FUTURE GENERATIONS

A CHALLENGE FOR MORAL THEORY

Gustaf Arrhenius

Dissertation for the Degree of Doctor of Philosophy in Practical Philosophy presented at Uppsala University in 2000.

ABSTRACT

Arrhenius, G. 2000: *Future Generations: A Challenge for Moral Theory*. viii+234#? pp. Uppsala.

For the last thirty years or so, there has been a search underway for a theory that can accommodate our intuitions in regard to moral duties to future generations. The object of this search has proved surprisingly elusive. The classical moral theories in the literature all have perplexing implications in this area. Classical Utilitarianism, for instance, implies that it could be better to expand a population even if everyone in the resulting population would be much worse off than in the original.

The main problem has been to find an adequate *population theory*, that is, a theory about the moral value of states of affairs where the number of people, the quality of their lives, and their identities may vary. Since, arguably, any reasonable moral theory has to take these aspects of possible states of affairs into account when determining the normative status of actions, the study of population theory is of general import for moral theory.

A number of theories have been proposed in the literature that purport to avoid counter-intuitive implications such as the one mentioned above. The suggestions are diverse: introducing novel ways of aggregating welfare into a measure of value, revising the notion of a life worth living, questioning the way we can compare and measure welfare, counting people's welfare differently depending on the temporal location or the modal features of their lives, and challenging the logic of axiological and normative concepts. We investigate the concepts and assumptions involved in these theories as well as their implications for population theory.

In our discussion, we propose a number of intuitively appealing and logically weak adequacy conditions for an acceptable population theory. Finally, we consider whether it is possible to find a theory that satisfies all of these conditions. We prove that no such theory exists.

Keywords: future generations, population theory, population ethics, axiology, welfare, moral theory, social choice theory, impossibility theorems.

Gustaf Arrhenius, Department of Philosophy, Uppsala University, Drottningatan 4, S-753 10 Uppsala, Sweden.

Printed in Sweden by *Reprocentralen Ekonomikum*, Uppsala, 2000.

For Adeze

ACKNOWLEDGEMENTS

I have received a great deal of help from a variety of individuals during the years of work that culminated in this thesis. I've been fortunate to have been surrounded by a group of helpful and gifted philosophers, and they have taught me the immense value of co-operation in research. To begin with, my supervisor Sven Danielsson has been very supportive – I don't think anyone can ask for more from a supervisor. Although his detailed and trenchant criticism sometimes made me despair over the complexity of the subject and my inability to get things right, without his firm advice on how to write, or how not to write, this thesis would have been in much worse condition. I would especially like to mention his invaluable advice on how to conceptualise and structure the problems discussed in this work.

My advisor, Wlodek Rabinowicz, has a special place in my career as a philosopher. Many years ago, quite bored with my studies in medicine, I strolled by the Department of Philosophy in Uppsala and decided to sign up for some courses. Wlodek was my first philosophy teacher, and his enthusiasm and inspiring lectures enticed me to leave medicine for philosophy. Later on, he was the supervisor of my Master's thesis, and his advice and guidance during those years have made an everlasting impression on me. In a remarkably short period of time, Wlodek read the last draft of this essay, and his comments spared me many unnecessary errors.

I owe my greatest intellectual debt to my friends and colleagues Krister Bykvist and Erik Carlson. I've been in continuous dialogue with them during the last seven years or so about philosophical questions in general, and population theory in particular. They never got tired of my confused questions, crazy proofs, and the countless of drafts of chapters with which I burdened them. They have been a constant source of intellectual challenge and philosophical comraderie.

The essential part of this thesis consists of a number of theorems. I wouldn't have succeeded in producing these without the advice of Jan Odelstad, Rysiek Sliwinski, Howard Sobel, and especially Kaj Børge Hansen. Kaj Børge took upon himself the Herculean task of checking all the theorems in this essay, and he suggested many substantial improvements. I would also like to thank Rysiek for all his help in practical matters, and for his contagious good spirit.

An earlier version of a part of this thesis was successfully defended at the Department of Philosophy, University of Toronto. My thesis committee in Toronto, Wayne Sumner, Howard Sobel and Arthur Ripstein, were always willing

to discuss not only topics from my thesis but also philosophy in general, and I've learned a lot from those occasions. Both Wayne and Howard continued to help me with my work after my stay in Toronto and they have read and commented in detail on several earlier versions of the present work. From Wayne, I've learned most of what I know about theories of welfare and much more. Howard has continuously and assiduously sent me extremely detailed and helpful comments of my work, all of which, I'm sure, I have not been able to address. The suggestions from my external examiner, Thomas Hurka, were also valuable and insightful. During my years in Toronto, I've benefited much from discussions with Chandra Kumar and John Gibson. Chandra has influenced my philosophical thinking far more than he can imagine. Financial support through travel grants from the Swedish Institute, The Royal Swedish Academy of Sciences, The Swedish-American Foundation, Uppsala University, and Stockholm Nation during my stay in Toronto 1994-97 is gratefully acknowledged.

Over the years, I've discussed parts of this thesis with Thomas Anderberg, John Broome, Bruce Chapman, Danny Goldstick, Gert Helgesson, Bernard Katz, Karsten Klint Jensen, Andrew Latus, Per-Erik Malmnäs, Joshua Mozersky, Tomasz Pol, Jesper Ryberg, Peter Ryman, Pura Sanchez, Michelle Switzer, and Folke Tersman. With Jan Österberg, I've discussed most of the thesis and his work in population axiology has been a source of inspiration. I'm grateful to them all for the constructive criticism and the stimulating discussion they have offered.

Earlier versions of parts of this thesis were presented at Svenska Filosofidagarna, Göteborg, June, 1999; "Utilitarianism Reconsidered", ISUS, New Orleans, April 1997; "Utilitarianisme: Analyse et Histoire", Association Charles Gide pour l'Étude de la Pensée Économique and the University of Lille, January 1996; The Learned Societies Congress, Canadian Philosophical Association, UQAM, Montreal, June 1995; the University of Copenhagen, Dept. of Philosophy, October 1997, and at the University of Calgary, Dept. of Philosophy, January 2000. I would like to thank the participants at these occasions for their stimulating criticism. The long written comments on one of my papers from Charles Blackorby, Walter Bossert, David Donaldson and from Derek Parfit were also very helpful.

Without the intellectual and emotional support of Adeze Igboemeka, I doubt that this thesis would ever have been written. Without her linguistic advice, it would have been unreadable. I dedicate this essay to her.

Uppsala, April 17, 2000.

Gustaf Arrhenius

TABLE OF CONTENTS

1

Introduction

Many would agree that the present generation is profiting from the earth's resources at the expense of future generations. In combination with a steadily increasing population, this could result in a future world crowded with people whose lives are all of poor quality. Assume that we have an opportunity to avoid this overpopulation and to create a world with a sizeable but smaller population in which every person enjoys very high welfare. Which future is the better one? I think most of us find it evident that the latter future is superior to the former.

Let's call these two futures respectively the A- and the B-future. Suppose that we, sensibly, opted for the B-future. Things didn't turn out exactly as we had planned, however. We did succeed in securing very high welfare for as many people as we planned but we were not able to slow down population growth. As a consequence, the population expanded by a vast number of lives that were of poor quality but still worth living. Was this result a failure? How does this future, the C-future, compare to the B-future?

The number of well off people in the B- and C-future are roughly the same. The difference between these two populations is that in the C-future, there is a number of "extra" people whose lives are of poor quality but worth living. Could the existence of these extra lives which are worth living make the C-future worse than the B-future? Would it have been better if these "extra" people with lives worth living had never existed? It might strike one as implausible to claim that the C-future is worse than the B-future, merely because there are additional people with lives worth living.

How does the C-future compare to the A-future? Since we didn't succeed in slowing down population growth, the size of these populations are approximately the same. In the C-future, there is a number of people with very high welfare. However, on inspection, it turns out that the people with poor quality of life in the C-future are worse off than the people with poor quality of life would have been in

the A-future. As a matter of fact, since there are so many people with very low welfare in these futures, the total gain for the worst off would have been much greater than the total loss for the best off, if the A-future had come about instead of the C-future. Moreover, in the A-future there is equality – everybody has more or less the same quality of life – whereas in the C-future, the majority is much worse off than the minority. It turns out that there is an equal distribution of welfare in the A-future, and a greater average and total welfare as compared to the C-future. In other words, to claim that the A-future is worse than the C-future seems to be committing oneself to an untenable anti-egalitarianism.

By now, we have contradicted ourselves. We have claimed that the B-future is better than the A-future, the C-future is not worse than the B-future, and the A-future is not worse than the C-future. If the A-future is not worse than the C-future, and the C-future is not worse than the B-future, then the A-future is not worse than the B-future. But we have claimed that the B-future is better than the A-future, that is, that the A-future is worse than the B-future.

The above paradox is a simplified version of the "Mere Addition Paradox" introduced in Derek Parfit's seminal contribution to population ethics.[1] What is the significance of a paradox like this? At one end of the spectrum, we find people who think that it is little more than an intellectual puzzle with, at most, some entertainment value. At the other end, we find people who think that it is a disturbing practical problem that we actually face today.

It might be that paradoxes such as the one above represent some practical problem in the world today or some problem that we are likely to face in the future. If this is the case, then the problems discussed in this essay are certainly of considerable significance. In the introduction to works in theoretical moral philosophy, it is not unusual to see claims to the effect that the problems discussed and the answers delivered are of great importance to the solution of practical, real-life questions. Unfortunately, these claims often don't amount to more than hand-

---

[1] See Parfit (1984), pp. 419ff. For an informal proof of a similar result with stronger assumptions than Parfit's, see Ng (1989), p. 240. A formal proof with slightly stronger assumptions than Ng's can be found in Blackorby and Donaldson (1991). It should be stressed that the above paradox is not identical to Parfit's Mere Addition Paradox. As we shall see later on, Parfit denies that the C-future is not worse than the B-future, and perhaps also that the A-future is not worse than the C-future.

waving without any substantiating arguments. It is clear that a lot of work has to be done, involving numerous arguments and results from economics, political science, anthropology, history, environmental sciences, and so forth, to show that the above paradox represents a practical problem. Such an extensive investigation could show that we actually face, in some sense, cases which are at least similar enough to the above paradox to give us reason to take it seriously as a practical problem. No arguments to this effect will be presented in this essay, however, since I think that irrespective of how such a project would turn out, there is another aspect of the above paradox and its cognates which make them into important philosophical problems.

In discussions of moral questions, we cannot avoid appeals to intuitions: "From your position it follows that this-and-this is good, but that is counter-intuitive, so you're wrong." Such appeals are as common in everyday discussions as in lofty philosophical debates. We test a moral view by checking whether it complies with our considered beliefs regarding different cases. Most often, these are not actual but hypothetical cases which we can easily grasp and have firm intuitions about. Of course, such intuitions may be mere prejudices – we are all strongly affected by the particulars of the cultural environment in which we grew up and live. There will always be moral beliefs that we will find grounds for abandoning when we scrutinise them in light of facts and other moral beliefs. Still, there is a number of such intuitions which are widely shared and which we tend to hold on to even after critically reflecting upon them: One shouldn't inflict pain on people unnecessarily; it is better that people are better off rather than worse off, and so forth. Such intuitions are often referred to as "firm, considered convictions", "solid beliefs", "moral convictions we share and have confidence in", and so forth. If a moral theory (view, outlook) is inconsistent with such beliefs, then that constitutes a reason to reject it.

Appeal to considered intuitions is part of the core of the methodology of the dominant tradition in modern moral philosophy. A necessary but presumably not sufficient condition for a moral theory to be justified is that, apart from being internally consistent, it should be consistent with considered moral intuitions. One tries to find, as it is often put, a "reflective equilibrium" among more or less general principles and beliefs about more or less particular cases.

The examples of considered moral beliefs I gave above have the character of truisms. A common worry about testing moral theories against such beliefs is that

this method isn't powerful enough. The moral intuitions that are firm enough to stand up to critical scrutiny will only weed out the wildest of moral theories.[2] The paradox above suggests something else, however, something much more troubling. If the evaluations above stand up to scrutiny, that is, if we find it impossible to give up any one of them, then our considered moral beliefs are mutually inconsistent. And of course, the same would hold for any moral theory which implies these beliefs. Since consistency is a necessary condition for moral justification, we seem to be forced to conclude that there is no moral view which can be justified. In other words, paradoxes of the above kind might challenge some of our deepest beliefs about moral justification.

Since inconsistency is a hard bullet to bite, the sensible reaction to a paradox is to reconsider the involved beliefs. That the above evaluations are inconsistent is a *prima facie* reason to give up at least one of them. In this sense, paradoxes, or apparent paradoxes, can be useful in that they give structure to our thinking. If we can specify a set of conditions in an exact manner and prove that they are mutually inconsistent, then we know that we have to jettison at least one of the involved conditions to retain consistency.

As the above paradox is presented, however, it is hopelessly vague and, at most, of rhetorical value. It doesn't force us to any conclusion but is rather an invitation to philosophical analysis. To understand what it involves, we have to clarify concepts such as "welfare", "population", and so forth. This is the purpose of the next chapter. And of course, we have to critically investigate the moral evaluations and presuppositions which constitute the above paradox. As we shall see in the following, all three of these evaluations have been challenged in the literature and at least two of them have been criticised on pretty good grounds. Likewise, some of the presuppositions might be questioned, as some authors have suggested. In chapters 3 - 9 we shall discuss these different arguments and the proposed population theories in the literature. In our discussion, we shall propose a number of intuitively compelling and logically weak adequacy conditions for an acceptable population theory. In chapters 10-11, we shall consider whether it is possible to find a population axiology that satisfies all of these conditions, or a population

---

[2] For this worry, see, for example, Griffin (1986), p. 2.

morality that satisfies the corresponding normative conditions. We shall prove that no such theory exists.

2

Basic Concepts and Presuppositions

2.1 Welfare and the Value of Life

As should be evident from the introduction, "welfare" is a term that will be used often in this essay. This concept has, not surprisingly, acquired a number of different meanings. On the one hand, we need to narrow down the possible meanings of this expression so that we know what the examples and principles that we shall discuss involve. On the other hand, we want to avoid taking a stand on controversial issues about welfare which don't affect the nature of the problems that we are going to discuss – we don't want to narrow the scope of our discussion unnecessarily.

Roughly, a person's welfare has to do with how well her life is going. Welfare concerns how good or bad a life is for the person living it, how good or bad her life is *for* her.[1] This is still pretty vague but we can narrow it down by stating some dimensions of the value of a life which welfare doesn't capture.

