare of actual persons, for not having children*, but *not*, arising from the same considerations, *for having them*, so long as the persons in questions are those who would be brought into existence by the contemplated act. There is no moral objection *against* having children who would be *happy*, for the duties we would then have to them would be discharged satisfactorily. But if it would be *impossible* to fulfill the duty to *promote their welfare*, then we *ought to avoid conception*.[48]

As we saw above, Total Utilitarianism implies the Repugnant Conclusion. Narveson tries to avoid this implication by asserting that nobody is harmed by not being born to a happy life, whereas existing people are harmed if we lower their welfare. Another attractive feature of Narveson's theory is that it embraces what has been called the "Asymmetry." On this view, the fact that a person's life would not be

[47] One principle that we have not discussed, Rawls' Maximin, could perhaps be applied to Mere Addition or Pareto Addition cases, but Parfit argues convincingly for a rejection of Maximin in different number cases. See Parfit (1984), pp. 422-23. At any rate, Maximin violates both the Mere Addition Principle and the Pareto Addition Principle.

[48] Narveson 1973, p. 73. Our emphasis, except the first one.

worth living constitutes a strong moral reason for not bringing her into existence, while the fact that a person's life would be worth living provides no or only a weak moral reason for bringing her into existence. This view is strongly approved by common sense but is not espoused by any of the impersonal principle we have considered so far. According to Total Utilitarianism, for example, we have, *ceteris paribus*, a moral duty to produce a lot of happy children.[49]

We can interpret the Person Affecting Restriction in several ways. Temkin has put forward a general description:[50]

> (1) One situation is worse (better) than another if there is *someone* for whom it is worse (or better), and *no one* for whom it is better (or worse), but not vice versa, and
>
> (2) One situation *cannot* be worse (or better) than another if there is *no one* for whom it *is* worse (or better).

We can now get different versions by interpreting "someone for whom it is worse" in different ways. We can believe, as Narveson seems to do, that for an act to be worse for someone, there must exist a complainant.[51] Thus, an alternative would be worse for people only if there are or will be people for whom it is worse, and causing to exist cannot benefit but can harm a person. A better way to elaborate this is to make a distinction between *necessary* and *contingent* persons. A person is a *necessary person,* relative to a set of alternative *worlds,* exactly if she exists in all alternative worlds. A person is a *contingent person* exactly if she does not exist in all alternative *worlds*.[52] We can then interpret Narveson's theory to mean that we cannot make a contingent person worse off by choosing an alternative in which she does not exist. On the other hand, if we choose an alternative where a contingent person would exist, and there is an alternative where she would have it better, then we have harmed her. Finally, choosing an alternative where a contingent person would have a life not worth living would always harm her.

Suppose we are comparing two outcomes X and Y. Another interpretation of "someone for whom it is worse" could be that the occurrence of X rather than Y would be either worse or bad for the X-people. This is what Parfit has called the "narrow sense" of "worse". He also states the "wide sense" of "worse": X is worse for people if the occurrence of X would be less good for the X-people than the occurrence of Y would be for the Y-people.[53] Further on, we can combine these two senses of "worse" with the claim that a person is benefited if she is caused to exist, that is, to

[49] There is, however nothing intrinsic to impersonal theories that bars them from incorporating the Asymmetry. Negative Utilitarianism (se Ch. 3) is the well-known example of an impersonal theory that is compatible with the Asymmetry. We shall, in section 6 below, discuss theories that embrace the Asymmetry in the compelling cases.

[50] Temkin 1986, p. 166. As Temkin points out, the "not vice versa" clause is only necessary on the view that causing someone to exist can benefit that person, even though failing to cause someone to exist harms no one.

[51] Narveson never explicitly states the view that there has to be complainant but there are ample indications that this is what he in fact requires, such as the quote above. See also Narveson (1978), pp. 43, 50 and 55-56. Cf. McMahan (1981) who convincingly argues for this interpretation of Narveson's theory.

[52] The concepts of necessary and contingent persons are from Österberg (1992). We shall discuss Österberg's theory below.

[53] See Parfit (1984) p. 395-6.

give up the Asymmetry. We shall first discuss Narveson's and Österberg's theories, then the two other interpretations of the Person Affecting Restriction.

5.2. Narveson's Theory

How could Narveson's theory avoid the Repugnant Conclusion? He summarises his theory as follows:[54]

> (1) New additions to population ought not to be made at the expense of those who otherwise exist, even if there would be a net increment in total utility considered in person-independent terms. But (2) new additions ought to be made if the benefit to all, *excluding* the newcomer, would exceed the cost to all, *including* him or her, as compared with the net benefit of any alternatives which don't add to population [i.e., if the benefit minus the cost would exceed the net benefit of any alternative]. Finally, (3) within those limits, the decision whether to add to population is up to the individuals involved in its production, provided that if they have a choice of which child to produce they produce the happier one, other things being equal.

This theory would avoid Repugnant Conclusions in cases involving nonexclusive Different Number Choices, such as when we must decide whether to add to an existing population or not. In such cases, there would be an expense for the necessary people; their welfare would be diminished. Moreover, Narveson's theory embraces the Mere Addition Principle. Finally, assuming that Narveson would use a totalling principle when it comes to evaluate alternatives in which we can diminish the number of existing people, this theory would not imply a Reversed Repugnant Conclusion.

But when we consider exclusive alternatives, problems arise.[55] In cases such as example 4, "Different Energy Systems - Different People" in Chapter 1, all people are contingent. It is not clear what Narveson's theory would imply here; the clause (3) above is silent about what kind of aggregation principle Narveson has in mind when it comes to contingent people. If we were to use a totalling principle, Narveson's theory would imply the Repugnant Conclusion when considering exclusive alternatives. If we, on the other hand, were to use an averaging principle, we would get the Reversed Repugnant Conclusion and violate the Mere Addition Principle.[56] In other words, we get all the problems associated with the impersonal principles we considered in section 4.

We would also get conclusions similar to the Repugnant Conclusion in nonexclusive cases. If we adopt a totalling principle for contingent people, then we should add a huge amount of people with low welfare rather than a smaller amount

[54] Narveson (1978), p. 55-56.

[55] For definitions of Same People Choices, exclusive and nonexclusive Same Number and Different Number Choices, see Ch. 2, section 2.2 and Ch 4., section 1.

[56] It seems that Narveson favours an averaging principle, since his "concern is that whatever people there are be as happy as possible... The concern that there be more people, simply to maximize instances of happiness in the universe, seems, of another order" (Narveson 1978, p. 55). Cf. McMahan (1981, p. 103). On the other hand, as McMahan points out, Narveson rejects averaging principles as population principles in Narveson (1973). However, we do not think, as McMahan does, that this bars Narveson from accepting an averaging principle for contingent people; the argument that McMahan refers to is directed against the impersonal Average Utilitarian Principle.

with high welfare as long as the total welfare of the added people is higher in the former case.

This shows that Narveson's theory is untenable as it stands, but perhaps we could amend it by combining it with a better impersonal principle, such as one of the Variable Value Principles we are going to consider in the next section. There are, however, other problems with Narveson's theory.

5.3. Adam and Eve

Consider the following case:

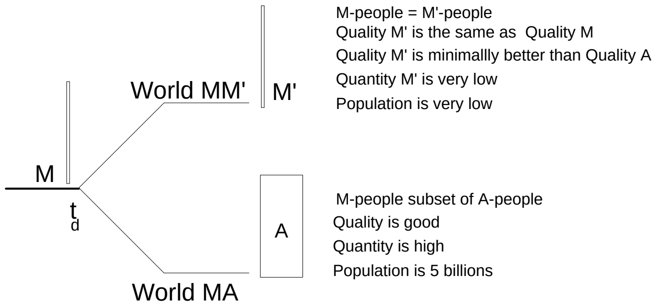


Figure 5.1-1 The Adam and Eve case.

Narveson's theory would here prefer M' to A although the M-people would continue to exist in world MA with just a slight reduction of their welfare. This claim is hard to believe: How could this small loss for the M-people outweigh the great increase in total quantity of welfare? This is tantamount to holding that it would be worse if, instead of Eve and Adam, a billion billion other people lived, all with a quality of life that would be almost as high - a quite extreme perfectionist view one might say. Consider also the case that either Eve and Adam continue to live alone or they live on with the same welfare together with a billion billion people *as happy as them* plus one slightly unhappy individual. Here, Narveson's theory would still prefer the solitude of Adam and Eve - a quite extreme negativism one might say.

Narveson would accept the Adam and Eve case:

> Now, if we agree that something is intrinsically valuable, then no doubt we should do something to promote it. But is this the sort of ground on which we should promote human happiness? - - - I am inclined to say instead that we should promote people's happiness, and reduce their unhappiness, where possible, because they are people and that is the way people should be treated. *It is not, as it were, because people are nice things to have around, still less that happiness is a nice thing to have around,* although that is probably true enough. Intrinsic value is the particular home, one supposes, of the aesthetic. And we can well imagine people discussing the question of what sort of world is nicest or most interesting, some extolling the virtues of vast barren wastelands and rugged mountains, with a smallish and hardy populace to do combat with its challenges, others favoring a more social sort of place with lots of cities full of varied people with diverse tastes and customs and so on. - - - As between the first and second, however, I find it overwhelmingly plausible to say that the issue between them, hence the choice between them, was a matter of taste. Morally speaking, so far as the descriptions go, there seems nothing to choose between them. No doubt there is, in an obvious sense, *more happiness in the second than in the first. . . But it seems to me simply odd to count that as a reason for thinking that the second situation is morally better than the first.* - - - it seems repulsive to think that the goodness of a community is a function of its size, e.g., that America is a happier country than Canada because it is so much bigger, demographically.[57]

[57] Narveson (1973), pp. 72, 80.

This reasoning also holds, of course, for the Adam and Eve case.[58]

5.4. An Ambiguity

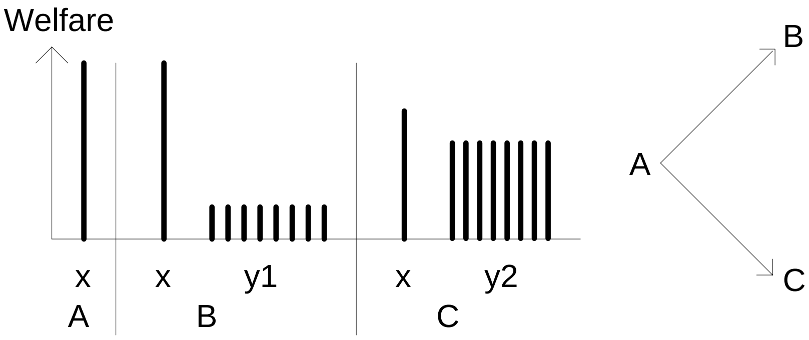


Figure 5.1-2. An ambiguous case.

Applied to the case above, Narveson's theory is ambiguous. It seems that Narveson assumes that we always have the possibility to avoid addition when he says that we should compare the net benefit of an alternative with "the net benefit of any alternative which don't add to the population." As the present case illustrates, this is not always the case. Let us introduce a new distinction: We have an *avoidable addition* when there exists an alternative where no contingent people are added; we have a *unavoidable addition* when there is no alternative where contingent people are not added. Two revisions of Narveson's theory are now available:[59]

> ***Theory N1***: In cases with unavoidable additions, if the welfare of the contingent people is positive in all of the alternatives, and there are some alternatives where the necessary people's welfare is not decreased, then we should choose the one of the latter alternatives where the value of the necessary and the contingent people is maximised on some impersonal principle. In the other cases, where the necessary people's welfare is decreased or the welfare of some or all the contingent people is negative, we should choose the alternative where the welfare of the necessary people, plus the eventual contingent people with negative welfare, is maximised. Within these limits, we should choose the alternative where the value of the contingent people is maximised on some impersonal principle.

> ***Theory N2***: When it comes to unavoidable additions, the contingent people are on the same footing as the necessary people and one should maximise the value of the population according to some impersonal principle.

[58] Note that the Adam and Eve case is not a Reversed Repugnant Conclusion (see fig 4.1-2). It is not true that for any possible population of at least five billion people, all with a good quality of life, there must be some imaginable population with just slightly better quality, whose existence would be better, even though this population is very much smaller and the total quantity is very much lower. If the M-population in fig. 5.1-1 would consist of the A-people with slightly higher welfare, then Narveson's principle would prefer world MA to MM'.

[59] Theory N1 is probably the revision most in line with Narveson's intuitions. The two theories above are not the only possible extrapolations of Narveson's theory to cases involving unavoidable addition, but they are certainly the most charitable ones.

The "impersonal principle" mentioned above could be one of the Variable Value Principles we shall discuss in the next section. In the examples considered here, these principles would behave like the Total Utilitarian Principle.

Both of these new theories have problems of their own. Theory N1 yields the conclusion that alternative B is better than C in the case above. This seems awkward: the x-person's loss in outcome C is considerably less than the difference between the welfare of the y1- and y2-people. This case looks similar to the Adam and Eve case but there is one important difference. In the Adam and Eve case we had the option not to add any people at all, an option which is not obtainable in the present case. If the y1- and y2-people had been identical, then theory N1 would pick out alternative C. Why should the different identities of people play such an decisive role when we cannot avoid adding them? Indeed, how much positive welfare the lives of the y2-people have is irrelevant; even if they had ecstatic lives, theory N1 would rank B as better than C.

Theory N2 works better in this case: it would value C better than B. But what is left of the "Person Affecting Restriction"? Exclusive Same Number and Different Number Choices and nonexclusive Different Number Choices with unavoidable additions are taken care of by an impersonal principle, and when it comes to nonexclusive choices with avoidable additions theory N2 needs help from an impersonal principle. The only thing left of the Person Affecting Restriction is that when it comes to nonexclusive Different Number Choices with contingent additions, additions should not be made at the expense of necessary people. Moreover, this only creates problems similar to the Repugnant Conclusion. Consider a case where we have three alternatives. In the first alternative the original population of, say, a billion people with very high quality of live continue as before: no changes in their welfare and no additions are made. In the second alternative a billion people with very high quality are added. In the third alternative, one person in the original population gets a slight increase in his or her welfare, the rest of the original population continues as before and a billion billion people with welfare close to zero are added. Here, theory N2 rules that the last addition "ought to be made" because "the benefit to all, excluding the newcomer, would exceed the cost to all . . . as compared with the net benefit of any alternatives which don't add to population."[60] This is even worse than the Repugnant Conclusion implied by Total Utilitarianism: we ought to add the billion billion people with welfare close to zero even if the total welfare of these people is lower than the total welfare of the billion people with very high quality. Indeed, the added people could have zero welfare.

We also have a version of the Negative Repugnant Conclusion 2:[61] Assume that we have a big population of people with high welfare. Additions of very unhappy lives can now be compensated by small increases in the welfare of the original people, as long as these small increases add up to more welfare than the negative welfare of the added unhappy people.

Narvesson's person affecting principle does not solve any problems, it only creates new ones.

[60] Narveson (1978), p. 55-56.

[61] See Ch. 3, section 5.2, for the intrapersonal version of this conclusion.

5.5. Pessimism Utilitarianism

Jan Österberg has proposed a theory similar to Narveson's. It is based on the following principles:[62]

> PU1: That there is an (extra) being who is happy is not intrinsically better than that there is no (extra) being.
>
> PU2: That there is a being who is happy to the degree n+m is intrinsically better than that there is a being who is happy to the degree n.
>
> PU3: That there is an (extra) being who is unhappy is intrinsically worse than that there is no (extra) being.
>
> PU4: That there is a being who is unhappy to the degree n+m is intrinsically worse than that there is a being who is unhappy to the degree n.

This theory embraces, of course, the Asymmetry: We are not making the world better by causing a happy person to exist, but we make it worse if we bring about an unhappy person.[63] Österberg launches the following principles to evaluate alternative worlds:

> Given a set M of alternative worlds, call the beings who exist in all alternative worlds necessary beings (relative to M) and the beings who do not exist in all alternatives contingent beings (relative to M). Let *n* be the number of contingent happy individuals who exist in the alternative or those alternatives which have the smallest number of these beings.
>
> (1) The positive intrinsic value of a world V is the sum of the happiness of the happy necessary beings in V plus the sum of the happiness of the *n* happy contingent beings in V who are the least happy.
>
> (2) The negative intrinsic value of a world V is the sum of the unhappiness of the unhappy beings.
>
> (3) The intrinsic value of a world is the positive intrinsic value minus the negative intrinsic value.

Like Narveson's theory, this principle implies that it is better that Adam and Eve continue to exist in solitude in the three versions of the Adam and Eve case above. Österberg's theory will also imply the Reversed Repugnant Conclusions in exclusive Different Number Choices. Suppose that we have one alternative A with five billion happy people and another alternative B with one slightly more happy person and that this is an exclusive case. The number of necessary persons will then be zero and *n* will be equal to one. Österberg's theory will then only count the welfare of one person in A and compare it with the welfare of the slightly more happy person in B. Finally, Österberg's theory implies similar versions of the Negative Repugnant Conclusions as Narveson's theory. This is especially noteworthy, considering the fact that Österberg's theory is supposed to be a negativist theory, i.e., a theory that gives more weight to unhappiness than to happiness.

[62] Österberg (1992)

[63] Österberg, unlike Narveson, argues for this asymmetry on conceptual grounds, see Arrhenius (1992).

5.6. Parfit's Narrow and Wide Person Affecting Restrictions

Suppose we have two outcomes X and Y. We can then state the Narrow Person Affecting Restriction as follows:

> *The Narrow Person Affecting Restriction*: One alternative X is worse (better) than another alternative Y if the occurrence of X rather than Y would be worse (better) for the X-people.

Parfit combines this principle with the claim that we can benefit somebody by causing her to exist and calls this principle the "narrow principle."[64] But whether or not we make this claim, this principle will imply contradictions. Assume that both the x- and the y-people have lives not worth living and that the x- and y-people are different people, that is, an exclusive Same Number Choice. Then alternative X is worse than Y for the x-people and alternative Y is worse than X for the y-people. This is a contradictory conclusion which cannot be amended in any way.

The "wide" interpretation of "worse for people" avoids this conclusion:

> *The Wide Person Affecting Restriction*: One alternative X is worse (better) than another alternative Y if the occurrence of X would give a lower (higher) total net benefit to the X-people than the net benefit given to the Y-people by the occurrence of Y.[65]

If we combine this restriction with the Asymmetry, then we have a vague statement of Narveson's principle. As we have seen, Parfit rejects the Asymmetry. But if causing to exist can benefit, then the wide person affecting restriction just restates the impersonal Total Utilitarian Principle in a person affecting form. Consequently, this principle will imply the Repugnant Conclusion. The Z-people in figure 4.1-1 will *together* receive a greater benefit than the A-people, even though each individual in Z will receive a smaller benefit than each individual in A.

The Narrow Person Affecting Restriction both with and without the Asymmetry implies contradictions. The Wide Person Affecting Restriction without the Asymmetry just restates the Total Utilitarian Principle.[66] We should reject both of these principles. The Wide Person Affecting Restriction combined with the Asymmetry is a vague variant of Narveson's principle, a principle which we explicated and rejected above.

5.7. The Asymmetry

[64] Parfit (1984), p. 395. Parfit also points out the contradiction we present below.

[65] See Parfit (1984, p. 394-96) for similar definitions. An illuminating discussion of these principles can also be found in McMahan (1981). Temkin (1986, p. 166) defines what he calls the "Person Affecting Principle" which is similar to the Wide Person Affecting Restriction. The use of the "total net benefit" clause in the definition of the wide principle is necessary to avoid ambiguity in different number cases. We could of course use other aggregation principles, like averaging, but that will have no importance for the arguments pursued here.

[66] If we combined the "wide" sense of "worse for people" with some other aggregation principle, such as the Average Utilitarian Principle, then we would have a restatement of that principle in person affecting terms.

We mentioned above that one of the positive features of the Person Affecting Restriction is that it embraces the Asymmetry, i.e., the fact that a person's life would not be worth living constitutes a strong moral reason for not bringing her into existence, while the fact that a person's life would be worth living provides no or only a weak moral reason for bringing her into existence. Perhaps this is the main argument for a Person Affecting Restriction. Common sense beliefs seem to strongly support the Asymmetry. However, Narvesson formulates his view on a normative level, in terms of action and duties, and one can suspect that values other than welfarist ones are involved and blurs our intuitions in such cases, i.e., liberal values. Perhaps we think it is a too strong demand to say that we have a moral duty to procreate. Rather, the decision to procreate is up to every individual to freely decide by themselves as long as they do not bring a suffering individual into existence. Let us formulate two versions of the Asymmetry in axiological terms:

> *The Strong Asymmetry*: *Ceteris paribus*, adding a contingent person with positive welfare neither increases nor decreases the value of a population. Adding a contingent person with negative welfare decreases the value of a population as much as if this person had been a necessary person.
>
> *The Weak Asymmetry*: Increasing the welfare of necessary people or unavoidable contingent people increases the value of a population more than adding a number of avoidable contingent people with a positive welfare equal to the increase. Decreasing the welfare of necessary people or unavoidable contingent people decreases the value of a population as much as adding a number of avoidable contingent people with a negative welfare equal to the decrease.

The Strong Asymmetry is the one defended by Narveson. The Weak Asymmetry is one we think is more intuitive and this version avoids the problems that the first one ran into. It will preserve the more compelling intuitions we have about contingent persons. Let say that we can choose between increasing the welfare of the existing people or add a number of contingent people with a welfare equal to the increase. Here the Weak Asymmetry would opt for the former alternative. If we can choose between inflicting suffering on an existing person or creating a suffering person, then these two outcomes are equally bad if the suffering is the same in the two cases. Lastly, if somebody chooses to procreate and the person thereby created happened to be a happy person, then that choice would make a population better. In the last case, the Strong and the Weak Asymmetry conflict, the former principle yields that bringing the happy person into existence did not increase the value of the population. This seems quite awkward to us. Compare a family of prospective parents with good welfare to a family where the welfare of the parents has been slightly decreased but there is a child with good welfare.

The Weak Asymmetry, combined with act-consequentialism, would give us some reasons to procreate, but not as strong as Total Utilitarianism. In a pluralist ethic, we could combine the Weak Asymmetry with some liberal principle that assigns value to the autonomy of the prospective parents. Such an ethic would capture all intuitions about the moral duties to procreate or not, but still hold that, *ceteris paribus*, one makes a population better if one adds a happy person.

We should also note that Narveson does not *justify* the Asymmetry, he only restates it when he appeals to the intuition that there has to be a complainant for an act to be wrong.[67] The Weak Asymmetry, on the other hand, can be justified on impersonal welfarist grounds by giving more weight to suffering and quality of welfare than to happiness and quantity of welfare. As we argued in chapter 3, there are many good reasons to give more weight to suffering than to happiness, and as we have argued in this chapter, there are many good reason to give more weight to quality of welfare than to quantity of welfare.

6. Variable Value Principles

One of the ideas behind *Variable Value Principles*[68] is that the value a new entity contributes to the value of a whole depends on the number and the quality of the original entities that make up the whole. This is analogous to the idea of diminishing marginal value used in economics: the more money a person already has, the lesser good an extra pound will do her. This is represented by a *strictly concave function,* where the slope in a given point (x_i) is the marginal benefit of that income, that is, the extra good a person would obtain from an extra pound.

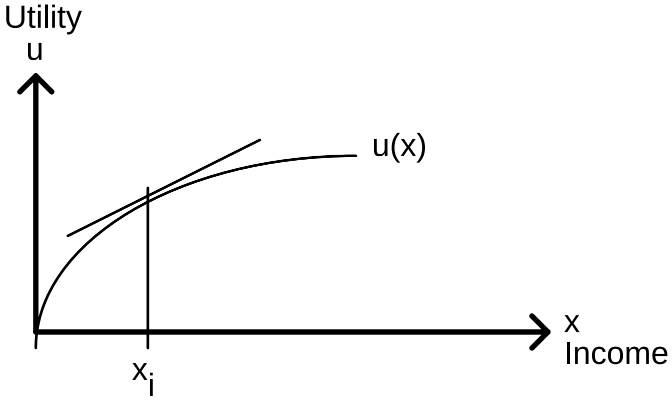


Figure 6-1. A diminishing marginal value graph.