Welfare should be distinguished from the aesthetic value of a life such as how beautiful or dramatic a life is, that is, the aspects of life that capture our imagination in novels and plays. Of course, such aspects of a life can affect how good that life is for the person living it, but that's all right. Obviously, someone can lead a very beautiful and dramatic life but still – to allude to the etymology of "welfare" – fare pretty badly.

We shall also distinguish welfare from the ethical value of life understood as how well or poorly a life squares in regard to some moral standard of how we ought to live our lives. On reasonable accounts of how we ought to live our lives,

[1] I am essentially following Sumner's (1996), p. 20ff., explication of welfare. Notice that by claiming that welfare has this perspectival character, we are not trying to exclude theories which make welfare logically independent of people's attitudes. We shall leave the field open for other ways of explaining this property of welfare such as, for example, Moore's (1903), p. 98, "private ownership theory". Cf. Sumner (1996), p. 47 and fn. 14 below.

there is no necessary tie between a moral life and a happy life. Granted, there is a time-honoured tradition in ethics of trying to show that morality and self-interest coincide. But coincidence is not the same as conceptual identity.[2] It is clearly conceivable (not to say likely, given the present state of the world) that a "morally upstanding citizen" should lead a life full of dissatisfactions and disappointments.

We shall also distinguish welfare from the *contributive value of a life* – this distinction plays a crucial role throughout this essay and in the theories we are going to discuss. By the contributive value of a life we shall mean the value that a life confers to a population of which it is a member. More exactly, the contributive value of a life x relative to a population A, of which x is a member, is the difference in value between A and the population consisting of all the A-lives except life x. An example might help here. Let's say that the only difference between populations A and A' is that Scott exists in A but not in A'. If population A is better (worse, equally as good) than (as) A', then the contributive value of Scott's life relative to population A is positive (negative, neutral).[3]

The contributive value of a life is a central matter at stake in population axiology. Some theorists, most notably classical utilitarians, hold that the contributive value of a life equals its welfare. Others deny this but still hold that the contributive value of a life is determined by its welfare and is independent of other people's welfare. An example would be theorists who stipulate a positive level of welfare which a life has to attain to have positive contributive value. Another group of theorists believes that the contributive value of a life is context-sensitive, that is, dependent on the welfare of other people.[4] An example of a theorist in the last group is the average utilitarian, that is, a utilitarian who ranks populations according to their average utility. According to her, a life has positive (negative, neutral) contributive value relative to a population A exactly if it has higher (lower, the same) welfare than (as) the average welfare of the rest of the A-population. One

---

[2] Sumner (1996), p. 24, makes this point.

[3] If the value of populations can be measured on a scale that allows us to talk meaningfully about numerical difference in value (we shall shortly talk about this in more detail), then the contributive value of Scott's life can be represented by the numerical difference between the value of population A and A'.

[4] This is analogous to reasonable conceptions of the ethical value of a life which makes this value partly dependent on other people's welfare, for example, on how much an individual has contributed to the welfare of others. Cf. Sumner (1996), p. 24.

might also hold that the contributive value of a life depends on other factors apart from welfare. An example would be those theorist who believe that desert is an important consideration. They consider it bad that people have lower welfare than they deserve, and that such lives might have negative contributive value although they are good for the people living them.

It would be all too hasty to dismiss the three latter groups of theorists as conceptually confused. On the contrary, to identify the contributive value of a life with its welfare on conceptual grounds is to conflate a theory about what makes people's lives good with a theory about what makes populations good.[5] How people's welfare correlates with their contributive value is a substantial axiological question which has to be settled by investigating our considered beliefs over different cases, which is, indeed, exactly what we are setting out to do in this essay.

After having stated what welfare conceptually is not, let's turn to some substantive ideas about what welfare is. There are, not surprisingly, a number of different views on what makes a life better or worse. In the last twenty years or so, great advances have been made in this field and the number of theories has multiplied. This is not the place to give a complete survey of all those theories and such a list would anyway be pretty tiresome. Let's therefore just bring up some paradigmatic types or components of such theories.

Experientialist theories make a person's welfare solely a matter of her mental experiences. Classical hedonism, in which welfare is a function of experiences of pleasure and pain, is the standard textbook example. Of course, pleasurable and painful experiences should not be understood as restricted to only bodily pleasures and pains. Hedonistic welfare also includes complex intellectual pleasures and pains such as the pleasure of solving a chess problem and the grief of a loved one's death. Not so much in vogue these days, hedonism had its heyday in the eighteenth and nineteenth century with such famous proponents as Jeremy Bentham, John Stuart Mill and Henry Sidgwick. This is not to say that there are no contemporary defenders of hedonism – in the last few years a number of such attempts have been made, most notably by Fred Feldman and Torbjörn Tännsjö.[6]

---

5 See Temkin (1993a), pp. 258ff, (1993b), and (1994), pp. 354-5, for the same point.

6 Sumner (1996) gives a good account of classical hedonism in ch. 4 (in ch. 6, Sumner develops a sophisticated theory of welfare that retains some important traits of hedonism). Feldman (1997), essays 5 - 6, discusses a number of different formulations of hedonism and develops his own

The dominant approach today is perhaps that of the *desire theorist*. A person's welfare is a question of her desires, wants or preferences being fulfilled. One version of this account of welfare has been especially popular among economists: Revealed Preference Theory. An individual's preferences are simply identified with the "preferences" that she "reveals" in her choices, and her welfare increases exactly if these preferences are satisfied. As many authors have showed, this is a pretty dubious theory of welfare, and it has yielded its place to attitudinal explications of desires.[7] This still leaves room for a pretty wide spectrum of theories, since there is no consensus on which kind(s) of desire count(s). Some theorists only count the desires that a person actually has, others count only informed desires or the desires that a person would have if she were well-informed. Some of them make a distinction between rational and irrational desire, others among desires for the past, present and the future or among desires that are located in the past, present or the future. One might stress autonomously formed desires, another prioritises so-called global desires, that is, desires about the character of one's whole life, and so forth. At any rate, the desire theory has become very influential, and it has been promulgated in ethics by such prominent philosophers as Brian Barry, James Griffin, Richard Hare, John Harsanyi, Peter Singer, Joseph Raz and Robert Goodin.[8]

Objective List Theories are so called since they are made up by a list of things that, purportedly, are good or bad for a person irrespective of her subjective attitudes towards these things.[9] A person's welfare is determined by her possession of these things. The good things might, for example, include the development of one's abilities, knowledge, appreciation of true beauty, friendship, good health, nourishment, personal security, freedom, dignity, and so forth; and the bad things

---

account in essay 7. Feldman gives a propositional analysis of hedonism which moves his account closer to the position of desire theorists. This move is criticised in Sumner (1998). For an unabashed defence of classical hedonism, see Tännsjö (1998).

[7] For arguments against a revealed preference account of welfare, see Mongin (1997) and Sumner (1996), ch. 5.

[8] See Barry (1989), Griffin (1986), Hare (1981), Harsanyi (1992), Singer (1993), Raz (1986), Goodin (1991). For a detailed analysis of the very concept of a desire, see Bykvist (1998). Rabinowicz and Österberg (1996) make an interesting distinction between "object" and "satisfaction" versions of the desire theory where the former explication moves the position of the desire theorist closer to the objective list account of welfare.

[9] This label is from Parfit (1984), pp. 4 and 499.

might be losing liberty or dignity, bad health, malnutrition, sadistic pleasure, being deceived, appreciation of kitsch, and the like. Some objective list theorists have remained content with just stating that their list is "self-evident" whereas some have tried to give an explanation of the specific items on the list of goods.[10] A typical kind of objective list theory consists of theories centred around "basic needs". Here, the list of goods – nourishment, exercise, rest, companionship, and so forth – allegedly springs out of some aspect(s) of human nature and predicament, or from some kind of consensus among some selected group of people.[11] An approach similar to the latter is used by John Rawls to derive his influential list of primary goods: rights and liberties, opportunities and powers, income and wealth, and a sense of one's own worth.[12] Another example of an influential objective theory of welfare is Amartya Sen's theory of functionings and capabilities. A functioning is, roughly, anything that a person succeeds in doing or being, for example, working as a brick-layer and being well-nourished; a capability is an opportunity to achieve a particular functioning, for example, the opportunity to work as a brick-layer if one so chooses. A person's welfare consists in her collection of functionings and capabilities.[13] We shall also include perfectionist theories under the heading of objective list theories. According to these, the goodness of a life depends on how well it manifests the "essential" properties of human beings or, as

---

[10] Finnis (1980) is an example of the former kind of theorist. His list includes seven items: knowledge, preservation of life (health), play, aesthetic experience, sociability (friendship), practical reasonableness, religion. Each of these items are "equally self-evidently a form of good" (p. 92). As Sumner (1996), p. 45, points out, one can question whether a mere list of human goods is a theory of welfare rather than just an inventory of its sources.

[11] For the latter kind of theory, see Braybrooke (1987).

[12] Rawls (1971), pp. 62, 92. These "ideas of the good may be freely introduced … so long as they belong to a reasonable political conception of justice for a constitutional regime. This allows us to assume that they are shared by citizens and do not depend on any comprehensive doctrine." See Rawls (1988), p. 263. It should be stressed, however, that Rawls developed his list of primary goods as a part of his theory of justice and not as a theory of welfare. At other times (1971), p. 93, he claims that "… good is the satisfaction of rational desire". Cf. Sumner (1996), p. 57.

[13] See Sen (1980, 1993). As Sumner points out, a person's set of functionings and capabilities is very large and for Sen's theory to have any credibility, he needs to tell us how to sort out the trivial functionings and capabilities from the important ones. Since Sen's solution to this problem involves personal evaluation of the functionings and capabilities, and personal evaluations have, arguably, a subjective character, his theory might not after all be an objective list theory. See Sumner (1996), p. 62-66. Cf. fn. below.

it is often expressed, human nature. Proposed examples of such properties are rationality, knowledge, autonomy, love and friendship, and so forth.[14]

One might, of course, advance a theory of welfare that incorporates components from all or some of these three types and thus straddle the distinctions made above. Perhaps welfare consists in experiencing pleasure in objectively good things that one desires.[15] Consequently, it should be clear that we are not claiming that the three categories above are analytical or salient in any way. Such a classification would be useful if we had some axe to grind, for example, that all theories belonging to one of the categories shared the same flaw and therefore should be discarded. But we don't have any axe to grind in this case. The classification above should just be seen as descriptive of the current philosophical thinking about welfare. In general, the nature of the problems we shall discuss is not dependent on any specific theory of welfare. In some particular cases, when we discuss a population axiology proposed by a certain theorist, there might be some specific problems of that theory that are dependent on the favoured conception of welfare of that theorist. In such cases, we shall be careful to point this out. For the general results that we shall derive in this essay, however, theories of welfare belonging to any one of the above categories, or combinations thereof, will do.

One can meaningfully talk both about the welfare of a whole life and of parts of a life. Perhaps we are more used to speaking about the latter ("My teenage years

---

[14] Our classification of perfectionism as an objective list theory of welfare might be controversial. Sumner (1996), pp. 23-24, claims that perfectionism, in contrast to other objective list theories, is not even in the running as a theory of welfare since "[w]hatever we are to count as excellences for creatures of our nature, they will raise the perfectionist value of our lives regardless of the extent of their payoff for us". I find this dismissal too hasty – it all depends on which aspects of human nature a perfectionist theory picks out for us to develop to as high a degree as possible. For example, it could conceivably be rationality understood as maximisation of preference satisfaction. The leading contemporary proponent of perfectionism, Thomas Hurka holds that "perfectionism cannot concern well-being …[and] cannot define the "good for" a human because the ideal is one he ought to pursue regardless of his desires" (1993, p. 17). This would, of course, not only exclude all objective list theories from the race but also mental state theories. This is also too quick a conclusion – it is not self-evident that welfare is tied to desire satisfaction. Sumner (1996), however, provides a convincing argument that all pure objective theories, that is, theories which make welfare logically independent of people's attitude (understood in a wide sense and not restricted to propositional attitudes) cannot make sense of the subject-relative character of welfare.

[15] Parfit (1984), pp. 501-2 suggests this. Sumner's (1996), ch. 6, theory of welfare includes traits from both experientialist and desire theories.

were horrible", "I really enjoyed the seventies", and so forth), but both philosophers and social indicator researchers have come to stress evaluations of whole lives.[16] We shall join this trend. By talking about the welfare of whole lives, we avoid "bookkeeping" problems for welfarist theories which stress, for example, fulfilment of projects and life plans as an essential part of a good of life – such aspects of well-being might not be located in any specific part of a life (Does the goodness of a fulfilled life-plan only occur at the point of time when the plan is fulfilled?). Moreover, since life-expectancy plays a pretty important part in any reasonable theory of welfare, we would like to compare alternatives which involve whole lives of different length. It is not the case, however, that we think that this way of framing the discussion is crucial for the results that we are going to derive. It just gives our discussion a clearer form. With some care, our questions could also be formulated in terms of welfare during some shorter period of time.

2.2 Orderings of Lives

Utterances such as "She had a good life", "He is better off pursuing a career in French than in physics", "Unemployed people are worse off in Mexico than in USA", "What a terrible life", and the like belong to our ordinary language. These utterances involve comparisons of the welfare of lives, or, as we also can put it, orderings of lives in regard to their welfare. The kinds of orderings of lives that are possible is a somewhat controversial topic. We shall approach it from two angles.

We shall begin by stating some weak assumptions regarding the orderings of lives which are sufficient for the *adequacy conditions* that we will discuss for reasonable population axiologies. There are two reasons why it is important that these ordering assumptions be as weak as possible. Firstly, the logically weaker these assumptions are, the harder they are to reject, that is, the intuitively stronger they get. As we mentioned in the introduction, we are going to prove some impossibility theorems. The importance and credibility of such theorems are directly proportional to the logical weakness and the intuitive strength of the assumptions and conditions on which they are based. Secondly, by making our assumptions as weak as possible, we can weed out unnecessary shrubbery that stops us from seeing the core character of the problems that we are to discuss. This

---

[16] See Sumner (1996), ch. 6 and Griffin (1986), pp. 34ff.

is one of the main objectives of any philosophical inquiry and thus also of this essay.