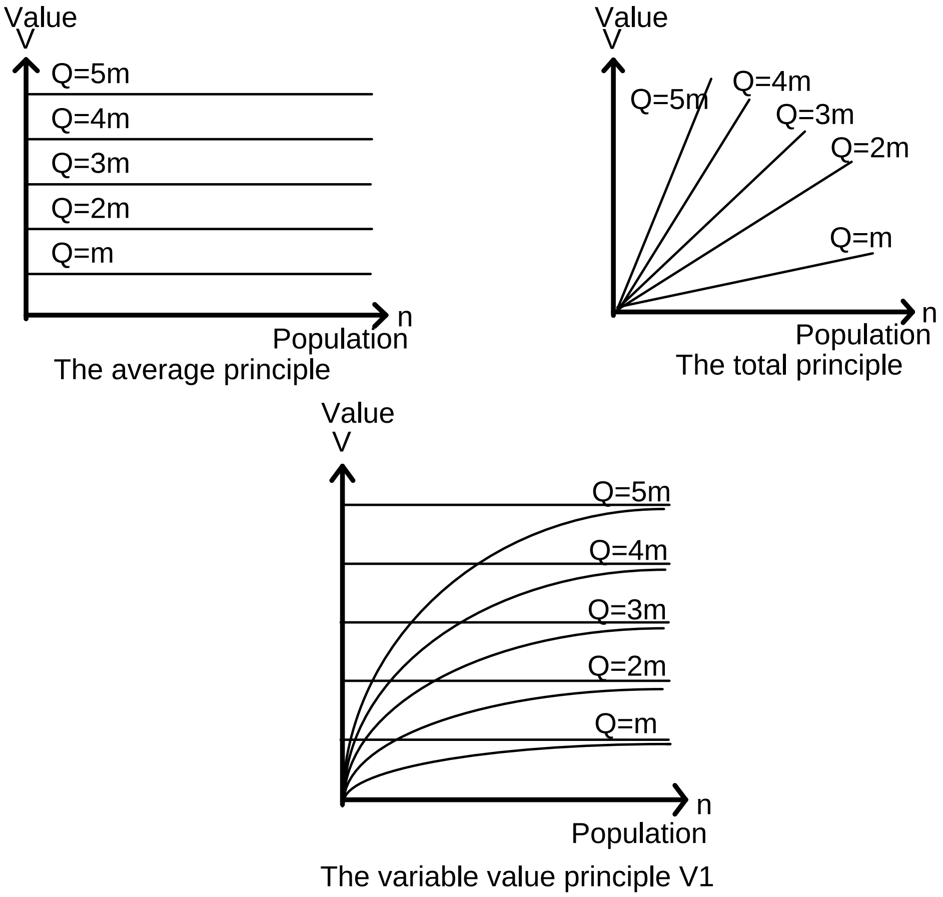
Variable Value Principles are sometimes called "compromise theories" since a Variable Value Principle can be said to be a compromise between the Total and Average Utilitarian Principle. When it comes to small populations with good welfare, a Variable Value Principle behaves like Total Utilitarianism and assigns most of the value to increase in total quantity of welfare.[69] When it comes to large populations

[67] Österberg (1992) makes an interesting attempt to justify the Asymmetry on conceptual grounds, by invoking an asymmetry in our use of deontic modalities. This attempt, however, is not successful, as shown in Arrhenius (1992).

[68] To the best of our knowledge, Hurka was the first to propose a Variable Value Principle (Hurka 1983). Sider (1991) and Ng (1989) have also proposed theories of this kind, while Hudson (1986) has attacked the idea. Parfit (1984, p. 402) mentions a Variable Value Principle but ignores it on basis of the faulty reason that such a principle applied to large population sizes would be equal to linear quantity limiting principles, i.e., the kind of principles we discussed in section 4.

[69] Hurka (1983, p. 497) argues that this is not good enough. When we have small populations, the contributing value should be greater than the mere sum of people's welfare, to open up for the possibility that the contributing value can outdo the lowering

with low welfare, the principle mimics Average Utilitarianism and assigns most of the value to increase in individual quantity of welfare.



The average principle

The total principle

The variable value principle V1

Figure 6-2. Three population principles (from Hurka 1983)

Variable Value Principles assign asymptotically increasing value to total quantity of welfare and linear increasing value to individual quantity of welfare. If we keep the individual quantity of welfare constant and increase the population size, then the value of a population will converge on a value limit asymptotically -- a doubling of the population size without any increase in individual quantity of welfare will always do less than a doubling of the value. A doubling of the individual quantity of welfare, on the other hand, will always double the value of the population, which is reflected in the even spacing of the asymptotes. We can construct a Variable Value Principle in at least two different ways as suggested by Ng and Sider respectively.

6.1. Ng's Principle

One way to construct a Variable Value Principle is to dampen the increase of the linear function *n*, the population size, by transformation with a concave function

of total utility for the sake of population growth. Excluding the possibility that Hurka assigns intrinsic value to population growth as such, his argument rests on a conflation of intrinsic and instrumental value.

f(n), as suggested by Ng.[70] He calls his principle "theory X'."[71] Whereas the Average Utilitarian Principle maximises the average welfare *Q*, and the Total Utilitarian Principle maximises *n* x *Q*, this principle maximises *f(n)* x *Q*. A suitable concave function could look like the one below:

$$f(n) = \sum_{i=1}^{n} k^{i-1} \qquad 1>k>0$$

The weighing coefficient *k* represents how quickly the values of additional people approach zero. The smaller *k* is, the quicker the values of additional people decline. When *n* approaches infinity, *f(n)* approaches 1/(1-*k*), which is of finite value. For example, for *k*=0.99, *limes f(n)*=100. This means that when it comes to large populations, the value yielded by the function X'=*f(n)* x *Q* is not increased when the average welfare is decreased but the total quantity of welfare is increased by addition of more people. When it comes to large populations, X' approaches *m* x *Q* where *m* is a constant (100 in the example above); that is, X' behaves like Average Utilitarianism with large populations and thereby avoids the Repugnant Conclusion.

A problem with Ng's principle, however, is that it violates the Mere Addition Principle. For example, if *k*=0.99, then a Mere Addition of one billion people to a population of one billion, would make the outcome worse, even if the quality of the added people were as high as 75 percent of the original people. This can be shown in a more general way by the following figure:

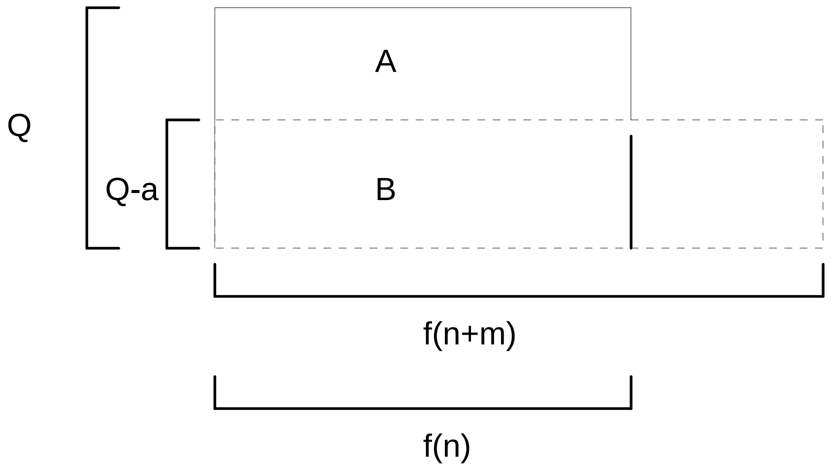


Figure 6.1-1.

In figure 6.1-1, the length of the horizontal lines represents the *dampened* number of people and the height of the vertical lines represents the average welfare *Q*. The values of the populations A and B are thus represented by the areas of the blocks since, according to Ng's principle, the value of A is *Q* x *f(n)* and the value of B is (*Q*-*a*) x f(*n* + *m*).

[70] Ng (1989).

[71] Parfit calls the theory of beneficence that can solve all problems related to future generations and population ethics for "theory X." Hence Ng's paraphrase "theory X'."

The difference between A and B is that in B *m* persons with positive welfare are added to the population. These added people have a welfare that is well below the welfare of the A-people. Hence, they lower the average by *a* units. In the figure above, the lowering of the average is so great that, although the number of people increases and the horizontal line is prolonged, the area of block B is smaller than the area of block A. Consequently, the Mere Addition of *m* persons with positive welfare makes population B worse than population A. A similar conclusion can be made about the Pareto Addition Principle. Even if the average welfare of the A-people in population B is increased, the average welfare of the whole B population can be lower than in population A. Consequently, we get cases where the area of block B is smaller than the area of block A.

The violation of the Mere Addition Principle is granted by Ng but he holds that if we avoid functions of extreme concavity (that is, choose a value of *k* closer to 1), then the Mere Addition Principle can be preserved for more compelling cases:

> If the chosen function f(n) is not of extreme concavity, the Mere Addition Principle can be preserved for more compelling cases. By more compelling cases, I mean cases where the average utility of the added people is not very much lower than those of the preexisting people, and the number of preexisting people has not become very large, so that most people find it very compelling to agree that the situation with the added people is better than the original situation. (Some people are content to say that, at least, the situation cannot be worse.)[72]

It is true that Ng's principle complies with the Mere Addition Principle in these more "compelling" cases if he avoids theories of "extreme" concavity; yet this will have as a consequence that theory X' behaves more like Total Utilitarianism even with large populations and yield conclusions similar to the Repugnant Conclusion.[73] At any rate, this principle would still not comply with the Mere Addition Principle when the population is sufficiently large and, consequently, imply the Sadistic Conclusion: Theory X' yields that one sometimes can make a population better by adding people with negative welfare rather than positive. By adding a few people with positive but much lower welfare than the original people, or many people with slightly lower welfare, the average welfare will decrease more than when adding one person with negative welfare. When it comes to large populations, where *f(n)* is close to the constant *m* and theory X' mimics Average Utilitarianism, theory X' yields that it is better to add the unhappy person.

Ng's principle also has counter intuitive consequences when it comes to populations with negative welfare. An uncontroversial condition of acceptability is the negative counterpart of the Mere Addition Principle:

[72] Ng (1989), p. 249.

[73] A better way to proceed is to use an asymmetric concave function, a function that is more curved towards the end than in the beginning. This could reflect an intuition that the value of quantity starts to decrease at a certain level; when adding people with the same quality, they contribute the same value to the population as long as the population "has not become very large" (relative the average quality of the population). This could be achieved by combining Ng's function with Total Utilitarianism: Let the value of quantity increase linearly up to a certain limit and, when the limit is passed, let the increase slow down asymptotically. Such a principle accepts all Mere Additions as long as the quantity is below the limit. However, it shares with theory X' all the negative features mentioned below.

> *The Negative Mere Addition Principle*: For any population, if by Mere Addition one adds a number of individuals with negative welfare to create a new population, then this new population is worse than the original one.

Ng explicitly claims that he sees no reason for an asymmetrical weighing of positive and negative welfare. The average of negative welfare should be treated in exactly the same way as the average of positive welfare:

> I find the asymmetrical treatment of utility and disutility unconvincing. No matter how great is the disutility, it can always be compensated by a sufficient big amount of utility. This is true for most of our personal choice and I see no reason for its rejection in social choice.[74]

Assume that the average welfare of the A-people is negative in figure 6.1-1, that *Q* is less than zero. In B, we have added persons who will be better-off but still unhappy. In cases where the average is negative, the best population is the population that is represented by the *smallest* area. Ng is therefore forced to judge the B-population as better than the A-population, despite the fact that the only difference between A and B is that B consists of all the unhappy A-people plus *m unhappy* people! Consequently, theory X' does not fulfil the very compelling Negative Mere Addition Principle.

Ng believes that theory X' is what Parfit is after. He does not believe that it is Parfit's theory X, since Parfit requires that theory X espouse the Mere Addition Principle and, as we saw above, theory X' violates this principle. However, Ng claims that, disregarding the Mere Addition Principle, theory X' meets all of Parfit's requirements and may be exactly the theory he is after.[75] This is unmistakably false. Parfit rejects Average Utilitarianism exactly on the ground that it does not give enough weight to negative welfare, referring to an example similar to the one we used.[76] In cases like those, theory X' and Average Utilitarianism go hand in hand.

Common to Average Utilitarianism and Theory X' is that both give less weight to suffering than Total Utilitarianism does. Although not all of us are convinced negativists who regard suffering as morally more important than happiness, surely an acceptable theory of beneficence must at least give as much weight to suffering as it gives to happiness.

6.2. Sider's Principle

[74] Ng (1989), p. 247, fn. 13.

[75] See Ng (1989), p. 245.

[76] Parfit (1984), p. 422. Parfit describes what he calls "Hell Three": "Most of us have lives that are much worse than nothing. The exceptions are the sadistic tyrants who make us suffer. - - - The tyrants claim truly that, if we have children, they will make these children suffer slightly less. On the Average Principle, we ought to have these children. - - - This is another absurd conclusion."

A second way of constructing a Variable Value Function is to dampen each person's contributing value. Sider has proposed a theory of this kind:[77]

[77] Sider (1991), p. 269. Sider's version differs from the one defined above in that his principle operates over whole possible worlds and life utilities. This leads to further problems as will be discussed below, in section 6.5, "Egyptology, Futurology and Astronomy."

Group a population into two ordered sets:

P: ($p_1 . . . p_i . . . p_n$) - the people with positive or zero welfare, *in order of descending welfare* - in case of ties, any order for those tied will suffice.
N: ($p_1 . . . p_j . . . p_m$) - the people with negative welfare, in order of ascending welfare.

Let u_i be the welfare of p_i from P.
Let v_j be the welfare of p_j from N
Let k be some real number greater than but close to 1.

$$GV = \sum_{i=1}^{n} \frac{u_i}{k^{i-1}} + \sum_{j=1}^{m} \frac{v_j}{k^{j-1}}$$

This principle will not violate the Mere Addition Principle. Adding one person with positive welfare u, $u_i \geq u \geq u_{i+1}$ has the consequence that u is inserted into the summing sequence between u_i and u_{i+1}. In both cases the summing sequence will be the same up to u_i. Then we have (follows from definitions):

$$u/r^{i+1} \geq u_{i+1}/r^{i+1}$$
$$u_{i+1}/r^{i+2} \geq u_{i+2}/r^{i+2}$$

and so forth. Finally, there is one extra positive term in the new sequence. When all terms in an ordered sum A are greater or equal to their counterparts in another ordered sum B and there is an extra positive term in A, then, of course, the sum A must be greater than the sum B.[78]

GV avoids the Repugnant Conclusion by being a convergent sum. When there is perfect equality, GV approaches *Q/(1-1/r)* which is of finite value; that is, applied to large population sizes, GV mimics Average Utilitarianism. When it comes to small populations, GV mimics Total Utilitarianism. Consequently, this principle avoids a Reversed Repugnant Conclusion.

While this principle may seem promising, it is nevertheless flawed.

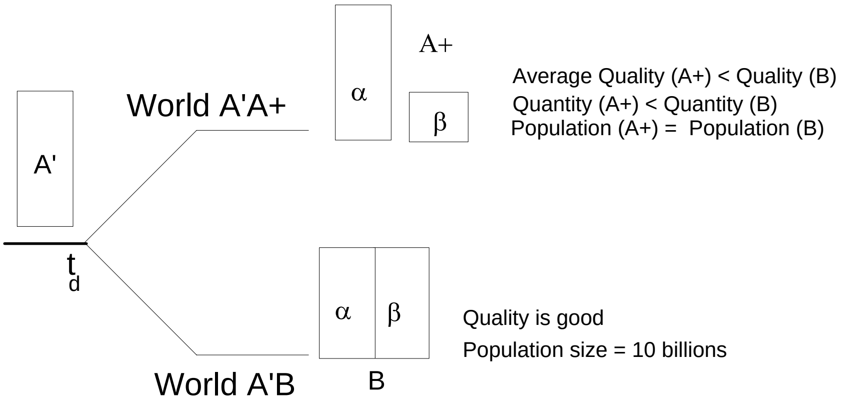

A+
Average Quality (A+) < Quality (B)
Quantity (A+) < Quantity (B)
Population (A+) = Population (B)
World A'A+
α
β
A'
t
d
α
β
Quality is good
Population size = 10 billions
World A'B
B

Figure 6.2-1.

[78] For a formal proof, see Sider (1991). Cf. Ch. 3, section 8.2.

Suppose that when we have a population of 5 billion people with good quality, the value of extra quantity is close to zero (i.e., $u_{n+1} / k^{n} \approx 0$, where $u_{n+1} \leq$ good quality). In the figure above, population B has higher total welfare, higher average welfare, and it is more equal than population A+; yet, Sider's principle would rank A+ as better than B. The reason is that the welfare of the β-people will count for much less than the welfare of the α-people. On GV, the α-people's welfare will be dampened much less than the β-people's. Consequently, the small losses for the α-people cannot be outweighed by the greater gains of the β-people. This would, perhaps, be acceptable if this was a nonexclusive Same Number Choice. We could then argue that we should not improve the welfare of the contingent people at the expenses of the necessary people's welfare. As we argued above in section 5, this argument is doubtful when we have unavoidable additions. Even worse, Sider's principle would rank A as better than B in exclusive cases. We can buttress this objection with the following example:

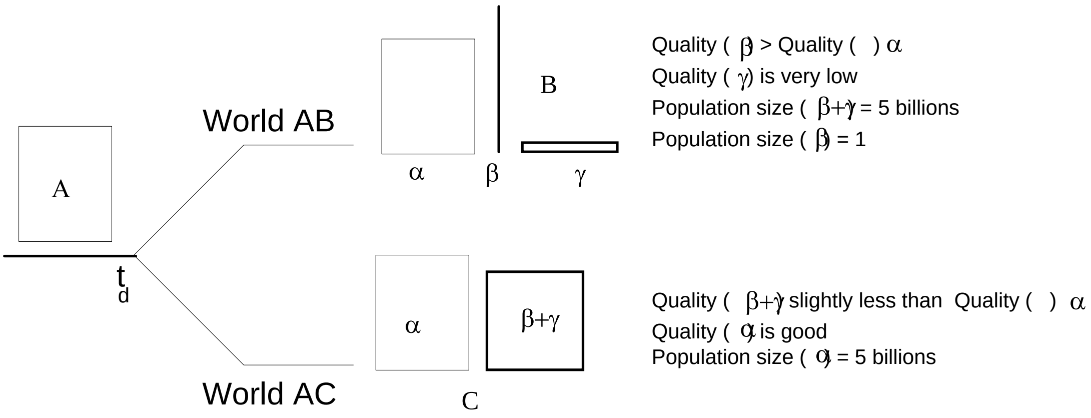

World AB
A
t_d
α
β
γ
B
Quality (β) > Quality ( ) α
Quality (γ) is very low
Population size (β+γ) = 5 billions
Population size (β) = 1
World AC
α
β+γ
C
Quality (β+γ) slightly less than Quality ( ) α
Quality (α) is good
Population size (α) = 5 billions

Figure 6.2-2. An anti-egalitarian case.

Here, GV ranks B as better than C for the same reason as above. In alternative B, β's welfare will not be dampened at all but the welfare of the γ-people will be strongly dampened. This evaluation is grossly anti-egalitarian. We could also imagine a Same People Choice where the original population and one of the alternative populations are like alternative C but with perfect equality. Still, GV would rank B as better than C, although C contains more welfare and is more equal. Sider's principle makes it a duty to enforce inequalities even when such enforcement lowers the total quantity of welfare! GV implies the following conclusion:

> *The Anti Egalitarian Conclusion*: For any possible population of at least two persons with positive welfare, there must be some imaginable population with positive welfare, which has the same number of people, less total quantity of welfare and less equality, whose existence, if other things are equal, would be better.

In fact, Sider himself does not advocate GV as a "theory X" and his reason is, *inter alia*, that GV proclaims unjust distributions of welfare.[79] If we look on the

[79] Sider (1991), p. 270, fn. 10.

negative side of welfare our reasons for not advocating GV become even stronger. Assume that the world is crowded by lots of people, all living in the same hell full of illness and pain. Let us say that we ponder whether to add two more people or not. One of these added people will have a life barely worth living. The other one will have the kind of hellish life that is commonplace in this world. Since the number of unhappy lives is great the negative value of the extra unhappy life will be small - the weight assigned to her life will be small. The extra happy life will be the only happy life in this world and therefore must be assigned the weight 1. Consequently, the negative value of the extra unhappy life will be outweighed by the positive value of the life barely worth living. According to Sider's principle, it is better to add the life barely worth living and the hellish life than to refrain from creating them.

6.3. Outline of a Possible Theory

Every theory we have so far discussed has been seriously flawed in one or another way. Even so, there are further problems that we have not yet discussed. These problems will be easier to discuss if we first present what we think is the best possible theory for evaluating the value of a population. We shall then discuss a problem that cannot be avoided by any welfarist principle of beneficence when it comes to future generations. After that, we return to the problem of the relevant demarcation of a population, a problem we partly discussed in section 3.2, "The demarcation of a population," when discussing how to evaluate actions that affect more than one population, that is, the complete principle of beneficence with respect to future generations.

In Chapter 3, section 8.1, "Theory WUN," we presented our proposal for the best possible theory to solve problems of intra- and interpersonal compensations. Now, our population theory will be an extension of this theory. Thus, our value function for calculating the value of a population looks as follows:

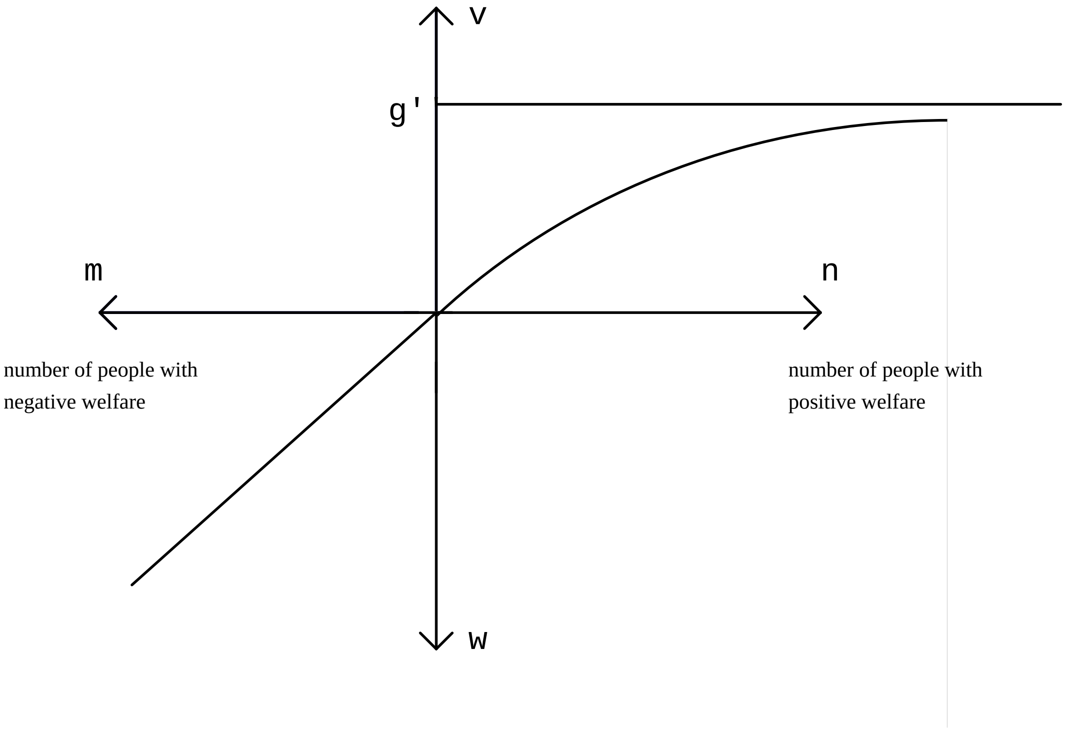


Figure 6.3-1. The value function PV.

The figure shows how the two functions *v* and *w* would behave when we vary the number of people but keep the welfare of every person constant (but not zero). A particular person's welfare is calculated according to the value functions we presented in Chapter 3, section 8.1, where we used an analogous function to the one above to calculate the value of particular moments and used those values to calculate the welfare of a particular period of a person's life.[80] To calculate the value of a population, the *population value,* group the people within the population into two sets: One set comprising the people with positive or zero welfare and one set comprising the people with negative welfare. Order the values of the former set, the *p*-values, in order of descending positive value ($p_1, p_2, \ldots, p_n$), where p_1 consequently is the greatest positive value and p_n the smallest positive value - in case of ties, any order for those tied will suffice. The set of negative values ($n_1, n_2, \ldots, n_m$), the *n*-values, need not to be ordered. Let *k* be some real number greater than but close to 1. The population value is then:

$$PV = v + w = \sum_{i=1}^{n} \frac{p_i}{k^{i-1}} + \sum_{j=1}^{m} n_j$$

[80] In Ch. 3, section 8.1, we talked about the *life time value* for a particular person. As we pointed out, however, the relevant partition of a life is a difficult problem. When it comes to intra- and interpersonal compensations, whole lives or quite long parts of lives seem to be the partition that matches our considered beliefs best. When it comes to population problems, our intuitions are tied to short time slices. This raises a problem: If our intuitions are tied to different ways of partitioning lives, how can we then construct an unified theory of beneficence? For further discussions of this topic, see Ch. 5.

This implies that there is a limit g'' for the positive value of a population:

$g'' = g' \text{ x } 1/(1\text{-}1/k)$

Here, g' stands for the maximum welfare of a life or a period of a life.[81] Irrespective of how large a population is, and how great positive welfare each person has, the value of the population cannot exceed g'' but approaches g'' asymptotically. Even if, *per impossibile*, each person's period consisted of an infinite number of positive moment each having an infinite positive utility, the welfare of each person's period would still not exceed g'. Moreover, irrespective of how many such heavenly lives a population consists of, the value of the population cannot exceed g''. There is, however, no limit to the negative value of a set of unhappy people.