We shall also state the assumptions of orderings of lives underlying the different population axiologies that we shall discuss and criticise. These assumptions are stronger than the ones needed for our adequacy conditions. To understand these axiologies properly, it is important to know what assumptions of orderings of life they involve. Likewise for the *principles* and *conclusions* that we are going to discuss. Some of these have been proposed in the literature as adequacy conditions and are compelling whereas others are deficient in some respect (there are good reasons for rejecting them or they can be replaced by weaker conditions that show the same point or they are too vaguely formulated). Most of them, however, are of our own making and are convincing, or so we shall argue, and we shall use them in our critique of the population axiologies proposed in the literature. The simple reason that they are called principles and conclusions rather than conditions is that we shall not use them in the impossibility theorems in chapters 10-11. Moreover, the properties of the ordering of lives that some of these principles and conclusions presuppose are more demanding than for the adequacy conditions. For example, some of the principles require that talk about total welfare is meaningful. This is acceptable, of course, as long as we only use them for criticising theories which presuppose orderings of lives that are at least as strong as these principles and conclusions.

For the convenience of the reader, we have listed in appendix A the conditions, principles and conclusions that we refer to in several chapters.

2.2.1 Comparative Ordering Presuppositions for the Adequacy Conditions

Since we are going to discuss how the ranking of populations depends on the welfare of their respective members, we need to make *comparative* welfare statements such as "p has higher (lower, the same) welfare than (as) q", or as we also could put it, "life p is better (worse, equally as good) for the person living it than (as) life q is for the person living it". We shall assume that at least some lives can be ordered by the comparative relation "_ has at least as high welfare as _" where each blank is to be filled in with the name of a life. We use this relation to define the two relations "_ has higher welfare than _" and "_ has the same welfare as _". x has higher welfare than y if and only if x has at least as high welfare as y and it is not the case that y has at least as high welfare as x. x has the same welfare as y if and only if x

has at least as high welfare as *y* and *y* has at least as high welfare as *x*. We shall assume that the relation "has at least as high welfare as" quasi-orders all possible lives with positive welfare, that is, it satisfies at least two standard properties of this type of relation: reflexivity and transitivity.[17] Thus, for all *x*, *x* has at least as high welfare as *x* (reflexivity), and for all *x*, *y*, and *z*, if *x* has at least as high welfare as *y*, and *y* has at least as high welfare as *z*, then *x* has at least as high welfare as *z* (transitivity).

Notice that we haven't assumed, and none of our definitions above imply, full comparability (or completeness as this property is also called): For any *x* and *y*, *x* has at least as high welfare as *y*, or *y* has at least as high welfare as *x*. In other words, we haven't ruled out that there might be lives which are incommensurable in regard to their welfare.[18] We shall leave the door open for the existence of incommensurable lives of both the intra- and interpersonal kind (lives of the same person can be incommensurable and lives of different people can be incommensurable).

We shall use the relation "has at least as high welfare" to define a *welfare level*. Roughly, a welfare level is a set of lives with the same welfare. More exactly, by a welfare level 𝔄 we shall mean a set such that if a life *a* is in 𝔄, then a life *b* is in 𝔄 if and only if *b* has the same welfare as *a*. In other words, a welfare level is an equivalence class on the set of all possible lives with respect to the relation "has at last as high welfare as". Let *a** be a life which is representative of the welfare level 𝔄. We shall say that a welfare level 𝔄 is higher (lower, the same) than (as) a level 𝔅 if and only if *a** has higher (lower, the same) welfare than (as) *b**; and that a life *b* has welfare below (above, at) 𝔄 if and only if *b* has higher (lower, the same) welfare than (as) *a**.

Lastly, notice that in our discussion above we have assumed that welfare is at least sometimes interpersonally comparable. Without this assumption, claims such as "John is better off than Chandra" wouldn't be meaningful, and, to put it bluntly, most of our talk in moral, political and economical questions would be nonsense. Without interpersonal comparability of welfare, one cannot say much in population

---

[17] We're using Sen's terminology for orderings. See Sen (1970), p. 9.

[18] See, for example, Griffin (1986), ch. 5, Raz (1986), ch. 13, and Broome (1999), ch. 9, for a discussion of incommensurability in welfare and some arguments for the existence of this phenomenon.

axiology, and, unsurprisingly, no one has proposed a population axiology without interpersonal comparability. At any rate, as "Arrowian" impossibility theorems show, *without* interpersonal comparability, one cannot even find a reasonable theory for ordering same-sized populations.[19] To escape Arrowian impossibility results, one has to reject the non-comparability assumption involved in these theorems and introduce some kind of interpersonal comparability. It is from this juncture that our discussion proceeds: Can one find a reasonable theory for ordering populations given that at least some interpersonal comparisons of welfare are possible?[20]

2.2.2 Categorical Ordering Presuppositions for the Adequacy Conditions

We also need to make *categorical* welfare statements, that is, statements of the general form "*p* has such-and-such welfare". We shall assume that there are possible lives with positive, neutral, or negative welfare, or, as we also could put it, lives that are good for, bad for, or neutral in value for the person living it (for variation, we shall also use the expressions "a life worth living"/"a life not worth living" as synonyms with the former expressions). We shall say that a welfare level **A** is positive (negative, neutral) if and only if *a** has positive (negative, neutral) welfare.[21]

The assumption that there are lives with positive or negative welfare is standard in the literature on population axiology and it is so commonsensical that it is hard to find any further arguments for it. Values in general have a positive/negative polarity – things are good or bad, beautiful or ugly, attractive or repugnant, agreeable or disagreeable, and so on – and welfare in particular displays this feature: lives or periods of lives can be happy or unhappy, wonderful or horrible, pleasant

---

[19] See Sen (1970), pp. 123-5, 128-30, and Roemer (1996), pp. 26-36 for Arrowian impossibility theorems with different measurement assumptions but no interpersonal comparability of welfare. Arrow's original theorem appears in Arrow (1963).

[20] Whether or not interpersonal comparisons of welfare are possible, and to what extent they are possible, might of course depend on the theory of welfare in question. Arguably, such comparisons seems to be less problematic on an objective list account than on a desire account. We shall not pursue this question further here, however.

[21] Notice that we are not assuming that the above partitions of possible lives into lives with positive, neutral, and negative welfare is exhaustive. There might be some peculiar lives that cannot be grouped into any of these sets.

or unpleasant, satisfying or dissatisfying, fulfilling or disappointing, tormenting or soothing, and so on.[22]

We are not claiming, of course, that it is apparent how this classification of lives looks in every detail. Where exactly the cut-off point between a life with positive and a life with negative welfare should be drawn is a difficult question, and different substantive theories of welfare will probably yield somewhat different answers. For example, whereas a hedonist might think that a life consisting of a few happy days is a life worth living, an objective list theorist might find such a life below the threshold of a life with positive welfare (of course, these two theorists would probably also disagree on the ordering of lives). Admittedly, the intuitive force of examples that we are to discuss is linked to our understanding of lives with positive and negative welfare. And if we were to radically revise these notions – for example by claiming (implausibly) that there are no lives worth living – then many of these examples would lose their force and many of the adequacy conditions that we are to propose would lose their relevance. As a matter of fact, a few of the solutions proposed for some of the problematic cases in population axiology seem to turn on some kind of radical revision of our understanding of a life worth living. We shall discuss these proposals in detail below. However, as long as a theory of welfare, as a reasonable theory should, roughly respects our common-sense intuitions about the value of life, I do not think that the solution to the problems discussed in this essay essentially turn on where we exactly draw the line between lives of positive and negative welfare and precisely how we spell out our theory of welfare.

In the literature on population axiology, there is an abundance of other categorical welfare concepts in use such as "terrible" or "dreadful" lives, lives "barely worth living", "very very happy" lives, and so forth. Likewise, for convenience we shall employ a few undefined categorical welfare concepts in the informal discussion of population axiologies, such as "very high positive welfare", "very low positive welfare", "slightly negative welfare", and the like. We take it that the intuitive meaning of these concepts are clear enough for the informal discussion. An important question is whether these concepts play an essential role in the discussion of population axiology. In chapter 10, we shall show that none of these concepts are necessary for the results that we shall derive. One property of

---

[22] For the same point, see Sumner (1996), pp. 35-6.

them is worth nothing here, however. Like positive and negative welfare, they involve several welfare levels – people with, say, very high welfare can be at different welfare levels. In other words, "very high positive welfare" is a set of welfare levels, or, as we shall put it, a *welfare range*.

2.2.3 The Relation between Comparative and Categorical Welfare Statements

It is fair to ask how the concepts "positive", "negative", and "neutral welfare" are related to the comparative concepts introduced above. For a starter, we can reduce the number of primitive concepts by defining lives with positive or negative welfare in terms of lives with neutral welfare and the relation "has at least as high welfare as":

> (*) A life has positive (negative) welfare if and only if it has higher (lower) welfare than a life with neutral welfare.

I presume that no one would reject the above definition but some people might think it is not enough. They would like to see all categorical value concepts reduced to some comparative concepts. For example, Edwin T. Mitchell claims that "value judgements … have the form 'A is better than B', or they can be reduced to this form".[23] Since judgements about welfare arguably are value judgements – they are judgements about what is good or bad for a person – then any categorical welfare statement has the same meaning as some comparative value statement.

It is important here to distinguish Mitchell's claim from the much stronger claim that categorical concepts are meaningless, or that we cannot clearly grasp

---

[23] Mitchell (1950), p. 114, as quoted in Hansson (1998). Broome (1999), ch. 10, p. 164, claims that "… there is nothing more to goodness than betterness". What Broome seems to have in mind is that all judgement about intrinsic value are reducible to comparative judgements. I'm unclear about his position regarding welfare judgements. He seems to think, however, that on a naturalist account of welfare, this is not true (Broome (1999), p. 170): "Let us distinguish a person's wellbeing from her good. Let us treat her wellbeing as a natural property; it is made up of the good and bad things in her life. Wellbeing in this sense has a natural zero given by the natural zeros of the good and bad things. - - - Wellbeing has a natural absolute zero; goodness does not. Goodness is still reducible to betterness." Below, we shall propose a reduction of categorical welfare statements to comparative welfare statements that is compatible with a naturalist account of welfare.

what they amount to, unless they are defined in terms of some comparative concepts. I don't see any foundation for this claim in regard to categorical welfare judgements. It is true that in many fields we can do without any categorical welfare concepts, such as in the theory of general equilibrium, but of course, this doesn't give us any reason to believe that we can do without them in population axiology, nor that we cannot understand what these concepts amount to unless they are defined in comparative terms. Even if one could construe "neutral welfare" in terms of, for example, "has at least as high welfare as", it doesn't follow that the latter concept is in some sense more primitive or fundamental than the former.[24] As a matter of fact, the linguistic evidence for some other categorical concepts seem to point in the opposite direction.[25] And of course, one can equally well turn the question around, take the categorical predicates as primitives, and ask whether one can reduce the comparative welfare concepts to categorical ones. So I don't think that an affirmative or negative answer to the question whether one can reduce the categorical welfare concepts to some comparative concepts would affect any of the arguments in this essay regarding the plausibility or implausibility of different population axiologies.

Having said this, one might wonder whether a discussion of different possible reductions of the categorical welfare concepts into comparative concepts would just be an unnecessary diversion from the main task of this essay. But I think it is worthwhile to investigate this matter a little bit further since these concepts play a crucial role in the discussion to follow. Moreover, such an investigation will clarify how we should understand, and, most importantly, not understand the categorical welfare concepts. For example, one might think that which lives that will count as lives with positive or negative welfare depends on which comparative concepts one uses to define the categorical welfare concepts. This would thus have implications for how we should understand the adequacy conditions that we shall discuss. As we shall argue, however, this is not true for the only two reductions that have some plausibility, since these will be formulated solely in terms of the comparative welfare statements introduced above. Of course, this discussion will also be of

---

[24] For the same point, see Hansson (1998), pp. 122-3.

[25] Hansson (1998), p. 123, fn. 1, reports that in a survey of 123 languages, the categorical concept "tall" is the basic form in all of these languages and the comparative form "taller is derived from the former concept. See also Roberts (1984), p. 66.

some philosophical interest in itself and it might show us how to reduce the number of primitive concepts in our discussion. We shall first look at some proposals which we shall reject, and then consider two promising candidates.

The first proposal we have already touched upon:

> (1) A life has neutral welfare if and only if it has neutral contributive value relative to all possible populations.

As we said above, the identification of the welfare of a life with its contributive value implies that a number of theorists are conceptually confused. An example would be the critical level theorists who stipulate a positive level of welfare that a life must reach to have positive contributive value and that lives with neutral welfare have negative contributive value. Given the above definition, this cannot make sense. And its combination with average utilitarianism implies that there are no lives with neutral welfare. According to average utilitarianism, any life can have positive (negative) contributive value relative to some populations, that is, relative to a population consisting of lives with lower (higher) welfare than the life in question. As I said above, I think it would be all too hasty to dismiss these and other theorists as conceptually confused. Again, to identify the welfare of a life with its contributive value on conceptual grounds is to conflate a theory about what makes people's lives good with a theory about what makes populations better or worse. I think we can invoke a version of Moore's "open question argument" here: How people's welfare correlates with their contributive value is an open axiological question which has to be settled by investigating our considered beliefs over different cases.

A restricted version of the above definition doesn't fare much better:

> (2) A life has neutral welfare if and only if it has neutral contributive value relative to an empty population.

This definition would save the average utilitarian from the problems above but still leave the critical-level theorists out in the cold. The same goes for those theorists who claim that an addition of people with positive welfare have neutral contributive value whereas an addition of people with negative welfare have negative contributive value.

Another proposal which one sometimes hears mentioned is the following:

> (3) A life has neutral welfare if and only if it has the same welfare as a life of complete unconsciousness.

This proposal has the advantage that it seems to agree with certain welfarist axiologies. A hedonist, for example, would probably agree that an unconscious life has neutral welfare. But again, coincidence is not conceptual identity. And it is clear that those that are not hedonist might reasonably object to (3) without being conceptually confused. For example, many objective list theorist would probably consider an unconscious life bad for the person living it. So, this definition is also subject to a version of Moore's open question argument, that is, one can reasonably ask whether an unconscious life is not worth living.

The next proposal draws on Rawls's famous "veil of ignorance" in the "original position". Let's say that you are in a position involving information constraints analogous to those suggested by Rawls for the parties in the original position: you don't know your place in society, your class position, social status, fortune, natural assets, abilities, intelligence, strength, your own particular conception of the good, particulars of your life-plan, and the like.[26] We can summarise this by saying that you don't know what kind of life you are living. Let's also say that if you are fully informed, you not only know all the true empirical facts but also the correct theory of welfare, be it an experientialist theory, a desire theory, an objective list theory or some combination thereof; and that you are rational if you maximise your lifetime welfare. We can now define a life with neutral welfare in the following manner:

> (4) A kind of life is of a neutral kind if and only if a person, who doesn't know which kind of life she is living but who is otherwise fully informed and rational, would be indifferent between living that kind of life and not continuing to live. A life has neutral welfare if and only if it is of a neutral kind.[27]

---

[26] See Rawls (1971), pp. 136-7.