These conditions taken together do not describe a particular theory. Rather they define a set of theories, some of then differing a lot. By varying the weight k and the limit g' we get different population principles, ranging from the most extreme ones, where g' is low and k is large to the more moderate ones, where g' is large and k is close to 1.[82] Moreover, we have not yet said anything about how to calculate the value of a whole world. An easy solution would be a mere totting up of the population values but as we shall see in section 6.5, "Egyptology, Futurology, and Astronomy," there are further problems hidden here. Let us first take a look at one remaining problem with this principle on the population level.

6.4. Anti Egalitarianism

We saw above that Sider's principle GV implies the Anti Egalitarian Conclusion. We can contrast GV with the following principle:

> *The Non-Anti Egalitarianism Principle*: If a population A consists of the same number of people as an alternative population B, and there is perfect equality in A, and higher total welfare in A than in B, then A is better than B.[83]

Ng's theory embraces this compelling principle but Sider's does not. As we saw above, Ng's principle X' violates the Mere Addition Principle. This is not surprising because no principle can simultaneously embrace the Mere Addition Principle, avoid conclusions similar to the Repugnant Conclusion, and embrace the Non-Anti Egalitarianism Principle.[84] Consider the following alternatives:

> A: One billion people all with high welfare.
> A+: The population in A plus a Mere Addition of one billion trillion people with low but positive welfare.

[81] See Ch. 3, section 8.1, where g was defined as the limit for the maximum value of a moment and g' as the limit for the maximum value of a life or a period of a life.

[82] More exactly, g' is a function of g and α, where g is the maximum value of a moment and α is the weight given to positive moments. That is, by varying g and α, we can vary g'. See Ch. 3, section 8.1.

[83] See Ng (1989), p. 238. Ng formulates his principle in terms of "same set of individuals" and "higher total utility."

[84] Ng makes this point in Ng (1989), p. 240.

> B: The same population as in A+ but with slightly higher total welfare *equally* shared by all.

The Mere Addition Principle implies that A+ is not worse than A. It follows, given comparability, that A+ is at least as good as A. The Non-Anti Egalitarianism Principle yields that B is better than A+. Since B is a repugnant alternative relative to A, B is not better than A. If B is better than A+ and A+ is at least as good as A, then B must be better than A. Hence, these valuations imply a contradiction: B is better than A and B is not better than A - we have to jettison one of the underlying conditions that lead to this contradiction. We cannot see any good reason to give up comparability.[85] Hence, we are left with the Mere Addition Principle, the Non-Anti Egalitarianism Principle, and the avoidance of the Repugnant Conclusion. One of these conditions must go.

Our opinion is that among these conditions the avoidance of the Repugnant Conclusion is the most important one. If we cannot construct a welfarist theory that avoids the Repugnant Conclusion, then, we hold, one has to accept that there is no welfarist theory that can accommodate our considered beliefs about moral duties to future generations. Thus, our choice is between the Mere Addition Principle and the Non-Anti Egalitarianism Principle.

As we saw in sections 4.3, "Mere Additions," and 6.1, "Ng's Principle," there is a vexatious relation between the Mere Addition Principle and the Sadistic Conclusion - if a principle violates the former then one can suspect that it implies the latter. If one can make a population worse by adding people with positive welfare, then one can construct cases where an addition of one person with slightly negative welfare decreases the value of a population less than an addition of many people with low positive welfare. This holds for all welfarist principles that assign a finite weight to negative welfare and state that the negative welfare of lives or periods of lives is not dependent on the positive welfare of other lives or periods of lives. Consider a case where a population A contains n persons with positive welfare and another population B contains $n + m$ persons with positive welfare and the m-people are added by Mere Addition, i.e., the conditions for the Mere Addition Principle are fulfilled. Suppose we have a principle which yields that the addition of the m-people lowers the population value with d units - the value of population B is d units lower than the value of population A. Ng's theory X' is an example of a principle which can imply such valuations. Consider a third population C which differs from A only in respect to an extra person with negative welfare. Since a person's negative welfare is independent of other people's welfare, we can easily suppose that the unhappy person in C has a negative welfare v smaller than d. Moreover, when negative welfare is given a finite weight w, then we can construct cases where w x v is smaller than d. Thus, a principle that ranks B worse than A will rank C as better than B, i.e., we can make an outcome better by adding people with negative welfare rather than positive welfare, a clear instance of the Sadistic Conclusion.

It might be objected that if negative welfare is given infinite or great weight then these situations will never occur in realistic cases. With such great weight on

[85] Moreover, sacrificing comparability does not solve our problems. See Parfit (1984), pp. 430-37.

negative welfare, however, the problems that confronted the strong negativist and the lexical negativist will reappear. In short, such theories permit all too few compensations.[86]

One might also argue that the Sadistic Conclusion could be avoided if one makes a persons negative welfare dependent on other people's positive welfare. To let the value of unhappy people vary with the value of other happy people, however, is an utterly strange axiological principle.

Where does this leave us? We could abandon the Non-Anti Egalitarianism Principle. Abandoning this principle, however, leads us to the dreadful Anti Egalitarian Conclusion. Compare the following populations A and B. A contains two persons with welfare $m > 0$. B contains one person with the welfare $m+x$ and another person with the welfare $m-z>0$, $0<x<z$. Consequently, there is perfect equality in A as well as a higher total of welfare. There are only two interesting differences between these populations. First we have the difference between the highest welfare in B and A: $m+x-m=x$. Then we have the difference between the lowest welfare in B and A: $m-z-m=-z$. The difference in population value between B and A is thus: $x/k^0-z/k^1 = x-z/k$. Now, for any $k>1$, there exists an x and z such that $z/k<x<z$, that is, we can always construct a population B that has higher population value than population A even though B is more unequal and has less total welfare. This can be generalised to any population A with at least two persons with positive welfare: one can always subject the two persons with the highest and lowest welfare to the same process as above. This is a major drawback of PV. To judge a more equal population with higher total welfare worse than a less equal population with less total welfare is not reasonable.

We now have a triad of objectionable conclusions: the Repugnant Conclusion, the Sadistic Conclusion and the Anti Egalitarian Conclusion. One could argue that this amounts to an impossibility proof for the existence of an acceptable welfarist theory. We think that such a conclusion is reasonable, but in the absence of any other theory that can accommodate our beliefs about the weight of evil and moral duties to future generation, we think that it is more reasonable to develop the best possible welfarist theory to deal with such problems. Such a theory avoids the Repugnant and the Sadistic Conclusion but implies the Anti Egalitarian Conclusion.

6.5. Egyptology, Futurology and Astronomy

Sider and Ng use a population concept that encompasses all people that will ever live. As we argued in section 3.2, this goes against the ordinary language meaning of "population" to which our intuitions about population questions are tied. There are, however, further arguments against the "timeless view." Consider the following conclusion:

> *The Egyptology Conclusion*: Whether it would be good or bad that a person comes into existence, will depend on facts about the welfare and the number of the Ancient Egyptians.

[86] See Ch. 3, section 4.1 and 5.1.

This will hold for all principle that gives any weight to quality of welfare, which violates the Mere Addition Principle and makes use of a timeless population concept. To calculate the quality we use the number of people as denominator; thus, the number of Ancient Egyptians will affect the quality. Furthermore, their welfare will affect the numerator. If there was a huge population in Egypt and the welfare was very high, then it could be the case that it would be bad to bring somebody into existence today even though that person would have a welfare higher than anybody else in the existing population. This is a devastating objection to the timeless view. As Parfit nicely puts it, "research in Egyptology cannot be relevant to our decision whether to have children."[87]

Sider's principle does not imply the Egyptology conclusion: Whatever welfare the Ancient Egyptians enjoyed, it will always be good to add a person with positive welfare. The value Sider's principles assigns to the addition of a specific person will, however, be affected by the welfare of the Ancient Egyptians: How good it would be that a person comes into existence, will depend on facts about the welfare and the number of the Ancient Egyptians. The value of two persons can differ dramatically even though their welfare does not differ that much. If one of the persons has a welfare slightly higher than the Ancient Egyptians and the other one has a welfare slightly below, then the value of the first person will be dramatically higher than the value of the second. There is, however, a worse problem with Sider's principle in combination with the timeless view. To avoid a Repugnant Conclusion when evaluating the first generation of people, one needs a weighing coefficient such that the value of extra quantity is close to zero when we have five billion people with good quality. With such a coefficient, the value of quantity will be strongly dampened in all subsequent generations and we get a Reversed Repugnant Conclusion of the worst sort: One person with slightly better welfare can outweigh any number of people with slightly less welfare. In other words, Sider's principle in combination with the timeless view implies either a Repugnant Conclusion or a Reversed Repugnant Conclusion of the worst sort.

It could be tempting to argue that only future people's welfare should count. The following conclusion should dispel that illusion:

> *The Futurology Conclusion*: Whether it would be good or bad that a person comes into existence, will depend on facts about the welfare and the number of the people living billions of years from now, even though this decision will not affect these future people's welfare.

We get the same problem as above with the only change that future populations take the place of the Ancient Egyptians. We have to reject the timeless and the future oriented view. Should we then use theory PV to calculate the value of all time slices and then just tot them up? We get the following conclusion:

[87] Parfit (1984), p. 420.

> *The Time Repugnant Conclusion*: For every possible world A with good quality there exists another possible world B with quality close to zero but with greater quantity of welfare because in this world there will exist much more people who are spread over a much longer time span than the fewer persons in world A. Therefore, B is ranked as better than A.

This is implied both by theory PV and Total Utilitarianism, but it is a more acute problem for the former theory. When people are more thinly spread over time, less dampening will take place when calculating the value of the different time slices. Consequently, a world which contains *less* quantity of welfare but with people more thinly spread over time could be ranked as better by PV than a world with more quantity but with people less thinly spread over time. We could avoid this conclusion by discounting the value of extra time slices with happy people in the same way we dampened the value of extra happy people to avoid the Repugnant Conclusion. This solution, however, would introduce new but familiar problems. If one were to use the type of concave functions Ng made use of above one would get variants of the Egyptology and Futurology Conclusion: Whether the birth of a child is good or bad will partly depend on the fact of how long sentient beings have been existing in the universe, and on how long they are going to exist. One would also get a version of the Sadistic Conclusion: It could sometimes be better to add a population with negative value rather than several populations with positive value. Using a concave function of Sider's type would avoid these problem but introduce a version of the Anti Egalitarian Conclusion: For any possible world with at least two populations with positive value, there must be some imaginable world with positive population values, which has the same number of populations, less total population value and less equality among populations, whose existence, if other things are equal, would be better. Moreover, the relative ranking of two populations would depend on the welfare of the Ancient Egyptians or populations in further future.

We seem to be trapped between the Egyptology, Futurology and the world version of the Anti Egalitarian Conclusion on the one hand, and the Time Repugnant Conclusion on the other hand. This problem is of a more general character, however, as another problematic conclusion shows :

> *The Astronomy Conclusion*: Whether it would be good or bad that a person comes into existence, will depend on facts about the welfare and the number of the people living billions of light years from here, even though this decision will not affect these people's welfare.

This conclusion cannot be avoided by changing the way we add up the value of time slices or by adopting the timeless view. This points to the crux of the matter when one tries to calculate holistic values, values that depend on the structure of a whole: one must define the relevant whole to calculate. We have to deepen our analysis of what is the relevant demarcation of a population.

6.6. The Relevant Population Concept

In section 3.2 we argued for a demarcation of population in time. The Astronomy Conclusion shows that this is not a sufficient demarcation of a population. We have further beliefs about the appropriate demarcation of a population in a moral or axiological context. The Oxford English Dictionary, second edition, gives two definitions of "population" that is relevant in this context:

> 1. A peopled or inhabited place.
>
> 2. `The state of a country with respect to numbers of people'; the degree in which a place is populated or inhabited; hence, the total number of persons inhabiting a country, town, or other area; the body of inhabitants.

Both of the above definitions connect the concept of a population to a physical area. This could be one way to proceed. We could argue that the relevant demarcation is the planet that a population inhabits. When calculating the value of our different possible acts, the relevant population would thus be the people inhabiting the planet Earth. This would not be sufficient, however, when we consider an act that affects people both on Earth and a future Mars which is inhabited. We would then have to consider the effect of our acts on the Martians. We could still hold that planets are the relevant demarcation but that we have to calculate the effect our acts have on the population on Earth and the population on Mars. When computing the value of the effect our act has on the population on the planet Earth, we only use the number of people on Earth in the denominator and we proceed in the same way with the population on Mars. To determine the desirability of the act, we sum the value of both the resulting populations in some way. We can compare this to the work of an art museum curator. The demarcation of an art collection could be said to be the space where it is exhibited. When our curator is contemplating an addition to the collection it is irrelevant to her decision, from an aesthetic point of view, how other collections in other museums are composed. On the other hand, if she is in charge of two museums, and offered to buy two paintings by the bulk where each of them might be suitable for one of the collections, we think that she should evaluate the effect on each collection in isolation. To get the overall aesthetic value of the purchase, our curator should then "sum" the respective value increase or decrease of the two collections.

A problem with this approach is that to use planets as demarcations for populations seems quite arbitrary. We could imagine that two populations inhabit a planet in the sense that they have no contact whatsoever and they do not affect each other in any way. It seems awkward that the welfare and the number of the first population should have any bearing on whether it would be good, or how good it would be, to create a child in the second population. It seems that when it comes to moral questions we have beliefs about the relevant demarcation of a population which are better captured by the concept of a *community*. The Oxford English Dictionary give three meanings of "community" which is relevant in this context:

> 1. A body of persons living together.
>
> 2. A body of people organised into a political, municipal, or social unity.

3. Life in association with others; society, the social state.

All of the above explications stress some kind of continuous interaction among the members of a community. A tentative definition of a population could then be a set of people with continuous overlapping chains of interactions. I live in the same community as you if you have continuous interactions with me or with somebody else who has continuous interactions with me or with somebody else who has continuous interactions with me or . . . and so on. On this explication most people on earth live in the same moral community. An extreme hermit would not be part of this community since she would not have any interactions at all.

We think that this is the right path to walk but we doubt one can give an exact definition of what counts as a community. Just to take one problem: What should count as "continuous interactions"? This is not an axiological or moral question but rather a metaphysical one. We can compare our problems to define the relevant conception of a population with our conception of a person. Here we have a quite clear ordinary language meaning that we can make use of. When we start to scrutinise this concept, however, we can easily see that we get demarcation problems similar to the ones we have with the concept of a population. Derek Parfit has proposed a conception of the person that roughly consist in overlapping chains of memory.[88] With such a conception a person can survive for ever and one part of this life can be totally different from another part. A latter part of this life can be totally empty of memories of an earlier part and look upon that part as a totally different "person." Here it is no longer clear whether this life should be counted as one person's life or many people's lives. We leave unsolved the problem of providing an exact definition of the relevant axiological whole, both when it comes to persons and populations.[89]

Even if we cannot give an exact definition of "overlapping continuous interaction" we still think the concept can be useful because we have some clear examples and counterexamples of such interactions. Hermits and sentient beings on other planets, if they exist, do not have overlapping continuous interactions with us; we do have overlapping continuous interactions with most people on earth.

We are going to use a population concept which both stresses contemporaneity and overlapping chains of interactions among the members of the population. We think the right way to sum these populations is to use the same concave function we have used earlier, an asymptotic function that always gives some positive value to a Mere Addition of a new population with positive population value. Such a theory avoids the Sadistic, Egyptology, Futurology, Astronomy and the Time Repugnant Conclusion and embraces the Mere Addition and the Pareto Addition Principle. Moreover, unlike Sider's theory, it will not be trapped between the Repugnant and Reversed Repugnant Conclusion. We have to accept, however, the two versions of the Anti Egalitarian Conclusion, the population version and the world version.

[88] See Parfit (1984), pp. 215-17.

[89] Cf. our concluding paragraph in Ch. 3, section 8.3.

Chapter 5

SUMMARY AND ENERGY APPLICATIONS

1. Summary Chapter 3 "The Weight of Evil"

In Chapter 1, the examples 1, "The Uranium Mining", and 2, "The Power station", raised the question if and when happiness of some people can compensate the sufferings of others. In answering this question our point of departure was the firm intuition that unhappiness and suffering have greater weight than happiness. By taking this stand we revealed ourselves as members of the negative utilitarian family. The problem was then to find out which members of this family we want to join, and to spell out why we do not want to be as some of our siblings.

First we had the choice between the two main alternatives within the negativist family: strong negativism, according to which all weight is given to disutility, and weak negativism, according to which some weight is given to utility but more weight is given to disutility. It was argued that we should not join the strong pure negativists, who are exclusively concerned with minimising pure disutility. One important reason for not joining this group was that even strong pure negativists give far too small weight to disutility. The value of negative gains can always compensate the value of losses, irrespective of how small the gains are and how great the losses are. For if we have a sufficient great number of winners, compensation is a fact. Hence, the strong pure negativists do not avoid the Absurdity, the avoidance of which is a very compelling condition of acceptability.

Furthermore, we showed that the drawback of giving to small weight to disutility also holds for the strong mixed negativists, whose only objective is to minimise mixed disutility. This negativism does not fulfil the compelling Negative Pareto Principle. If we have two satisfactory futures for a person, where the only welfare difference between these futures is that each negative moment in the first future has less disutility than each negative moment in second future, then the strong mixed negativist must judge the futures as equally good. We therefore concluded that the strong negativisms are to be avoided.

So, we saw that if we want to be negativists we must join the weak negativists. In this camp we distinguished between the weighted negativists and the lexical negativists. The latter interpret the concept of weight as lexical weight. Disutilities have greater lexical weight than utilities which means that differences in disutility can never compensate differences in utility. Only when the disutility is unaffected can differences in utility make a value difference. We argued that this negativism is too rigid when it comes to trade-offs between unhappiness and happiness. It is not true that any suffering can be compensated by happiness. But it is neither true that marginal and trivial sufferings can never be compensated by happiness. If the suffering is trivial, then it should be possible to compensate it. Due to its lexical structure the lexical negativism cannot embrace this reasonable flexibility. We also showed that some impersonal versions of this negativism, despite their rigidity, do not avoid the Absurdity.

Having dismissed the lexical negativisms we set our hope to the weighted negativism according to which disutility and utility are weighted in the sense of being multiplied by a certain positive number; the utility being multiplied by a smaller number. The value of a state is then seen as the sum of the weighted disutilities and utilities occurring in the state. We must distinguish between different versions of this kind of negativism. The equal weighted negativism gives the same weight to every disutility and the same weight to every utility. The unequal weighted negativism gives different weights to different disutilities and/or different weights to different utilities. The equal weighted negativism was easily dismissed. First, it does not avoid the Absurdity. Second, it avoids neither the Positive Repugnant Conclusion, nor the Negative Repugnant Conclusions, where the latter implication is especially worrisome. For instance, the Negative Repugnant Conclusion (2) states that for any number of very happy moments, and for any number of very unhappy moments, there is a number of moments each having a positive utility close to zero such that the whole composed of the dull moments and the very unhappy moments has greater value than the whole composed of the very happy moments. The problematic feature here is, of course, that if one give *some* weight to positive moments, and moreover give the *same* weight to them, then for any life containing some sufferings the value of the life could in principle be made how great you want just by adding positive moments.

Before we investigated the possibilities of the unequal weighted negativism, we asked whether we should abandon our focus on sums of disutility, and instead be concerned with *average* disutility, i.e., disutility divided with the number of moments or persons. Our answer to this was in the negative, since the average approach led us to prefer a crowded hell to a less crowded one, when the units (persons or moments) occurring in these hells all have the same disutility. In other words, the average approach does not fulfil the Negative Mere Addition Principle that states that adding negative moments always makes the world worse.

By the method of elimination we ended up with the unequal weighted negativism. So, finally, in the last section we sketched a negativism of this kind, theory WUN, (Weak Unequal weighted Negativism), and showed that this theory of beneficence fulfils all the conditions of acceptability we had stated so far. As one could expect from a complete moral theory that always has something to say, WUN has its own drawbacks. The most important drawback is that WUN does not take a proper interest in the welfare of *parts* of lives. For instance, WUN does not prevent that the welfare of one part of a life with negative lifetime value is sacrificed for the welfare of another part of another person who also has a negative lifetime value. Admittedly, WUN is in this respect inadequate, but we think that a theory that gives some special concern to periods must nevertheless use a value function similar to the one used in WUN. That is, in the same way that WUN dampens positive lifetime values and gives greater weight to negative lifetime values, the period oriented theory must dampen positive *period values* and give greater weight to negative *period values*. This means that even if WUN is too lifetime oriented, it incorporates a general idea on how to weigh positive values against negative ones, which may be used in other contexts where the focus is on other units of a life.

Furthermore, even if we want to give more weight to periods of different persons' lives, it seems to be wrong to give *all* weight to periods and completely abandon the lifetime perspective. To give all weight to periods means that instead of moving from moment value to lifetime value, we move from moment value to period value. Each life is divided into periods and the value of a period is a function of the value of the moments within that period. The value of a world is then seen as a function of the value of the periods in that world. Hence, the theory does not discriminate between aggregation of values of different periods from the *same* person's life and aggregation of values of different periods from *different* persons lives. This means that *intra*personal interperiodical compensations are treated exactly in the same way as *inter*personal interperiodical compensations. We think that this impersonal feature makes this theory counter-intuitive. (For more on this theory, see below section 5.3)

2. Summary Chapter 4 "Moral Duties to Future Generations"

In Chapter 1, we depicted a scenario, "The Energy Consumption of Present People," which raised the question whether we, the present people, are justified in discounting the welfare of future people. Are happiness and sufferings in the further future of less importance than happiness and sufferings among present people? In section 2, "Social discount rates," we looked at the six most plausible arguments for a social discount rate and concluded that none of them could justify a general application of a social discount rate. Remoteness in time does correlate with a whole range of morally important facts, as does remoteness in space. None of these correlations, however, are of such a nature that they can justify that we care less about the effects our social policies have in the future, at some constant rate *n* percent per year. In some specific situations, there can be good reasons to care more for people close to us, in space or time, but generally, we ought to be equally concerned about the foreseeable effects of our policies, irrespective of when in time these effects occur.

Many moral theories make use of a Person Affecting Restriction: An act can only be good or bad if it is good or bad for a specific person. Theories that make us of a complainant invoke this restriction: an act is only bad if there is somebody who can complain that they have been harmed or made worse off. Theories of this kind run into problems with scenarios like the one we outlined in example 4, "Different Energy Systems - Different People." In such cases, the identities of future people are not determined but contingent upon our decision, i.e., the Non-Identity Problem. We rejected as strongly counterintuitive the claim that the specific identity of people should make a moral difference. We continued in section 5, "Person Affecting Restrictions" to show that there is no way to amend theories that make use of the Person Affecting Restriction by restricting their scope and combining them with an impersonal theory. Rather than solving any problems in population ethics, these theories create new, intractable, problems.

One of the best arguments for a Person Affecting Restriction, perhaps the main argument, is that it embraces the Asymmetry: The fact that a person's life would not be worth living constitutes a strong moral reason for not bringing her into existence, while the fact that a person's life would be worth living provides no or only a weak moral reason for bringing her into existence. We argued that a weaker asymmetry was more compelling, one that holds that a) we have stronger reasons to avoid creating unhappy people than we have for creating happy people; b) increase in existing people's welfare is more important than adding happy people; and c) increasing suffering by adding unhappy people or by making existing people (more) unhappy is equally bad. We pointed out that such an asymmetry could be retained in an impersonal theory that gives more weight to suffering and to individual quantity of welfare than to happiness and total quantity of welfare. We concluded that only pure impersonal theories could be possible candidates for a principle of benevolence towards future generations.

The two most popular impersonal theories, Total and Average Utilitarianism, have obvious drawbacks when it comes to population questions. Total Utilitarianism implies that a huge population, say 100 billion people, with very low welfare can be better than a population of, say 5 billion people, with a considerable higher level of welfare, an instance of the Repugnant Conclusion (example 5, "The Overcrowded Earth," in Chapter 1 illustrates this problem). Average Utilitarianism implies that a population consisting of only one person is better than an arbitrary large population with slightly lower welfare, an instance of the Reversed Repugnant Conclusion. The problem with Total Utilitarianism is that it only gives weight to total quantity of welfare; the problem with Average Utilitarianism is that it only gives weight to individual quantity of welfare. The solution then, would be to give weight to both total and individual quantity of welfare. One way is to give linear increasing value to both total and individual quantity of welfare. We investigated this possibility in section 4, "Linear value theories," and concluded first that such a solution would not avoid the Repugnant and the Reversed Repugnant Conclusion as long as there is no upper limit to the size of a population and to the quality of peoples lives. However, even if we could find a non-arbitrary reason to limit the size of a population and the quality of lives we need to consider, linear weighing principles have another serious flaw that cannot be amended. We showed that all principles that assign linear increasing value to quality and quantity of welfare violate the Mere Addition Principle and implies the Sadistic Conclusion: one can make a population better by adding an unhappy person rather than many happy persons. We reached the same conclusion for principles that assign linear increasing value to quality and quantity of welfare but make use of different kinds of limits to the value of quantity - lexical orderings and higher goods principles.