[27] Cf. Blackorby, Bossert, and Donaldson (1997), p. 201: a "[l]ife is worth living as a whole, for an individual, if and only if lifetime well-being (utility) is above neutrality. A fully informed,

Although I find this proposal promising, I'm afraid that it will stretch our imagination beyond its limits. Recall that we are discussing lifetime welfare. A rational and fully informed person might not care about pains in the past, she might be biased towards the future. Assume that she is evaluating a type of life in which the good and bad parts balance out apart from some intense pain in the early childhood years. If she knew that her early childhood years had already passed, she would be indifferent between leading this kind of life and not continuing to live. *Ex hypothesis*, she cannot know this behind her veil of ignorance. But it seems very hard to imagine how our evaluator could be fully informed and rational and believe that she is in her early childhood years. And if she cannot believe that, then she will assign neutral welfare to lives which intuitively don't fit the description.

A version of (4) can handle this objection better:

> (5) A kind of life is of a neutral kind if and only if a person, who doesn't know which kind of life she is living but who is otherwise fully informed and rational, would be indifferent between living an extra life of that kind and living no extra life. A life has neutral welfare if and only if it is of a neutral kind.

Although this proposal still demands quite a bit from our imagination, I don't know about any fatal objections to it. One might perhaps worry that we, because of evolutionarily ingrained instincts, tend to overvalue the prospects of living another life. But presumably a fully informed and rational evaluator should be able to disregard such evolutionary biases. And since she doesn't know what kind of life she is now leading, she won't be biased against certain kinds of lives because she is already living a similar life. Is it possible to deny that an extra life preferred by such an evaluator is a life worth living? It seems difficult to deny that but I must admit that this reduction puts quite high demands on our imagination and it is hard to see the conceptual connection. And isn't it based on a conflation of the value of a life for a person and the contributive value of one life to the value of a series of lives

---

selfish, rational person whose utility level is below neutrality prefers not to have any of his or her experiences."

for a person? Rather than as a criterion of a life with neutral welfare, I think (5) is best understood as an heuristic device which we can employ when we try to figure out whether a life is worth living or not.

Some people think that a person can benefit from coming into existence. The next proposal draws on that idea:

> (6) A life has neutral welfare if and only if it is equally as good for the person to live such a life as that she should never have lived at all.

This proposal, in combination with (*) above, yields that it can be better or worse for a person to live a life than not to live at all: A life with positive welfare is a life that is better for the person living it than not living at all, and *vice versa* for a life with negative welfare. However, a number of theorists have rejected this since they think it implies that it can be better or worse for a person not to exist than to exist, an implication which they consider absurd. John Broome, for instance, writes that

> The expression 'has value to the person whose life it is' might also suggest a third possible meaning: a life is worth living if it is better for the person that she lives than that she should never have lived at all. I have not mentioned this as a possible meaning before, because I think it makes no sense. At least, it cannot ever be *true* that it is better for a person that she lives than that she should never have lived at all. If it were better for a person that she lives than that she should never have lived at all, then if she had never lived at all, that would have been worse for her than if she had lived. But if she had never lived at all, there would have been no her for it to be worse for, so it could not have been worse for her.[28]

---

[28] Broome (1999), ch. 10, p. 168 (emphasis in original). Cf. Narveson (1967), p. 67: "If you ask, 'whose happiness has been increased as a result of his being born?', the answer is that nobody's has. - - - Remember that the question we must ask about *him* is not whether he is happy but whether he is happier as a result of being born. And if put this way, we see that again we have a piece of nonsense on our hands if we suppose the answer is either 'yes' or 'no'. For if it is, then with whom, or with what, are we comparing his new state of bliss? Is the child, perhaps, happier than he used to be before he was born? Or happier than his alter ego? Obviously, there can be no

I agree with the last part of Broome's argument: There is no sense in saying that a non-existing "person" can be better or worse off. But perhaps one can deny the implication that if a person can be better or worse off existing, then a "person" can be better or worse off not existing. As Nils Holtug puts it:

> When saying that a person has been benefited by coming into existence, I mean that this person is better off than if he had never existed. Of course, normally, if a person is better (worse) off in a situation X than in a situation Y, he is worse (better) off in situation Y. While this is normally true, it is not true when Y involves his nonexistence. And there is a perfectly natural explanation for that. The property of "being worse off", like other properties, does not apply to people in worlds in which they do not exist.[29]

It is clear that a state X is better than a state Y if and only if state Y is worse than state X. The critics of (6) assume that this logic also holds for "better for", that is, a state X is better for a person than another state Y if and only if state Y is worse for the person than state X. What Holtug suggests is that the logic of "better for" doesn't follow this pattern since it is not applicable to non-existing people. Rather, I think he suggests the following pattern:

> (i) If a person *p* exists in both state X and Y, then state X is better (worse, equally as good) for *p* than (as) state Y if and only if state Y is worse (better, equally as good) for *p* than (as) state X.
> (ii) If a person *p* exists in state X but not in Y, then state X can be better (worse, equally as good) for *p* than (as) state Y although state Y is not worse (better, equally as good) for *p* than (as) state X.

---

sensible answer here." (emphasis in original) See also Parfit (1984), pp. 395, 489, and Heyd (1988). Cf. the discussion in section 8.2.

[29] Holtug (1996), p. 77.

I agree with Holtug that we shouldn't assume that the logic of "better for" is the same as the logic of "better" and I think that he has given an explanation why this is not so: "better for" is only applicable when a person to which the "for" in "better for" refers to exists. Holtug has more explaining to do, however. If a state X is better than another state Y for a person *p*, we would expect X to be either good, bad, or neutral in value for *p*, that is, we expect it to have some kind of value or disvalue for *p*. We would also expect the same from state Y. Now, if a state X can be better for *p* than state Y although state Y is not worse for *p* than state X, then it is fair to ask what value Y has for *p*. Is it good for *p*? Bad? Or neutral in value? But if we claim that it is good, bad, or neutral in value for *p*, then we again have a piece of nonsense on our hands, since how could a state be good, bad, or neutral in value *for* a person when that person doesn't exist?[30] Holtug also has to revise this part of our logic. He has to claim that a state X can be better than another state Y for a person, although Y is neither good, bad, nor neutral in value for *p*. I think that Holtug probably would accept this, invoking the same explanation as the one he used above: "good for", "bad for" and "neutral in value for" are not applicable to people in worlds where they don't exist.

If we follow Holtug's suggestion, then it looks like we can endorse (6) without committing ourselves to any absurd ascription of welfare to non-existing people, and, together with (*), we can use it to reduce all categorical welfare statements to comparative welfare statements. Of course, one may consider Holtug's revision of the logic of "better for" an all too high price to pay. A reduction that doesn't involve such a radical revision of our logic would be clearly better. Notice also that the belief that a person can be benefited or harmed by coming into existence doesn't commit one to (6) and (*) since one can reject these and still claim that it can be good or bad for a person to come into existence. Let's turn to the last proposal.

Earlier we discussed different theories of welfare, that is, theories about which components make a life better or worse for the person living it. A straightforward proposal would be to say that a life has neutral welfare if and only if the intrapersonal aggregation of its welfare components is neutral. How to measure and

[30] This argument came out of a discussion with Krister Bykvist.

aggregate welfare intrapersonally is a contentious matter, however, so let's get rid of that part in the definition:

> (7) A life has neutral welfare if and only if it has the same welfare as a life without any good or bad welfare components.

This definition expresses, I think, the kind of conceptual connection we are looking for. Could one claim, for instance, that a life has negative welfare if it doesn't involve any bad things at all? Could a life without any good things be good for the person living it? That seems implausible.

This definition presupposes, of course, that we have a criterion determining which welfare components are good or bad for a person: It utilises the cut-off point between positive and negative welfare components. A hedonist, for example, would typically say that pain is bad and pleasure is good for a person, and, consequently, that a life without any pleasure and pain has neutral welfare. Similarly, a desire theorist holds that (some) fulfilled desires are good and frustrated desires are bad for a person, and that a life without any fulfilled or frustrated desires has neutral welfare.

The following might strike one as a problem, however. From a hedonist perspective, it is reasonable to say that there are experiences which are neither pleasurable nor painful, there are, so to speak, experiences which are neutral in value for a person (Of course, had she experienced pleasure at that moment, then she would have been better off, and the fact that she didn't is bad for her in that sense). Likewise, from the perspective of the desire theorist, there are states which are neutral in value for a person. For example, the mere absence of desires that would have been fulfilled or frustrated if one had had them, neither adds nor detracts from one's well-being – the fact that you don't have the desire that my office table should be asymmetrically shaped, which it is, doesn't count negatively towards your overall welfare. Moreover, there are desires whose fulfilment is, arguably, of neutral value for the person having them, such as the desire not to have headache. This might not hold for all kinds of welfarist axiologies, however. Take, for example, one of the standard items on an objective list of welfare components: Health. Here it seems like there are just two possible states of a person: Either she is healthy, which many would take as good for her, or she is unhealthy, which is bad for her, and no state in between which is of neutral value

for her. Consequently, if such an objective list theory is correct, then it seems that there cannot be any lives without any good or bad welfare components since a life without the good component "being healthy" will contain the bad component "being unhealthy". So (6) only works for welfarist axiologies which imply that there are possible states of a person which only involve components which are of neutral value for her.

I don't think this is a devastating problem for (7), however, since it seems probable that one can decompose or reformulate such components into ones that are well-behaved. For example, instead of talking about "health" in the abstract, we could talk about the different components that make up health. One such component is body temperature. It seems reasonable to claim that although having a fever is bad for a person, having a normal body temperature is of neutral value for her. I don't think that we would say that some extra days in a person's life with normal body temperature and in which the other welfare components balance out, increase that person's overall well-being. Of course, my claim that one can decompose health into components which can be of neutral value for a person implies that there is some state of a person's health which is of neutral value for her. So the problem is rather a verbal one, what to call such a state. For example, we could say that people are either in good, bad, or fair health.

So this proposal looks quite promising. We haven't, however, yet defined "neutral welfare" in terms of some comparative concepts, but in terms of a pair of categorical concepts, that is, in terms of good and bad welfare components. To reach this end, we need a way of defining good and bad welfare-components in terms of some comparative concepts. Here's a try:

> (**) A welfare component x is good (bad) for a person p if and only if p's life would have lower (higher) welfare without this component x, other things being equal.

If we combine (**) with (7) and (*), then we can deduce all of our welfare concepts from the basic ordering of lives in terms of the relation "has at least as high welfare as". We start by picking out the good and bad things in a person's life by determining which welfare components have decreased or increased the welfare of her life. By removing all the good and bad things from her life, we get a life that only contains welfare components with neutral value for her, that is, a life with

neutral welfare. Finally, a life with positive (negative) welfare is a life with higher (lower) welfare than a life with neutral welfare.

Let me now briefly summarise this section. We looked at seven putative reductions of the categorical concepts "positive welfare", "negative welfare", and "neutral welfare". We definitely rejected the first four of these, and suggested that the fifth proposal is best understood as an heuristic device for deciding which lives that have positive, negative or neutral welfare. The sixth proposal together with (*) could be used to reduce all categorical welfare statements to comparative welfare statements, but it also involved a quite radical revision of the logic of "better for". We found the seventh proposal the most promising: It expressed the kind of conceptual connection that we were looking for and it didn't involve any revision of the logic of "better for". In combination with (*) and (**), it could be used to reduce all categorical welfare statements to comparative welfare statements. I think, however, that (7) by itself the most plausible way to understand a life with neutral welfare. Most theories of welfare will tell us which components of a life count as positive and negative. Once we know which components count as good and bad, we can use (7) to define a life of neutral welfare and (*) to define lives with positive and negative welfare.

Notice also that on both the sixth and seventh proposal, which lives count as lives having positive or negative welfare will depend on which theory of welfare they are combined with, not on any special feature of these reductions.

2.2.4 Ordering Presuppositions of Population Theories and Measurement of Welfare

All of the population axiologies proposed in the literature presuppose stronger demands on the ordering of lives than we did above, and most of them assume that much more information about people's welfare is available. They talk about, for example, "the total sum of welfare" or that in a change from one population to another, "the worse-off half would gain more than the better-off half would lose".[31] These theories presuppose different kinds of *measurement* of welfare. In this context, measurement refers to some manner of assigning numbers to the objects one wants to measure (for example, mental states of happiness) in a way that makes

[31] For the latter statement, see Parfit (1984), p. 426.

it possible to represent the qualitative relations between those objects with quantitative (mathematical) relations between the assigned numbers.[32]

Assume that we could order all possible lives with the relation "has at least as high welfare as". This ordering could be represented by a function *f* that assigns numbers to each life.[33] For example, suppose that John has higher welfare than Chandra who has the same welfare as Steve. The relation between these three people's welfare could be represented by a function for which *f*(John) = 1, *f*(Chandra) = *f*(Steve) = 0. The significance of these numbers depends on some technical conditions on the ordering of lives (which need not concern us here).[34] Given that some conditions are fulfilled, it makes sense to speak about the difference in welfare between individuals, given that some other conditions are fulfilled, it makes sense to speak about sums of welfare, and so forth. In the jargon of measurement theory, welfare can be measured on different scales. There are a number of different scales. We shall only present the most common ones, using S. S. Stevens's classification.[35]

An *ordinal* scale is a scale which is unique up to an order-preserving transformation, that is, any transformation of the scale that preserves the order of the values yields another admissible scale. The admissible transformations are all functions satisfying the condition that x has at least as high welfare as y if and only if $f(x) \geq f(y)$, that is, strictly monotone increasing transformations. Claims such as "John has higher (lower, the same) welfare than (as) Chandra" make sense. One cannot say anything about the welfare differences of lives.[36]

---

[32] My account of measurement theory is taken from Roberts (1984).

[33] In the "standard approach" to measurement of welfare, one doesn't start out with orderings of lives but with orderings of those things that make life better or worse, the components of welfare. In a hedonist approach, for example, the welfare components are experiences of pleasure and pain which are ordered by the relation "is at least as pleasurable (painful) as". This ordering is represented by a function which assigns numbers to each type of welfare component. Orderings of lives are then derived from orderings of welfare components by means of some aggregation method. Whether one starts with orderings of welfare components or orderings of lives makes no difference for the discussion below.

[34] For these conditions, see Roberts (1984).