Having dismissed linear weighing principles, we turned to Variable Value Theories. These theories assign asymptotically increasing value to total quantity and linear increasing value to individual quantity of welfare. If we keep the individual quantity of welfare constant and increase the population size, then the value of the population will converge on a value limit asymptotically - a doubling of the population size without any increase in individual quantity of welfare will always do less than a doubling of the value. A doubling of the quality of welfare, on the other hand, will double the value of the population.

We first considered a theory suggested by Ng. His theory combines Average Utilitarianism with a concave function that uses the population size as input. According to this theory, the value of a population is the product of the average welfare and the dampened population size. We showed that Ng's theory has several flaws, most notably, it violates the Mere Addition Principle and implies the Sadistic Conclusion.

The second suggested Variable Value Theory, Sider's principle GV, was promising insofar as it was the first principle that both avoided the Repugnant Conclusion and embraced the Mere Addition Principle. As we pointed out however, this theory also has serious flaws. For example, it can sometimes prescribe that a hellish life can be compensated by a life barely worth living. Most important, it implies the Anti-Egalitarian Conclusion.

In section 6.3, we outlined what we think is the best possible population principle, theory PV. Our principle gives linear increasing value to increase in individual time slice value (individual quantity of welfare), linear decreasing value to total quantity of negative time slice value (total quantity of negative welfare) but asymptotically increasing value to total quantity of positive time slice value (total quantity of positive welfare). This theory shares one negative feature with Sider's theory: it implies the Anti-Egalitarian Conclusion. We showed, however, that no theory can avoid one of three conclusions: The Repugnant Conclusion, the Sadistic Conclusion and the Anti-Egalitarian Conclusion. We argued that of these three conclusions, the Anti-Egalitarian Conclusion was the least important one to avoid.

In section 3.2, "The demarcation of a population" we argued that the common language use of the word "population" and our intuitions in population ethics are tied to concept of population where people are contemporaries, the time slice view. In section 6.5, "Egyptology, Futurology and Astronomy" we buttressed this argument by showing that a timeless or a future-oriented population concept implies counterintuitive conclusions when combined with holistic value principles: The Egyptology and the Futurology Conclusion. We would have to accept that whether it would be good or bad that a person comes into existence, will depend on the welfare and the number of the Ancient Egyptians or people living billions of years from now, although these people's welfare would not be affected by this decision. Sider's principle in combination with the timeless or the future-oriented view implies either a Repugnant Conclusion or a Reversed Repugnant Conclusion of the worst sort. Should we then use theory PV to calculate the value of all time slices and just sum these values? No, because such a theory would imply the Time Repugnant Conclusion - one can construct worlds with lower individual and total quantity of welfare which are better than worlds with higher individual and total quantity of welfare.

Another problem showed that we had to deepen our analysis of the relevant demarcation of a population. The Astronomy Conclusion says that whether it would be good or bad that a person comes into existence will depend on facts about the welfare and the number of the people living billions of light years from earth, even though this decision will not affect these people's welfare. The counterintuitive character of this conclusion shows that we have more beliefs about the demarcation of the relevant whole when calculating holistic values than a demarcation in time reflects. We argued that the correct demarcation of the relevant axiological whole is partly captured by the concept of a community. This concept stresses some kind of continuous overlapping chain of interactions. We granted that we could not give an exact definition of what should count as a "continuous interaction" and that such a concept would run into analogous problems to those of a memory-based criterion of personal identity. However, a vague concept can be useful as long as it has clear examples and counter-examples. On our explication, most existing people on earth would be part of the same axiological whole but an extreme hermit, or a Martian, would not be part of this whole because they would not have any interactions at all with other people on earth. Our concept of the relevant axiological whole was then constructed as a combination of the concepts of a population and a community. First a demarcation in time extracted from the common language meaning of the word "population," then a demarcation extracted from the common language meaning of the word "community."

With this concept of the relevant axiological whole, we avoided the Egyptology, the Futurology and the Astronomy Conclusion. Moreover, unlike Sider's theory, it will not be trapped between the Repugnant and Reversed Repugnant Conclusion. This theory also embraces the Mere Addition and the Pareto Addition Principle and avoids the Sadistic Conclusion. To avoid the Time Repugnant Conclusion, however, we had to introduce another asymptotic function to aggregate population values. This meant that we had to accept another Anti Egalitarian Conclusion, the world version of this conclusion.

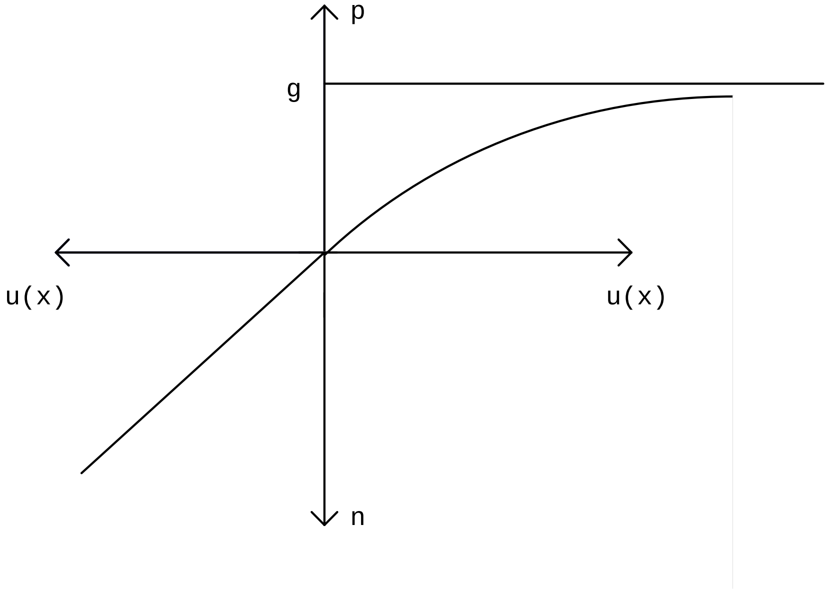
3. Our Theory

Our general theory of beneficence can be expressed by a set of rules as how to evaluate a possible world:

(1) Group the inhabitants in a possible world *w* into a set of exhaustive populations.

(2) For each population *p* in *w*, take for each individual *i* in *p*, the moments that occur in *i*'s life during the time *i* is a member of *p* and assign value to these moments according to the *Moment Value Function* (also described in section 8.1 in Chapter 3).

Moment Value Function



Aggregate the moment values in following way:

> the sum of the n-values plus ($\alpha^0 p_1$ + α^2 p_2+ α^2 p_3+ , . . . , + $\alpha^{k-1}p_k$), where k is the number of positive moments and 1 > α > 0, and the p-values are put in order of descending positive value

This aggregate, the *period value,* is the value of the period of i's life during which *i* is a member of *p.*

(3) For each population *p* aggregate the period values of the members of *p* in the following way:

> the sum of the negative period values plus ($\beta^0 P_1$ + $\beta^1 P_2$ + $\beta^2 P_3$ + , . . . , + $\beta^{l-1} P_l$), where *l* is the number of positive period values , and 1 > β > 0, and the positive period values are put in order of descending period value.

This gives us the *population value of p.*

(4) Aggregate the population values in *w* in the following way:

> the sum of the negative population values plus ($\chi^0 Pop_1$ +$\chi^1 Pop_2$ + $\chi^2 Pop_3$ + , . . . , + $\chi^{m-1} Pop_m$), where *m* is the number of positive population values, and 1 > χ > 0, and the positive population values are put in order of descending population value.

This gives us the *value of the world w*.

This is a most general scheme for a complete theory of beneficence. For, as discussed in Chapter 3 and 4, there are several ways to define the axiologically relevant wholes, both when it comes to periods and populations. Different definitions of a period and a population will give us different evaluations of worlds.

In Chapter 3, a period was defined as a person's lifetime and a population was for simplicity assumed to be the inhabitants of a whole possible world. In Chapter 4, the concept of an axiologically relevant population was shown to be problematic and subsequently defined as a set of people related to each other by overlapping chains of interactions and living during a period of time when the change in population size is insignificant to the problems in question. Admittedly, this definition lacks of precision. To give a complete theory of beneficence and not just a general scheme in which several theories fit, we would have to make the definition more exact. Furthermore, we would have to define the relevant periods within a life and decide the different weights used in the value calculus. (For more on this matter, see Chapter 3, section 8.1.) However, although our theory is vague in these respects, it has the precision needed to comment on the energy problems described in Chapter 1. So, we leave our theory at this unfinished stage, and turn to the energy applications.

4. Energy Applications

Now, when we have stated our general theory of beneficence, it is time to comment on the cases 1 to 5 given in chapter 1.

4.1. The Uranium Mining

Recall the story. People living nearby the mine suffer a lot due to the diseases they get from the radiation. But at the same time a vast number of people are benefited by the energy produced by the uranium from the mine. However, the nuclear consumers benefited by the uranium would be well-off without it, and are only slightly better-off with it. It was also assumed that the gains factually outweighed the losses. The crucial question was then whether the *value* of the gains compensates the *value* of the great losses.

It is clear that according to our theory factual outweighing is not sufficient for value compensation. To know whether the value of the gains compensate the value of the losses we need to know more about the welfare of the affected parties. So, let us therefore spell out the story somewhat more.

Assume that the lives of the people stricken with the illness are totally ruined, and hence unsatisfactory. Imagine that they have the most terrible kind of cancer, making them slowly die in pain. Here each difference in negative lifetime value equals *l*, and this is assumed to be a great difference, since they are each suffering from a horrendous kind of cancer. Furthermore, assume that the people benefited by the uranium live in a highly industrialised country with high living standard, and hence their lives are not just satisfactory, but also uniformly happy, i.e., good. Here each gain in positive lifetime value equals *g*, which is assumed to be a marginal difference, since the positive effects of the nuclear power is so evenly spread out among the inhabitants of the rich country.

If the positive lifetime values are dampened rather mildly, say with β = 0.99, the value limit of the gains in positive lifetime value is g x 1/(1-0.99) = g/ 0.01. This means that if the sum of differences in negative lifetime value is greater that g/0.01, then, irrespective of the number of inhabitants in the rich country, the value of the gains cannot compensate the value of the losses. So, even if there is a couple of hundreds suffering from the cancer, and 9 millions benefited by the uranium, then, given that the relation between the losses and the gains is as we have said, the value of the gains cannot compensate the value of the losses. This is in line with the opinion expressed in the quotation from Rescher (see Chapter 1, section1):

> We should surely not want to subject one individual to unspeakable suffering to give some insignificantly small benefit to many others (even an innumerable myriad of them)

Notice also that this result does not presuppose that happiness is given a very small weight. Hereby we avoid the problems that confront the strong negativisms.

On the other hand, if the sum of differences in negative lifetime value is smaller than g/0.01, then it is true that the losses *can* be compensated. But, of course, this does not mean that they are actually so in a world like ours. Perhaps, we have not a sufficient great number of winners, or perhaps, there are so many losers. Admittedly, our theory here shows some flexibility. But we think that this is reasonable. After all, happiness counts; it is just that more importance is attached to suffering.

4.2. The Power Station

Recall the story. In order to build a new power station some people have to move from the area where this station is planned to be situated. These people are highly attached to the area. Thus, if they move, they will be very frustrated. At the same time, if the station is built then a lot of other people will each gain some marginal welfare, and their gains will factually outweigh the losses of the others. The question was then: "Are we justified in building this power station?".

Again, we need to fill in with some details. In this case we do not think it is reasonable to say that the people attached to the area will have unsatisfactory lives if the are forced to move. It is more reasonable to say that they will have less satisfactory lives. Furthermore, assume, as in example 1, that the winners will have satisfactory lives anyhow.

With these details filled in, we know that, according to our theory, it is possible to compensate the losses. Irrespective of how great the losses in positive lifetime value are, they can be compensated by a sufficient great number of gains. Hence, if we have a sufficient great number of winners, then we are justified in building the power station. Is this counterintuitive?

It depends on whether forcing the people to move will make them very frustrated during a not insignificantly long period. If this is the case, and each of the winners is just marginally benefited, then, admittedly, our theory yields a counter intuitive result. But, as we already said, even if this is a major drawback of our theory, we claim that any adequate *period* oriented theory must make use of a value function similar to the one we have used.

On the other hand, if forcing the people to move will make them less happy, or marginally unhappy during some period, then it is not obvious that our theory give us the wrong result. Again, our theory has the reasonable feature of not being rigid when it comes to trade-offs.

Notice, however, that even if our theory judges it *possible* to compensate losses when all affected parties have good lives, this does not mean that actually the losses for the forced people are compensated. It is not sufficient that the gains factually outweighs the losses. To decide the case we need to know more about the number of the affected, and the distribution of happiness and unhappiness.