[35] See Roberts (1984), pp. 64-6.

[36] There are special cases, however, where it is possible to compare differences of value using only ordinal information. Suppose the welfare of three lives could be ranked, from greatest to least welfare, in the following order: x, z, y. Here the difference in welfare between x and y must be greater than the difference between x and z.

An *interval* scale is unique up to a positive linear transformation. This means that not only is the order of the scale values preserved but also the order and ratios of differences between scale values. The admissible transformations are functions of the form $f(x) = \alpha x + \beta$, $\alpha > 0$. Statements such as "the difference between John's and Chandra's welfare is greater (less, the same) than (as) the difference between Krister's and Erik's welfare" are meaningful. On this scale, we can meaningfully compare welfare differences and ratios of welfare differences.

A *ratio* scale is unique up to a similarity transformation, which means that the ratios of scale values are preserved. The admissible transformations are all functions of the form $f(x) = \alpha x$, $\alpha > 0$. Sentences such as "John has many times higher (lower) welfare than Chandra" are meaningful. One can meaningfully compare ratios of welfare.

Most of the theories proposed in population axiology seem to presuppose the ratio scale. With such a scale, talk of the total and average amount of welfare in a population makes sense. The interval scale hasn't gained much explicit attention, but seems to be implicit in some reasoning in population axiology (as is suggested by the statement quoted from Parfit above). On this scale, average welfare is well-defined but not total welfare.[37] A few theorists only make use of the ordinal scale. With this scale, neither average nor total welfare is well-defined. Let us also note that all of these theories presuppose complete interpersonal comparability of welfare.

Lastly, let's turn to a possible confusion. One might believe that talk about positive and negative welfare presupposes measurement on at least a ratio scale since "neutral welfare" has to be represented by zero under all transformations.

---

[37] An example: Let's say that population A consists of life p_1, population B consists of lives p_2 and p_3, and that on some assignment of numbers, p_1 has welfare $10u$ whereas p_2 and p_3 have welfare $5u$. The total welfare in A is the same as the total welfare in B, namely $10u$. Let $f(x)=x+10$. Then the total welfare of B is $2(5+10) = 30u$ which is greater than the total welfare of A which is $10+10 = 20u$. In other words, total welfare is not well-defined on an interval scale. The average welfare in A, before the transformation of the numbers representing individual welfare, is $10u$ which is greater than the average welfare of $5u$ in B. This inequality is preserved also after the transformation since the average welfare in A is then $10+10 = 20u$ whereas the average in B is $2(5+10)/2 = 15u$. For uniqueness theorems for these two scales, see Roberts (1984).

Thus, any condition which involves the significance of positive or negative welfare cannot be meaningfully formulated with weaker scales.[38]

This is not the case. Let's first clear up a possible but, I hope, unlikely source of confusion. Talk about lives with positive or negative welfare has nothing to do with numbers. It is shorthand for the more cumbersome expressions "a life which is good for the person living it" and "a life which is bad for the person living it". Our use of the terms "positive" and "negative welfare" is analogous to their use in sentences such as "That's a positive attitude", "I got a negative answer", "Smoking has a negative impact on people's health", and the like. We use these terms to convey the positive/negative polarity of welfare. We could have used other terms such as "good/bad welfare", "satisfactory/unsatisfactory welfare", "sufficient/insufficient welfare", "adequate/inadequate welfare", and so forth.

Positive and negative welfare are properties of lives and talk about positive and negative welfare doesn't presuppose measurement on *any* scale any more than talk about men and women presupposes measurement of gender on any scale. If one so wishes, one can represent these properties with numbers but the meaningfulness of these concepts certainly doesn't presuppose any kind of measurement.[39] Consequently, conditions that involve talk about positive and negative welfare don't necessarily presuppose measurement on a ratio scale. Here's one example: It would be good to decrease the number of people with negative welfare. Compare with: It would be good to increase the number of women in the parliament. Clearly, none of these statements require measurement on a ratio scale of the involved properties, welfare and gender.

We can reformulate the objection, however. We could claim that if we were to represent a relational system involving the set of all possible lives, the relation "has at least as high welfare as" and the predicate "has neutral welfare" with a numerical relational system, then the scale type defined by the class of admissible transformation would be a ratio scale.

---

[38] An anonymous referee of one of my papers made this claim and, in personal communication, several people have made similar claims.

[39] If we want to represent the gender properties of all possible people with numbers, then we could choose whatever pair of numbers x, y such that $x \neq y$. For example, (1, 0) or (1, -1) or (356, 518), and so forth, would do. Of course, this representation implicitly assumes, which might be controversial, that there are only two gender properties, *viz.*, male and female.

An example will show that this is not true. Assuming that we would like to represent lives with neutral welfare with zero (although we could choose some other number), and that the relation "has at least as high welfare as" is complete over the set of all possible lives (which we haven't assumed), a numerical representation f of the relational system in question should fulfil the following conditions:

(i) x has at least as high welfare as y if and only if $f(x) \geq f(y)$;
(ii) x has neutral welfare if and only if $f(x)=0$.

Clearly, the class of admissible transformation here consists of all function g such that $g(f(x)) \geq g(f(y))$ if and only if $f(x) \geq f(y)$, and $g(f(x))=0$ if and only if $f(x)=0$. An example is $g(f(x))=x^2$ for all $f(x) \geq 0$, $g(f(x))=-x^2$ for all $f(x) < 0$. But the class of admissible transformation defining a ratio scale consists only of functions of the form $g(f(x)) = \alpha f(x)$, $\alpha > 0$.

So what kind of scale type does this numerical representation involve? As a matter of fact, it doesn't seem to fit anywhere in Stevens's classification. Intuitively speaking, it seems to be very closely related to the ordinal scale. Statements that are meaningful on the ordinal scale are also meaningful on this scale, whereas some statements that are characteristically meaningful on the interval and the ratio scale – statements regarding welfare differences, ratios of welfare differences and ratios of welfare – are not meaningful on this scale. How to in an exact manner classify numerical representations of the kind of relational systems that we are dealing with here, how such a classification would relate to Stevens's classification, and uniqueness theorems for such a representation, are general problems in measurement theory that needs careful consideration but it would take us to far afield from the main topic to pursue this question further here.[40] The important point is this: None of the adequacy conditions that we shall suggest presuppose that one can meaningfully speak about total and average welfare. This is the important point since some theorists have suggested that the assumption that one can compare the total or average welfare of populations is the root of many of the

[40] For a discussion of how to understand Stevens's classification, see Wedberg (1968) and Odelstad (1990).

problems in population axiology, and that we can solve these problems by rejecting this assumption. As we shall show, this is not correct.

2.3 The Definition of a Population

How should we define a population? One definition presents itself directly: A population is a set of people. We are not interested in people as such, however, but in their lives, and on any reasonable conception of personal identity, the same person can lead several different lives. It would suit our purposes better to define a population in terms of lives: A population is a set of lives. As we shall understand it, a life is individuated by the person whose life it is and the kind of life it is, and two populations are identical if and only if they consist of the same lives. Since the same person can exist (be instantiated) and lead the same kind of life in many different scenarios, histories, or, as we shall call it, possible worlds, the same life can exist in many possible worlds. Moreover, since two populations are identical exactly if they consist of the same lives, the same population can exist in many possible worlds.

This definition includes, however, some odd sets of lives as populations, such as the set of all of my possible lives, infinite sets of lives, a set consisting of the first person and the last person born in a possible history of the universe, a set consisting of people that live light years apart, and so forth. We might want to put some constraints on the populations that we are to consider. In moral philosophy, where "intuitions" and "considered beliefs" play a part, one comes across different cases that are used as "tests" for moral principles. These are cases that we, supposedly, have firm beliefs about and we can test a principle by checking whether it complies with our considered beliefs in those cases. Most often, this kind of testing has to rely on hypothetical cases rather than actual ones. One could object that such examples are "unreal" or "artificial" or "hard to imagine" and therefore shouldn't have any implications for our moral beliefs and for our choice between competing moral theories.[41]

Let's call a possible population whose existence is compatible with the laws of logic a *logically possible population*; call a population whose existence is compatible with both the laws of logic and natural science (including "laws", if there are any, about

[41] See, for example, Hare (1981), pp. 5, 47-49, 113-116, 194-96, and Donagan (1977), pp. 32, 35.

human psychology) a *nomologically possible population*; call a population whose existence is possible given our present technology and resources a *technically possible population*.[42] A logically possible population may involve nomological and technical impossibilities; a nomologically possible population may involve technical impossibilities.

Some of the populations that we are going to consider involve technical impossibilities. But what is technically impossible today may not be so tomorrow. We don't know much about what will be technically impossible in the future. Just think about the rapid change in medical technology during the last thirty years or so. It doesn't seem wise to refrain from thinking about the moral and political aspects of future technology. The technically possible criterion rules out too many test cases.[43]

Is the restriction to only nomological possible cases reasonable? We shall contend that all the cases we are going to discuss are nomologically possible. I am, however, sceptical about this criterion. As we shall see in chapter 3, this criterion might make the truth or correctness of a moral theory dependent on some speculative facts that seem irrelevant from a moral perspective. It also allows for test cases that we should be sceptical about. We have epistemological reasons to discard some nomologically possible cases, namely cases which are so complex and multifaceted or involve such peculiar aspects that we cannot clearly conceive and grasp what facts they really involve. Our intuitions about such cases are bound to be unreliable. Let's call a population that doesn't involve such epistemological problems an *epistemologically unproblematic population*. Such a population might involve technical and nomological impossibilities but not, I suppose, logical impossibilities.[44] I think that this criterion, albeit vague, captures our misgivings about certain test cases much better than the nomological criterion. Nomologically

---

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

[43] Cf. Parfit (1984), p. 390.

[44] Might not all nomologically impossible cases be so peculiar that they are epistemologically problematic? If so, the restriction to nomologically possible test cases would be superfluous. I don't think this is necessarily the case, however. Imagine that someone was torturing innocent people with a device that is impossible to construct according to the laws of physics. Still, we have no problem to morally evaluate such a case. We shall in section 3.1.3 discuss a case whose nomological status is unclear but which is epistemologically unproblematic.

possible cases might still be epistemologically problematic. Consider populations of infinite size and populations involving people with "infinite welfare". It is not clear whether such populations are excluded by the nomological criterion, at least not the first kind. A swift glance at the history of mathematics, however, will show how many great minds have blundered when working on problems that involve some kind of infinity. Intuitions about cases that involve infinity are notoriously unreliable. Arguably, it is even questionable whether talk about "infinite welfare" makes any sense. Consequently, in line with the epistemological criterion, we shall assume that populations are finite sets of lives with finite welfare.

Many of the cases that we shall consider will fail to be likely cases, that is, cases that we often find in actual choice situations. This failure seems unavoidable if we want to present cases that we can easily grasp and that we will have firm intuitions about. In other words, this follows from our choice of epistemologically unproblematic cases. Realistic examples, such as probable future policy choice situations or historical cases, involve so many morally relevant aspects that all simplicity and transparency are lost. At any rate, it is reasonable to assume that in order to evaluate alternatives in regard to every morally relevant factor, one must evaluate each factor in turn, before one can make an overall evaluation where each factor is given its proper weight. Consequently, simplified cases that only involve a few morally relevant properties can be seen as aspects of more complicated real life cases.

What about the latter two of the odd populations we mentioned in the beginning of this section (the set consisting of the first person and the last person born in the history of the universe, and the set consisting of people that live light years apart)? Are these ruled out by the epistemological criterion? That is unclear. Anyhow, there are other reasons to be careful about our intuitions regarding such cases. How we value populations might depend on some structural features which depend on the location of lives in time and space. For example, our intuitions might depend on the possibility of some kind of interactions between people. An egalitarian might think that inequality is bad only if it holds between people that have some kind of interactions or at least some possibility of interactions. She could deny that it is bad from the perspective of equality that we are better off than the ancient Egyptians and people living on such a distant planet that we cannot affect their well-being in any way; but she could still hold that it is bad that contemporary Europeans are better off than contemporary Africans. An average

utilitarian might argue that whether it is good or bad to create a person with quality of life below average depends on what average we are talking about. She might not care about the average of the last millennium but rather about the average of presently existing people.

How to define the axiologically relevant temporal and spatial demarcation of a population is a tricky question that egalitarianism and other context-sensitive theories face. We shall leave this problem aside, however.[45] In the formulations of our arguments, we shall not appeal to intuitions that depend on people's location in time and space. In other words, the arguments and conditions that we shall discuss are relevant irrespective of whether the people in the populations we consider are contemporaries and have interactions in more or less the same fashion as the present people on earth. In most of our discussion we shall assume that the populations that we compare are possible alternative future populations since it makes the examples easier to grasp.[46]

Lastly, some matters of terminology. Let A, B, C, and so on, denote populations. Unions of populations, denoted as $A \cup B$, $A \cup B \cup C$, and so forth, are also populations, given that the aforementioned restrictions are satisfied, that is, that they are logically possible and epistemologically unproblematic. The number of lives in a population A (A's population size) is given by the function $N(X)$. We shall sometimes use expressions such as "a population with positive welfare", "a population with negative welfare", and so forth. Such expressions are elliptical for the more cumbersome phrases "a population consisting only of lives with positive welfare", "a population consisting only of lives with negative welfare", and so forth.

2.4 The Definition of a Population Axiology

A population axiology is an ordering of populations in regard to their goodness. The goodness in question is not their instrumental goodness but how good they are

---

[45] For a discussion of these problems, see Parfit (1984), 420-2, Arrhenius (1995), Arrhenius and Bykvist (1995), Blackorby, Bossert, and Donaldson (1995, 1997), Carlson (1998a).

[46] It doesn't seem defensible, however, to restrict the scope of a population axiology to only such populations. We do rank populations which consist of people who are in the past, for example, the population of Norway in the 18th century as compared to the population in the 20th century. As a matter of fact, many of the theories in the literature have been put forward as theories about how to rank whole possible worlds, that is, populations consisting of all the lives that will ever be lived. We shall return to the question of the scope of a population axiology in section 8.3.

in themselves, their intrinsic goodness, as it is sometimes expressed. More exactly, we shall define a population axiology as an "at least as good as" quasi-ordering of all possible populations. This is a minimal and very undemanding definition of a population axiology. Recall that a quasi-ordering is reflexive and transitive but not necessarily complete. In other words, we leave open the possibility that there might be incommensurable populations. Most of the axiologies in the literature exclude this possibility since they yield a complete ordering of all possible populations. We shall refer to those axiologies as *complete* population axiologies. Moreover, we are not committed to welfarism, the view that welfare is the only value that matters from the moral point of view. On the contrary, other considerations such as fairness, liberty, virtuousness, and the like may figure in the ranking of populations. We shall only assume that welfare at least matters when all other things are equal. Although we shall not defend this claim, this assumption is arguably a minimal adequacy condition for any moral theory.

3

Total and Average Utilitarianism

3.1 Total Utilitarianism and the Repugnant Conclusion

Utilitarianism is often taken to be the normative theory that tells us to maximise welfare. This theory can be broken down into two components: Consequentialism, which is the view that an action is right if and only if it maximises the good, and an axiological component which identifies the good with the sum total of people's welfare.[1] In this chapter, we shall discuss the axiological component of Utilitarianism. Likewise, many of the theories that we shall discuss in chapters 3-9 were originally formulated as normative theories, but we shall focus the discussion on the axiological part of these theories. We shall postpone the discussion of normative population theories until chapter 11. We shall refer to the axiological component of Utilitarianism as "Total Utilitarianism". According to Total Utilitarianism the contributive value of a person's life equals her welfare, and the value of a population is calculated by summing the welfare of all lives in the population:

$$TU(X) = \sum_{i=1}^{n} u_i = u_1 + u_2 + \ldots + u_n$$

In the above formula, n is the population size of X and u_i is the numerical representation of the welfare of the i:th life in population X. Since Total Utilitarianism sums welfare levels, it presupposes that welfare is measurable on a scale at least as strong as a ratio scale.

[1] See Sumner (1996), p. 3, for a similar definition of Utilitarianism. Sumner breaks down Utilitarianism into three parts: consequentialism, welfarism, and aggregation. The two latter parts correspond to what I call the axiological component of Utilitarianism.

Derek Parfit has put forward what seems to be a devastating argument against Total Utilitarianism. It implies the Repugnant Conclusion:

> *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.[2]

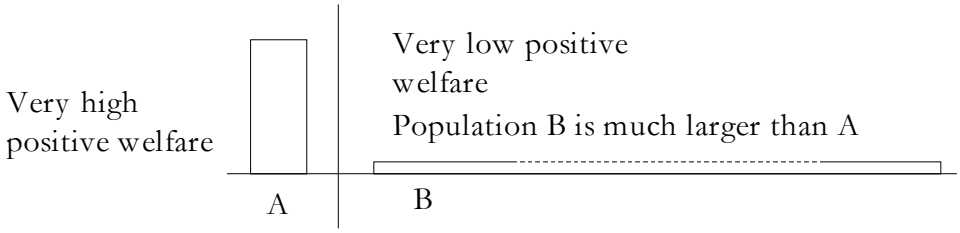

Diagram 3.1.1

The blocks in Diagram 3.1.1 represent two populations, A and B. The width of each block shows the number of people in the corresponding population, the height shows their welfare. People's welfare is much lower in B than in A –

---

[2] Parfit (1984), p. 388. Although it is through Parfit's writings that this implication of Total Utilitarianism has become widely discussed, it was already noted by Henry Sidgwick before the turn of the century: "… [I]f 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." See Sidgwick (1907) p. 415 (emphasis in original). Perhaps it can be said that the Repugnant Conclusion is anticipated in William Whewell's argument from 1852, that if quantity of pleasure in the effects is the test of conduct, then Jeremy 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". See Whewell (1852), quoted in Acton (1972). For other early sources of the Repugnant Conclusion, see Broad (1930), pp. 249-50, McTaggart (1927), pp. 452-3, and Narveson (1967).

everybody is worse off in B as compared to A – but since there are many more people in B, the total sum of welfare in B is greater than in A. Hence, the Total Utilitarian Principle ranks B as better than A – an example of the Repugnant Conclusion.[3]

As the name indicates, Parfit finds this conclusion very counter-intuitive and most commentators agree. There are, however, a few who disagree. We shall shortly take a look at their arguments for the acceptability of the Repugnant Conclusion, but let us first look a little bit closer at Parfit's formulation of it.

Parfit talks about populations "of at least ten billion". Why ten billion? It seems pretty arbitrary. What Parfit has in mind is that many people "would agree that, in the world now [with approximately five billion inhabitants], the value of extra quantity [of welfare] can be regarded as having reached its limit. This is why I supposed that, in my imagined outcome A, there would be ten billion people living. This makes my example relevantly similar to the actual world…"[4] The intuition that Parfit is referring to seems to be that in regard to small populations with positive welfare, both the average and the total amount of welfare matters and an increase in the latter can outweigh a decrease in former. For large populations with positive welfare, on the other hand, average welfare has priority and an increase in total welfare cannot outweigh a large decrease in people's welfare. Parfit thinks that those who accept this view would also agree that when a population reaches ten billion, then the value of total welfare has clearly reached its limit.

This explanation of why the Repugnant Conclusion is hard to accept is just one of many possible explanations. One might think, for example, that only average welfare matters, or that the existence of people with positive welfare has no value in itself, but given that people already exist, it is better that they are better off than worse off, and so forth. One might also deny that there is any explanation of the belief in the unacceptability of the Repugnant Conclusion and hold that it is just a basic non-inferential axiological belief. We shall come across several ideas that purport to explain the unacceptability of the Repugnant Conclusion. It is interesting to note that on some views, there are versions of the Repugnant Conclusion which

---

[3] For those axiologies that include animals in the welfare calculus, B could be a population of sheep or some other animal. See, for example, Singer (1993) and Blackorby & Donaldson (1992, 1997).

[4] Parfit (1984), pp. 402-4. He doesn't find this view defensible. See pp. 405-12.

would be deemed unacceptable even if population A only consists of one person. This holds, for example, for the first two views mentioned above. Views such as the one from Parfit quoted above, on the other hand, where there is some kind of limit to the value of total welfare, must posit some cut-off point where the value of total welfare has reached its limit. We might find such a cut-off point arbitrary: How could we determine such a limit? The only way to do this, I suppose, would be by testing one's intuitions over different cases. Surely, the borderline is going to be vague: There is going to be a grey area where one will be very unsure about one's beliefs and perhaps find the compared populations incommensurable. That the distinction between two things has a vague boundary is not, however, a knock-down argument against its correctness or usefulness. We don't want to grant Sextus Empiricus 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. It is just to construct a clear case that Parfit stipulates that the high quality population consists "of at least ten billion people", although he thinks that those who believe in a limit to the value of total welfare would agree that the value of total welfare has reached its limit already with five billion people.

How "large" the A-population must be in order for the Repugnant Conclusion to be clearly unacceptable might not only vary with different axiological ideas but also with what kind of populations we are considering. Parfit formulated his conclusion in the context of the global optimum population of contemporaneous people. Some theorists have assumed other contexts such as the optimal population in the history of the universe. Surely one might have different ideas about, for example, what the limit of the value of total welfare is in such a context as compared to the context which Parfit assumed. There are other contexts too, where versions of the Repugnant Conclusion are applicable but which are seldom discussed, for example, the size of a population in a country or the size of a family. Compare a couple with very high welfare that has one child with very high welfare to a couple with very low welfare that has twenty children with very low welfare (from exhaustion and lack of resources one might presume).

A theory could avoid the Repugnant Conclusion in an unsatisfactory way which I don't think was intended by Parfit: It could stipulate that at least one large population enjoying very high welfare is incommensurable with all larger

populations with very low welfare (that is, the former population is neither at least as good as, nor worse than, the latter populations), but no population with very high welfare is at least as good as all populations with very low welfare. Perhaps it can be reasonably believed that some populations with very high welfare are incommensurable with some populations with very low welfare, but that some of them are incommensurable with all larger populations with very low welfare seems, given that other things are equal, clearly counter-intuitive.

Let's formulate a condition that captures the intuition behind the Repugnant Conclusion and avoids the drawbacks of Parfit's formulation:

> *The Quality Condition*: There is at least one perfectly equal population with very high welfare which is at least as good as any population with very low positive welfare, other things being equal.

Avoidance of the Repugnant Conclusion implies that there is at least one population with at least ten billion members with very high welfare which is at least as good as or incommensurable with all larger populations with very low welfare. The Quality Condition is in one sense logically stronger than avoidance of the Repugnant Conclusion since it rules out axiologies that imply that at least one population with very high welfare is incommensurable with all larger populations with very low positive welfare but none is at least as good as all populations with very low welfare. It is a logically weaker and a more general condition in another sense since it doesn't specify any size of the population(s) with very high welfare which is at least as good as all populations with very low positive welfare. Consequently, it is applicable in different population contexts and is compatible with a wider range of axiological beliefs than is the avoidance of the Repugnant Conclusion. The Quality Condition is a weaker requirement in a second sense too, since it requires that there is perfect equality in the population(s) with very high welfare which is at least as good as all populations with very low welfare.[5]

---

[5] The Quality Condition doesn't imply avoidance of the Repugnant Conclusion, and vice versa, but given full comparability among populations and satisfaction of a weak dominance condition which we shall introduce below, avoidance of the Repugnant Conclusion imply satisfaction of the Quality Condition. For a proof, see Appendix B.

It is a very weak condition. A theory which implies that most but not all large populations with very high welfare are worse than some populations with very low welfare doesn't violate the Quality Condition. Likewise, neither a theory which yields that only one perfectly equal population with very high welfare is commensurable with all populations with very low positive welfare, nor a theory that deems all such populations to be equally good, violates the Quality Condition (nor do these theories imply the Repugnant Conclusion). If one holds that the Repugnant Conclusion is unacceptable, then it would be odd, one might argue, to accept theories such as the ones above. Arguably, the axiological intuition we have about the relation between high quality and low quality populations is much stronger than the Quality Condition. Perhaps we believe that if the high quality population is sufficiently large, then such a population and any larger high quality population is better than any low quality population. As true as this might be, one should remember that the Quality Condition is only a necessary and not a sufficient condition for an acceptable axiology and just one among a number of other conditions that we shall propose. Since we want our adequacy conditions to be as clearly acceptable for as many theorists as possible, in many instances they will be logically weaker than the corresponding intuitions that they trade on.

3.1.1 Other Things Being Equal

Parfit includes a *ceteris paribus* clause in his formulation of the Repugnant Conclusion: "if other things are equal". So did we in our formulation of the Quality Condition. What differences are supposed to be excluded by the *ceteris paribus* clause?

Roughly, the idea is that people's welfare is the only axiologically relevant aspect which may be different in the compared populations. In other words, the compared populations should be roughly equally good in regard to other axiologically relevant aspects. One might reasonably hold that welfare is not the only thing of importance, but that there are other things that have value in themselves which would not show up in our specification of the alternatives. Examples could be fulfilment of rights, autonomy, knowledge, cultural diversity or the genesis of a population. Assume, for example, that in population A, people live very similar lives. In population B, on the other hand, there is a vast number of different lifestyles and cultures. Under such circumstances, the Repugnant Conclusion might not look very appalling. Just as Parfit does, however, we can assume that with

respect to such considerations, the two alternatives do not differ in any axiologically relevant way. We can assume, for example, that there is roughly as much cultural diversity in population A as in B (it would not be absurd to claim, for example, that there is as much cultural diversity in Canada as in the United States).

For some considerations, the *ceteris paribus* clause does not seem to work as we wish. One might believe that desert is an important consideration: It is better that people get what they deserve than that they get less or more than they deserve. Assume that all the people in A and B are "terrible sinners" – they all deserve to be in hell with eternal torment. One might then think that B is better than A since it is better that "sinners" have a very low welfare rather than a very high welfare.[6] Here, the *ceteris paribus* clause seems to be satisfied since the people in A and B have the same desert factor: they all deserve eternal torment.

We shall take a closer look at desert in chapter 9, when we discuss Fred Feldman's desert adjusted utilitarianism. The objection above can, however, be met by interpreting the *ceteris paribus* condition along the lines suggested earlier: For two alternatives to be equal in regard to desert, they should be equally good in regard to desert, that is, equally good in regard to the contributive value of the fit between what people deserve and what they get. If the contributive value of the fit between people's welfare and desert is the same in A and B, for example, if the A-people have a higher desert factor corresponding to their higher welfare, and vice versa with the B-people, then the Repugnant Conclusion is, I suppose, as unacceptable as before.

One might think that the *ceteris paribus* clause rules out more than we want. Should we, for example, conclude that if other things are equal in population A and B, then A and B are of the same size? No, of course we don't intend the *ceteris paribus* clause to be read in such a restrictive manner (and neither did Parfit). How value varies with population size is one of the main questions of this essay. Likewise, we shall not conclude that if every part of the A-people's lives made them better off, then the same also holds for the B-people (although these parts did not make them as well off as the A-people). The A- and B-people might have a different mix of good and bad parts. Even if the A-people enjoy uniformly high welfare throughout their lives, this does not imply that the same holds for the B-

---

[6] A similar example can be found in Temkin (1994), p. 353-6.

people. The reason why the B-lives have very low positive welfare could either be, to paraphrase Parfit, that there are only enough ecstasies to just outweigh the agonies or that the good things in life are of uniformly low positive quality.[7]

Similarly, we do not demand that the lives in A and B are of the same length. Another reason why the B-people have very low welfare could be that their lives are pretty short, say, as short as the average life span in 17th century Europe, whereas the A-lives could be relatively long, say, as long as the average life span in late 20th century Europe. Every part of the A- and B-people's lives could be approximately equally good, but there could be much fewer good parts in the shorter B-people's lives.

It is worthwhile to stress this point. In the discussion of the Repugnant Conclusion, it is often assumed that the lives in population A and B are of the same length and that the B-people's lives are of uniformly poor quality. If we assume that the A-people enjoy uniformly high welfare throughout their lives, then there are at least six different ways in which the B-population could be said to enjoy very low positive welfare (this list is by no means exhaustive):

1. Same length of lives as the A-people, but much lower quality of the good things in the B-people's lives.
2. Same length of lives as the A-people, same quality of the good things but also very bad things in the B-people's lives.
3. Same length of lives as the A-people, but lower quality of the good things, and bad things in the B-people's lives.
4. Same quality of the good things as the A-people but much shorter lives in the B-population.
5. Same quality of the good things as the A-people but shorter lives and bad things in the B-world.
6. Shorter lives, lower quality of the good things, and bad things in the B-people's lives.