4.3. The Energy Consumption of Present People

Here we described a situation where our high energy consumption would cause sufferings for future people. So, in this example the happiness of present people stands against the sufferings of future people. The main question was whether we, the present existing people, are forbidden to continue to consume energy at the cost of sufferings of future people. In connection to this we asked whether the happiness and the sufferings in the remote future is worth less than the happiness and sufferings now and in the near future.

In our opinion, the mere timing of happiness and sufferings have no moral importance. We are not justified in using a temporal discount rate and discount the more remote welfare effects of our policies at some rate *n* per cent per year. For instance, nuclear waste may be dangerous for thousands of years, and global warming can radically change the conditions for life on earth. That these things will mainly harm people living in the remote future does not make them less bad. This is not to say that there are some morally relevant facts that correlates more or less with remoteness in time. For instance, one could argue that future welfare is hard to predict. But this does not mean that the future welfare is less worth; only that the further in the future some welfare effect occur, the harder is it to predict it.

We think that using a genuine temporal discount rate is as unreasonable as using a spatial discount rate, where one discount welfare effects at some *n* per cent per meter. Consequently, our theory does not discriminate between present and future people. This also means that *intergenerational* trade-offs are treated in exactly the same manner as interpersonal trade-offs between presently existing persons. So, the comments given to examples 1 and 2 are just as well applicable to the intergenerational case.

4.4. Different Energy Systems - Different People

Here we depicted a scenario where future people's identities were dependent on our choice of energy policy. We had a choice between two energy policies. People in the further future will be much better-off if we choose policy A rather than policy B. Policy B would, however, make the present people a little bit better-off. Both policies also affect the identities of the people in the further future. In fact, both energy policies are of such a character that there will completely different people alive if we choose policy B rather than policy A and vice versa.

Let us say that we choose policy B. The people whose lives will be of relatively poor quality cannot complain that they have been harmed, or that the choice was against their interest, or that they are worse off than they might have been had policy A been chosen, for if policy A had been adopted they would not have existed at all.

We rejected this claim and all principles that make use of a Person Affecting Restriction. We showed in Chapter 4, section 5, that all such theories had several counterintuitive consequences. The specific identities of people have no moral importance. In our theory, cases like this are on par with example 1, 2 and 3 above. All that matters is how people's welfare is affected by an energy policy. If the present people would get less welfare from policy B than the future people would get from policy A, then we ought to choose policy A.

4.5. The Overcrowded Earth

The main point of this example is that the current development on earth may very well lead to a future earth densely populated by people with very low but positive quality of life. We asked whether it would have been better, had we chosen another path leading to a future with a smaller population but with much higher welfare. This can sound like a rhetorical question, but the total happiness of the densely populated world would exceed the total happiness of the less populated world. Could this not be a reason to claim, along the lines of Total Utilitarianism, that it would be better to populate earth as much as possible although the individual quantity of welfare would be low?

We rejected this claim. Our theory implies that when a population is big, an increase in quality of welfare is more important than an increase in quantity of welfare. Assume that we could choose between doubling the population on earth or doubling the welfare of the existing people. When we have a big population, as we have today, our theory singles out the latter alternative. This does not mean that any small increase in the existing population welfare can outweigh population growth. On the contrary, our theory gives value both to quality of welfare and quantity of welfare but less value to quantity of welfare the bigger the present population is.

Our theory captures an intuitive asymmetry between the obligation to make people happy and make happy people. Let us say that we have the choice between bringing a happy person into existence or raising the welfare of an existing person with low welfare. If the welfare in both situations is of the same magnitude, our theory prescribes that we should help the existing person with low welfare rather than procreate.

Bibliography

Acton, H B (ed.), J S Mill: Utilitarianism, On Liberty and Considerations on Representative Government, Dent: London and Melbourne, Everymans's Library, 1972, reprinted 1987.

Alchourròn, B. E., Normative Systems, Wien, New York , Springer-Verlag, 1971.

Anglin, B, "The Repugnant Conclusion", Canadian J Phil, Vol 7, #4, Dec 1977.

Aristotle, The Republic, ed. Bloom (1968).

------------, Nichomachean Ethics.

Arrhenius, G., "Framtida generationer och de vedervärdiga slutsatserna, C-uppsats, Filosofiska Institutionen, Uppsala Universitet, 1992.

Barry, B, "Rawls on Average and Total Utility: A Comment", Philosophical Studies, #31, 317-25, 1977.

Bayles, M D, Ethics and Population, Cambridge, Mass, Schenkman, 1976.

Bergström, L., Frågor om livets mening, Filosofiska studier 36, 1984.

Bickham, S, "Future Generations and Contemporary Ethical Theory", J Value Inquiry, #15, pp. 169-77, 1981.

Bloom, A, The Republic of Plato, Basic Books Inc, New York, 1968.

Brentano, F, "Loving and Hating", 1907, Chisholm, ed. (1969).

Cameron, J R, "Do Future Generations Matter?", Ethics and the Environmental Responsibility, Aldershot, Avebury, 1989.

Chisholm, R M, The Origin of Our Knowledge of Right and Wrong, New York, Routledge and Kegan Paul, 1969.

Coombs, C. H., Dawes, R. M., Tversky, A., Mathematical Psychology, New Jersey, 1970.

Danielsson, S., Filosofiska invändningar, Sju kritiska uppsatser, Thales, 1986.

Dasgupta, P, "Lives and Well-Being", Social Choice and Welfare, 1988.

-----------------, An Inquiry into Well-Being and Destitution, Oxford University Press, 1993.

-----------------, "Savings and Fertility: Ethical Issues", Philosophy and Public Affairs, 1994.

Davies, K, "The Conception of Possible People", Cogito, #2, 53-60, Spr 1984.

Dostojevsky, F., The Brothers Karamazov, bk V, 1923

Elliot, R, "Future Generations: Lockes Proviso and Libertarian Justice", J Applied Phil, #3, 217-27, Oct 1986.

Fehige, C, "A Pareto Principle for Possible People", unpublished paper.

Feldman, F, Introductory Ethics, Englewood Cliffs, NJ,Prentice-Hall, 1978.

Glover, J, Causing Death and Saving Lives, Penguin Books, 1977.

Griffin, J., "Is Unhappiness Morally More Important than Happiness?", Phil. Quarterly, vol. 27, 1979.

Haksar, V, Individual Selves and Moral Practice, Edinburgh University Press, 1991

Hamilton E and Cairns H (ed.), Plato: The Collected Dialogues, Bollingen Series LXXI, Princeton University Press, 1961, twelfth printing July 1985.

Hanser, M, "Harming Future People", Phil Pub Affairs, #19(1), 47-70, Wint 90.

Hare, R M, "Ethical Theory and Utilitarianism", in H. D. Lewis, ed., Contemporary Brittish Philosophy, George Allen and Unwin, London, 1976, reprinted in Sen and Williams (1982).

Hare, R M, Moral Thinking, Clarendon Press, Oxford, 1981.

Heyd, D, "Procreation and Value: Can Ethics Deal With Futurity Problems?", #18, 151-170, Jul 1988.

Hudson, J L, "The Diminishing Marginal Value of Happy People",?, 1986

Hurka, T, "Value and Population Size", Ethics, Apr. 1983.

Jecker, N S, "The Extended Nonidentity Problem", Auslegung, #14, 185-96, Sum 1988.

Kavka, G, "The Paradox of Future Individuals", Phil & Public Affairs, #11(2), Spr 1982.

-------------- "The Futurity Problem, Obligations to Future Generations", ed. Sikora och Barry.

-------------- "Rawls on Average and Total Utility", Philosophical Studies, #27, 237-53, 1975.

Krantz, D.H., Luce, R.D., Suppes, P., and Tversky, A., foundations of Measurement, vol. 1, New York, 1971.

Lemos, N. M., "Higher Goods and The Myth of Tithonus", The Journal of Philosophy, Sep 1993.

Locke, D, "The Parfit Population Problem", Philosophy, #62, pp. 131-157, Apr 1987.

MacMahan, J, "Problems of Population Theory", Ethics, #92, Oct 1981.

Marglin, S, "The Social Rate of Discount and the Optimal Rate of Investment", Quarterely Journal of Economics, 1963.

Mattlage, A, "Response to Jecker's the Extended Nonidentity Problem", Auslegung, #14, 197-200, Sum 1988.

Mendola, J., "An Ordinal Modification of Classical Utilitarianism", Erkenntnis 33, 1990.

Mill, J S, Utilitarianism, 1863, in Acton, ed., (1987).

----------, "Whewell on Moral Philosophy", Westminster Review, 1852, Dissertations and Discussions, vol. II, in Priestley and Robson (1969-87), vol. X.

Narveson, J, "Utilitarianism and New Generations", Mind, #76, 1967.

-------------, "Future People and Us, Obligations to Future Generations", 1978, in Sikora och Barry (ed.).

-------------, "Moral Problems of Population", The Monist, #57, Jan 1973.

Ng, Y-K, What Should We Do About Future Generations? Impossibility of Parfit's Theory X., Econ Phil, #5(2), 235-253, Oct 1989.

Noonan, H., Personal Identity, London, Routledge, 1989

Ohlsson, R., The Moral Import of Evil, Filosofiska studier 1, Stockholm, 1979.

Parfit, D, Reason and Persons, Clarendon Press, Oxford, 1984.

-----------, "Overpopulation and the Quality of Life", in Singer (1986).

-----------, "Future Generations: Further Problems", Phil & Public Affairs, #11(2), Spr. 1982.

-----------, "On Doing the Best for Our Children", in Bayles (1976.).

-----------, "A Reply to Sterba", Phil & Public Affairs, #16, Spr 1987.

Partridge, E (ed.), Responsibilities to Future Generations, Prometheus Books, New York, 1981.

Popper, K., The Open Society and its Enemies, vol 1 and 2, Routledge, London, 1962.

Priestley, F E L and Robson, J M, The Collected Works of John Stuart Mill vol. I-XXI, University of Toronto Press and Routledge and Kegan Paul, Toronto and London, 1969-87.

Quinn, W, "The Puzzle of the Self-Torturer", Philosophical Studies #59, pp. 79-90, 1990.

Rawls, J, A Theory of Justice, Harvard University Press, Cambridge, Massachusetts, 1971

Raz, J, "Value Incommensurability: Some Preliminaries"

Rescher, N., Distributive Justice, Indianapolis/New York/Kansas City,1966

Roberts, F.S, Measurement Theory with Applications to Decisionmaking, Utility, and The Social Sciences, Massachusetts, 1979.

Ross, W D, Foundations of Ethics, Oxford, 1939.

Ross, W D, The Right and the Good, New York, Oxford, 1930.

Samuelson, P A, Economics, New York, McGraw-Hill, 1970.

Sartorius, Rolf E, Individual Conduct and Social Norms, Dickenson Publishing Company, 1975.

Scheffler, S., Consequentialism and its Critics, Oxford University Press, 1988.

Schopenhauer, A., ? vol 2 1969.

Sen and Williams: A K Sen and B Williams, Utilitarianism and Beyond, Cambridge University Press, 1982.

Sen, A, Collective Choice and Social Welfare, Mathematical Economics Texts 5, 1970.

Sher, G (ed.), J S Mill: Utilitarianism, Indianapolis, Hackett, 1979.

Sider, T S, "Might Theory X Be a Theory of Diminishing Value?", Analysis, 51 (4), 1991

Sidgwick, H, The Methods of Ethics, London, Macmillan, 1907.

Sikora, R I and Barry, B M (ed.), Obligations to Future Generations, Temple University Press, Philadelphia, 1978.

Sikora, R I, "Utilitarianism: The Classical Principle and the Average Principle", Canadian J Phil, Vol 5, #3, Nov 1975.

Singer, P, "Anglin on the Obligation to Create Extra People", Canadian J Phil, Vol 8, #3, Sep 1978.

------------ (ed.), Applied Ethics, New York, Oxford, 1986.

Smart, J .C.C, Williams B., Utilitarianism, For and Against, London, Cambridge University Press, 1973

Smart, R., N., "Negative Utilitarianism, Mind, vol 67, 1958.

Sterba, J, "Explaining Asymmetry: A Problem for Parfit", Phil & Public Affairs, #16, pp. 188-192, Spr 1987.

Swedish Council for Planning and Coordination of Research, Surprising Futures: Notes from an International Workshop on Longterm World Development, report 87:1, 1987.

Temkin, L S, "Intransitivity and the Mere Addition Paradox", Phil & Public Affairs, 1986, pp. 138-187.

Tännsjö, T, Göra Barn, Sesam förlag, Borås, 1991.

Whewell, W, Lectures on Moral Philosophy, 1st edition, 1852.

Woodward, J, "The Non-Identity Problem", Ethics, Jul 1986.

Österberg, J., unpublished paper, 1992.