Some of the most striking exemplifications of a B-population can be made by imagining that people have very short lives, that is, instances of 4-6 above. We

[7] See Parfit (1984), p. 388.

could imagine, for example, that the majority of the B-lives are led by somewhat happy children that only live for a few years.[8]

We shall also let populations vary in regard to the identity of their members (do the same people exist in the alternative populations?) and their temporal and modal features (are the people involved presently or actually existing people or are they people that will exist irrespective of how we act?). These aspects play no role for the theories we shall consider in chapters 3-7 but are crucial for the theories discussed in chapter 8. For example, we may ask whether the people in the A-population in Diagram 3.1.1 also are part of the B-population, although with much lower welfare. Some people might find the Repugnant Conclusion even harder to accept if this were the case. Since we want to investigate how the value of populations might be affected by these aspects, such differences among populations are not ruled out by the *ceteris paribus* clause.

3.1.2 Is the Repugnant Conclusion Unacceptable?

Torbjörn Tännsjö argues that the Repugnant Conclusion is not at all repugnant but rather "an unsought, but acceptable, consequence of hedonistic utilitarianism":[9]

> How we judge the Repugnant Conclusion must in the end and perhaps primarily depend upon where we think the line 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 normally most people are quite close to this level and that they perhaps often fall below this level.[10]

Richard Hare and Jesper Ryberg put forward a similar argument:

---

[8] Cf. Blackorby, Bossert, and Donaldson (1995), p. 1304, and Ryberg (1996), pp. 154-62. One might find this example very unrealistic. With the current child mortality rate in the underprivileged areas of the world, however, this example might not be too far removed from reality (assuming that the children that die at a young age in those areas enjoy positive welfare).

[9] Tännsjö (1998), p. 162.

[10] Tännsjö (1991), pp. 42-3 (my translation).

> Let us imagine we are actually living in the lap of luxury (as, relatively speaking, I and most of my readers are). Even so, it will be open to us, if we want to resist …[the Repugnant Conclusion], to do so by claiming that on average even our life is only just above the critical point at which we stop preferring to exist.[11]

> …[W]e regard the Repugnant Conclusion as repugnant at least partly because we regard low-average lives as bad lives in the sense that they are significantly worse than normal privileged lives. What we can conclude therefore, is that if a life in the more populous outcome in the Repugnant Conclusion is not a bad life [in the above sense] then the conclusion is not repugnant.[12]

Tännsjö et al. seem to consider facts regarding the present welfare of "most people" in the privileged parts of the world as relevant to how we should understand the Repugnant Conclusion. I find it hard to understand how such facts have anything to do with the acceptability of the Repugnant Conclusion. It might very well be that at present, most people in the privileged parts of the world have very low or negative welfare and that we are deluded about the welfare level of "normal privileged lives". But why should this piece of information make us re-evaluate the Repugnant Conclusion? As far as I understand, this information is only relevant for how we should evaluate the current states of affairs in the world as compared to other possible states of affairs. If Tännsjö et al. are right, then we can conclude that the current state of the world is worse than we thought, but nothing follows concerning the comparative value of a population with very high welfare and a population with very low positive welfare.

The unacceptability of the Repugnant Conclusion doesn't depend on the welfare of actual people. It is surely a logical and nomological possibility that people could enjoy very high welfare and we have no problem imagining such lives (as a matter of fact, if we are deluded, then we are imagining such lives all the time). Let's assume that Tännsjö et al. are right and that the current world population

---

11 Hare (1993), p. 74.
12 Ryberg (1996), p. 154.

consists of people with very low positive welfare. Which of the following two futures would be the best? In the first scenario we have a massive expansion of the population size but all the people still have very low positive welfare. In the second scenario, the population size remains the same but we have a major increase in people's welfare such that everybody enjoys very high welfare. The answer seems obvious. And it cannot be that Tännsjö denies the relevance of thought experiments that are logically or nomologically but not technically possible since he claims that the Repugnant Conclusion is "a mere logical possibility" but is "of the utmost relevance … to theoretical ethics".[13]

A possible explanation to why one would consider the welfare of presently existing people relevant for the acceptability of the Repugnant Conclusion could be that one believes that the reason people have found this conclusion unacceptable is that they consider populations with very low positive welfare "repugnant" or bad in themselves. If the people in the B-population in Diagram 3.1.1 enjoy the same welfare as people in the privileged parts of the world, then one might think that this population cannot be intrinsically bad or repugnant. This chain of reasoning seems to lie behind Christian Munthe's statement that "if the concrete living conditions in Z [a world consisting of lives barely worth living] are described like my current life, then my negative attitude toward that world would diminish considerably"[14] Presumably, Munthe's past negative attitude toward the Z-world was based on a belief that such a world is intrinsically bad.

This line of reasoning rests on a misunderstanding of what is primarily unacceptable about implications like the Repugnant Conclusion. The counter-intuitiveness of the Repugnant Conclusion doesn't essentially rest on categorical properties of populations with very low positive welfare, for example, that such populations are repugnant or bad in themselves. Indeed, that populations with very low positive welfare manifest such categorical properties is one possible explanation of our belief about the Repugnant Conclusion, but this is just one among many competing explanations. People who don't find such populations bad in themselves, which comprise the majority of the theorists in the field, still find the Repugnant Conclusion unacceptable. The unacceptability of the Repugnant

---

[13] Tännsjö (1998), p. 160. He doesn't give any arguments for these claims.
[14] Munthe (1992), p. 333.

Conclusion arises from the fact that any population with very high welfare is *worse* than some population with very low welfare. It is this *comparative aspect* of the Repugnant Conclusion that we find hard to accept.

It might be that Tännsjö has another argument in mind. At one point, he writes:

> The view that I am prepared to defend is somewhat pessimistic but still, I am afraid, realistic. My impression is that if only our basic needs are satisfied, then most of us are capable of living lives that, on balance, are worth experiencing. However, no matter how "lucky" we are, how many "gadgets" we happen to possess, we rarely reach beyond this level. If sometimes we do, this has little to do with material affluence; rather, bliss, when it does occur, seems to be ephemeral result of such things as requited love, successful creative attempts and, of course, the proper administration of drugs.[15]

Although he doesn't explicitly express it, perhaps Tännsjö's idea is that there are *no* possible lives with very high welfare. Would the truth of this matter make the Repugnant Conclusion acceptable? Yes, since it would neutralise the Repugnant Conclusion by making it an empty truth. If there are no possible lives with very high welfare, then the Repugnant Conclusion is vacuously true, since the antecedent – "[f]or any possible population … with a very high quality of life" – is false of every possible population. Consequently, if there are no possible lives with very high welfare, then Total Utilitarianism, which Tännsjö subscribes to, would only imply the Repugnant Conclusion in a trivial and uninteresting sense.

Is it plausible that there are no possible lives with very high welfare? Again, that the presently existing people in the privileged parts of the world have very low or negative welfare is at least conceivable, but it seems incredible that there are no logically possible lives with very high welfare (recall that Tännsjö thinks that "mere" logical possibilities are relevant).

Tännsjö is a hedonistic total utilitarian. The welfare of a life is determined by just summing the utilities of the happy and unhappy moments in the life.[16]

---

[15] Tännsjö (1998), p. 161.
[16] Tännsjö (1998), pp. 63-79.

Consider a population that consists of very short lives, say a minute of slight happiness. According to Tännsjö's theory, these lives enjoy positive welfare. It is hard to deny that such lives have very much lower welfare than the lives led in the privileged parts of the world. Irrespective of whether there are possible lives with very high welfare, Tännsjö's theory implies the following recasting of the Repugnant Conclusion: For any perfectly equal population with the same welfare as the people in the privileged parts of the world, there is a population of lives consisting of just one minute of slight happiness, which is better.[17]

3.1.3 Does Total Utilitarianism Imply the Repugnant Conclusion?

It is so easy to prove that Total Utilitarianism implies the Repugnant Conclusion that it might seem unnecessary. Let u_1 denote the numerical representation of some welfare level of people with very high positive welfare and let u_2 denote some welfare level of people with very low positive welfare. For any population of n people with very high welfare, there is a population of m people with very low positive welfare such that $nu_1 < mu_2$, namely a population consisting of $m > nu_1/u_2$ people with welfare u_2.

This proof implicitly invokes, however, an important assumption that, to the best of my knowledge, no one has ever formally noticed. One has to assume that for any value of n, there is a possible population with very low welfare u_2 *and* of size $m > nu_1/u_2$. It is easy to see how one could deny this: One could hold that there is a largest possible population, and that for any population with very low welfare, there is an equally large population with very high welfare. If this is true, then Total Utilitarianism doesn't imply the Repugnant Conclusion. Let's say that the limit of the size of a possible population is k. Then there is no population with very low positive welfare which is better than, for instance, a population with very high welfare and of size k.

Given that there are possible lives with very high and very low positive welfare, and that for any population with very low welfare, there is an equally large population with very high welfare, the following assumption is necessary for Total Utilitarianism to imply the Repugnant Conclusion:

---

[17] In section 10.3, we shall give an exact formulation of the Quality Condition which doesn't involve the concept "very high welfare".

> *The No-Limit Assumption*: For any possible population consisting of lives with a certain welfare, there is a larger possible population consisting of lives with the same welfare.[18]

Is this a reasonable assumption? It clearly does not involve any epistemologically problematic populations. Notice that the No-Limit Assumption does not imply that there are possible populations of infinite size.[19] Compare with the natural numbers: For any natural number n, there is a larger natural number $n + 1$, but, of course, every natural number is finite. Imagine a population of any finite size n. Do you have any problem imagining a larger population of size $n + 1$?

Does the No-Limit Assumption involve nomologically impossible populations? Probably not, since the latest verdict of modern physics is that the size of the cosmos is unbounded. Moreover, whether or not the cosmos has an upper boundary is dependent on contingent empirical facts, not on the laws of nature.[20]

---

[18] Strictly speaking, we only need to make the less general assumption that for any possible population with very low welfare, there is a larger possible population with the same very low welfare. If the No-Limit Assumption is problematic, however, then this assumption is problematic too.

[19] Recall that in section 2.3, we restricted the admissible populations, for epistemological reasons, to finite sets of lives. It should also be evident that the No-Limit Assumption doesn't involve any logically impossible populations since there is no greatest natural number.

[20] Those physicists who think that the universe will reach a maximum size and then contract back to a "Big Crunch" usually cite nomologically contingent features of the universe, such as its mass density, as responsible for this course of evolution. Since the physics of the Big Bang remains pretty mysterious to us, it of course remains possible that one day physicists will say that only a mass density high enough for collapse is consistent with Big Bang cosmology. It doesn't seem very likely, however. Recent astronomical observations suggest that the universe is unbounded, that is, there is not enough mass to cause a collapse so it will continue to expand indefinitely (see Glanz (1998)). However, given that the universe has a finite amount of energy, this expansion will continue to slow down until it is effectively zero (this is called the "heat death" of the universe; all energy eventually approaches zero). Again, I think that universes of arbitrarily high energy are nomologically possible, so we could still insist that even though any given universe will be finite in size, there is always a nomologically possible universe that is greater in size (owing to its greater energy). Indeed, some physicists have suggested that recent observations of the universe, indicating that its expansion is speeding up, demonstrate the need to reintroduce something Einstein called the "Cosmological Constant" or "lambda", a variable that represents the ability of space-time itself to create matter-energy which would account for the force that counteracts gravity (see Glanz 1998). If that is correct, then the universe could certainly grow in size indefinitely. I am grateful to Joshua Mozersky for explaining these intricate matters for me.

One might remark here that it would be very odd if the acceptability of Total Utilitarianism depended on some arcane and very speculative facts about the cosmos. I certainly agree and that is one of the reasons why I, in chapter 2, expressed scepticism about the restriction of test cases to only nomologically possible cases and emphasised an epistemological criterion instead.

As the attentive reader might have observed, the meaningfulness of the Quality Condition doesn't depend on the No-Limit Assumption. The condition that we shall introduce in the next section, however, implies the No-Limit Assumption.

3.2 The Quantity Condition

There are other axiologies apart from Total Utilitarianism that violate the Quality Condition and imply the Repugnant Conclusion. These can be characterised by a set of conditions. Consider the following condition:

> *The Quantity Condition*: For any pair of positive welfare levels 𝔸 and 𝔹, such that 𝔹 is slightly lower than 𝔸, and for any number of lives *n*, there is a greater number of lives *m*, such that a population of *m* people at level 𝔹 is at least as good as a population of *n* people at level 𝔸, other things being equal.

The Quantity Condition has some intuitive plausibility and should appeal to those thinkers that find some truth in the saying "the more good, the better". However, it implies the Repugnant Conclusion together with the following condition, which is, I believe, as uncontroversial as it gets in population axiology:

> *The Egalitarian Dominance Condition*: If population A is a perfectly equal population of the same size as population B, and every person in A has higher welfare than every person in B, then A is better than B, other things being equal.[21]

---

[21] Cowen (1996), pp. 774-5, sketches an impossibility theorem based on a condition – "the value pluralism axiom" – which *violates* the Egalitarian Dominance Condition. According to Cowen, this axiom implies "a maximum value for the social welfare that can result from very high levels of utility". Assume that a population A has reached this maximum value of utility, and that B is a perfectly equal population of the same size as A but with higher welfare. Assume further that

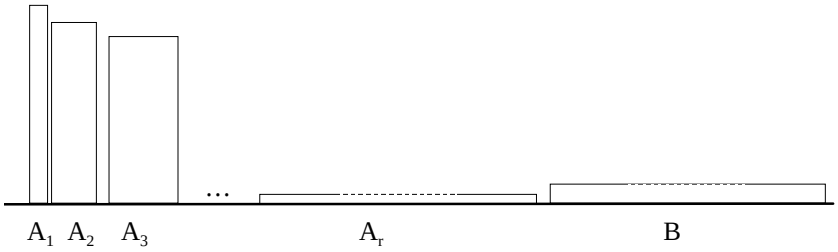

Diagram 3.2.1

In section 10.4, we shall formally prove that these two condition imply the negation of the Quality Condition. Here, we shall only give an intuitive demonstration that they imply the Repugnant Conclusion. Assume that A_1 in the diagram above is a population with very high welfare and that B is a population with very low positive welfare. According to the Quantity Condition, there is a population A_2 with slightly lower welfare than A_1 and which is at least as good as A_1; a population A_3 with slightly lower welfare than A_2 and which is at least as good as A_2; and so forth. Finally, we will reach population A_r with very low positive welfare. Assume that B is a population of the same size as A_r and with very low positive but slightly higher welfare than A_r. According to the Egalitarian Dominance Condition, B is better than A_r, and by transitivity, B is better than A_1. Since A_1 is an arbitrary population with very high welfare, this shows that for any population with very high welfare, there is a population with very low positive welfare which is better, that is, the Repugnant Conclusion. Consequently, any theory which satisfies the Egalitarian Dominance Condition and avoids the Repugnant Conclusion has to violate the Quantity Condition.[22]

I don't think, however, that the above demonstration amounts to a convincing impossibility theorem for an acceptable population axiology. Those who don't

---

these populations are equal in all other respects. Cowen's axiom implies, implausibly, that these two populations are equally good – a clear violation of the Egalitarian Dominance Condition. Needless to say, an impossibility theorem based on a condition which violates the Egalitarian Dominance Condition is not very convincing.

[22] The above demonstration assumes that the differences between any two welfare levels consists of a finite number of "slight welfare differences". Intuitively, this seems convincing but one may wonder how one should more exactly spell out this assumption. We shall clarify this issue in chapter 10.

accept the Repugnant Conclusion probably won't accept the Quantity Condition, and the explanations for the unacceptability of the former that we shall meet below, will also work for the latter. What the above demonstration shows, however, is that theories which imply that there is always some increase in the number of people with positive welfare that can outweigh a small decrease in individual welfare won't work.

3.3 Average Utilitarianism

The most popular theory among welfare economists, Average Utilitarianism, ranks populations according to the average welfare per life in the population. The value of a population is the sum of the welfare of all the lives in the population divided by the population size:

$$AU(X) = \frac{1}{n} \sum_{i=1}^{n} u_i = (u_1 + u_2 + \ldots + u_n)/n$$

As before, n is the population size of X and u_i is the numerical representation of the welfare of the i:th life in population X. Comparisons of average welfare presuppose that welfare is measurable on at least an interval scale.

Average and Total Utilitarianism are extensionally equivalent in same-number cases, that is, when the populations ranked are of the same size, these theories yield the same ranking. However, in different-number cases, that is, when the populations ranked involve populations of different sizes, they yield very different results. Whereas Total Utilitarianism is the paradigmatic context-insensitive theory, Average Utilitarianism is the typical context-sensitive theory. According to Average Utilitarianism, the contributive value of a life can vary in all respects: a life with positive welfare can have negative contributive value and a life with negative welfare can have positive contributive value.

Average Utilitarianism clearly avoids the Repugnant Conclusion and violates the Quantity Condition. Its violation of the latter condition has a quite vexatious character, however. It implies what we could call the Reversed Repugnant Conclusion:

The Reversed Repugnant Conclusion: For any population with very high positive welfare, there is a better population consisting of just one person with slightly higher welfare, other things being equal.

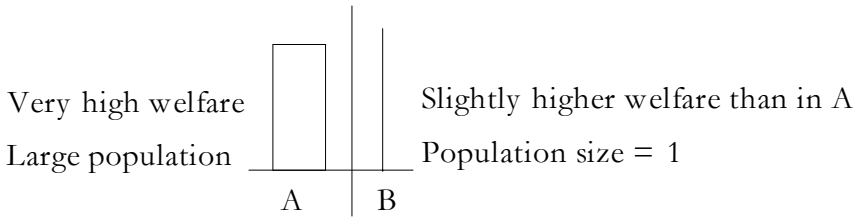

Diagram 3.3.1

According to Average Utilitarianism, it is worse if there is a lower average welfare and no value is put on the total amount of welfare. Consequently, an arbitrarily small increase in average welfare can outweigh an arbitrarily large decrease in the total sum of welfare.

Average Utilitarianism avoids the Repugnant Conclusion with a vengeance. Surprisingly however, it implies conclusions similar to the Repugnant Conclusion. Average Utilitarianism implies that for any population with very high welfare, it can be worse to add this population rather than a population with very low welfare.

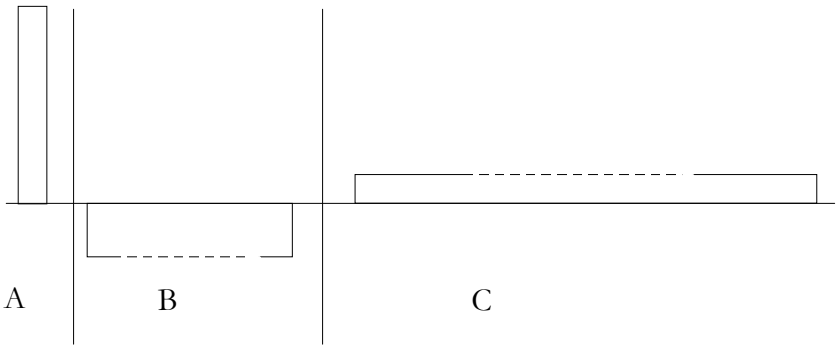

Diagram 3.3.2

Let A, $N(A)=k$, be any population with very high welfare and let v be the total sum of welfare in this population. For any value of v, there is population of n bad lives (population B in Diagram 3.3.2) such that the total sum of their welfare is $-w$, $w > v$. Likewise, for any value of w, there is a population of m lives with very low

positive welfare (population C in Diagram 3.3.2) such that the total sum of their welfare is z, $z > w$. If one adds A to B, then the average welfare would be $(v - w)/(k + n)$ which is less than zero, whereas if one adds C to B, then the average would be $(z - w)/(m + n)$ which is greater than zero. Consequently, according to Average Utilitarianism, it would be better to add the population with very low welfare rather than the population with very high welfare. Moreover, for any population with very high welfare, irrespective of its size, there are situations where it would be worse to add the population with very high welfare than a population with very low welfare. We have just shown that Average Utilitarianism violates the Quality Addition Principle:[23]

> *The Quality Addition Principle*: There is at least one perfectly equal population with very high welfare such that its addition to any population X is at least as good as an addition of any population with very low positive welfare to X, other things being equal.

The problem of finding an acceptable population axiology has often been conceived of as a problem of finding the right weighing of average and total welfare or, as it is sometimes expressed, between quality and quantity of welfare.[24] Thus, cases with increases in both the average and the total welfare have been seen as unproblematic. What Average Utilitarianism's violation of the Quality Addition Principle shows, I think, is that this way of framing the problem is unsatisfactory.

Those who find the Repugnant Conclusion worrisome and endorse the Quality Condition, would probably hold that a satisfactory population axiology should satisfy the Quality Addition Principle. Although we are inclined to agree, we shall adopt a weaker version of this principle as an adequacy condition:

---

[23] As we shall see in section 8.5, Average Utilitarianism violates the Quality Addition Principle also when all the involved lives enjoy positive welfare.

[24] See, for example, Parfit (1984), pp. 401-3, and Carlson (1998a). Carlson writes on p. 295 that the "problem is … to find a theory that strikes the right balance between quality and quantity". He later goes on to reject this claim in regard to cases that only involve negative welfare. See p. 298.

> *The Weak Quality Addition Condition*: For any population X, there is at least one perfectly equal population with very high welfare such that its addition to X is at least as good as an addition of any population with very low positive welfare to X, other things being equal.

This condition is implied by the Quality Addition Principle but it is a weaker condition: Although Average Utilitarianism violates the Quality Addition Principle, it satisfies the above condition. Let u_1 be the highest welfare level among lives with very low positive welfare and let u_2 be some very high welfare level. Consider first all populations X with average welfare greater than u_2. For such populations, any addition of lives with very low positive welfare decreases the average welfare more than an addition of one person with very high welfare u_2. Next, consider all populations X with average welfare greater than u_1 but less than or equal to u_2. For such populations, an addition of lives with very high welfare u_2 will not decrease the average whereas an addition of lives with very low positive welfare decreases the average. Lastly, consider all populations X with average welfare less than or equal to u_1. For any such X, there is an addition of lives with very high welfare which increases the average above u_1 whereas there is no addition of lives with very low welfare which increases the average above u_1. Consequently, Average Utilitarianism satisfies the Weak Quality Addition Condition.

Total Utilitarianism, on the other hand, violates this condition. We can see the Weak Quality Addition Condition as an extension of the Quality Condition. The former condition implies the latter since it implies that there is at least one perfectly equal population with very high welfare such that its addition to an empty population is at least as good as an addition of any population with very low positive welfare to an empty population. And I think that we should endorse the Weak Quality Addition Condition for similar reasons that we endorsed the Quality Condition: It would be repugnant if for every population with very high welfare, there was a population with very low positive welfare whose addition would be better. It is hard to see how one could endorse the Quality Condition and reject the Weak Quality Addition Condition.[25]

---

[25] In section 7.3, we shall deal with a possible (but not convincing) egalitarian objection to the Weak Quality Addition Condition.

Average Utilitarianism satisfies the Weak Quality Addition Condition but violates the Quality Addition Principle and implies the Reversed Repugnant Conclusion. There are other decisive reasons for rejecting this theory and we shall shortly take a look at two of them. Let us first, however, turn to a more promising theory.

4

Variable Value Principles

4.1 Introduction

Yew-Kwang Ng and Theodore Sider have proposed theories along the lines of Tom Hurka's idea of Variable Value Principles.[1] These principles are sometimes called "compromise theories" since a Variable Value Principle can be said to be a compromise between Total and Average Utilitarianism. With small populations enjoying high welfare, a Variable Value Principle behaves like Total Utilitarianism and assigns most of the value to the total sum of welfare.[2] For large populations with low welfare, the principle mimics Average Utilitarianism and assigns most of the value to average welfare.

In Diagram 4.1.1 below, the vertical axis indicates the value and the horizontal axis indicates the size of a population. The graphs show how the value of populations, according to Average Utilitarianism, Total Utilitarianism, and a Variable Value Principle, varies with population size given a fixed welfare level of the populations (*m*, 2*m*, and so forth). As we can see in the diagram, Variable Value Principles assign asymptotically increasing value to the total sum of welfare and linearly increasing value to the average welfare. If one keeps the average welfare fixed and increases the population size, then the value converges on a limit asymptotically: A doubling of the population size without any increase in the average welfare always results in less than a doubling of the value. A doubling of

---

[1] Hurka (1983), Ng (1989), Sider (1991). Parfit (1984), p. 402, mentions a Variable Value Principle but disregards it. This is because he thinks that such principles applied to large population sizes would amount to the same thing as theories which assign linear increasing value to the sum of welfare but put an upper limit to this value.

[2] Hurka (1983), p. 497, argues that with small populations, the contributing value of extra people should be greater than the mere sum of their welfare to allow for the possibility that the contributing value can outweigh the lowering of the total amount of welfare for the sake of population growth. Excluding the possibility that Hurka assigns intrinsic value to population growth as such, his argument seems to rest on a conflation of intrinsic and instrumental value.

the average welfare, on the other hand, always doubles the value of the population, which is reflected in the even spacing of the asymptotes in the diagram.

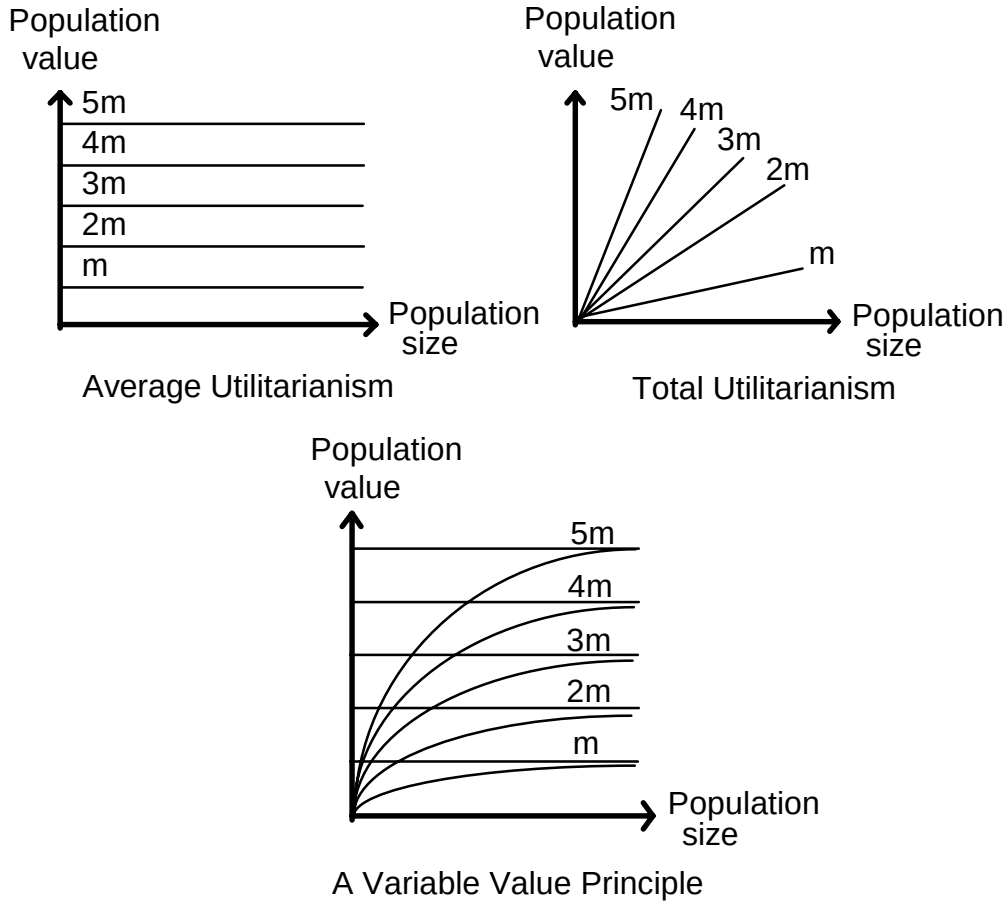

Diagram 4.1.1 Three Population Principles (after Hurka 1983)

4.2 Ng's Theory X'

Ng's Variable Value Principle, theory X', dampens the increase of the linear function *n*, the population size, by transformation with a concave function f(*n*). Whereas Average Utilitarianism ranks populations according to the average welfare Q, and Total Utilitarianism according to the total welfare *n*Q, theory X' ranks them according to f(*n*)Q. Ng's concave function looks like this:

$$f(\mathrm{n}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{k}^{\mathrm{i}-1} = \mathrm{k}^{0} + \mathrm{k}^{1} + \mathrm{k}^{2} \ldots \mathrm{k}^{\mathrm{n}-1} \quad 1 > \mathrm{k} > 0$$

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*) asymptotically approaches 1/(1-*k*), which is of finite value. This means that with large populations, the value yielded by the function *f(n)Q* is not increased when the average welfare is decreased but the total welfare is increased by an addition of more people. With large populations, *f(n)Q* approaches *mQ* where *m* is a constant. Consequently, theory X' behaves like Average Utilitarianism with large populations and thereby satisfies the Quality Condition and avoids the Repugnant Conclusion.

Since theory X' mimics Average Utilitarianism in cases that involve large populations, it shares a number of properties with this theory. Like Average Utilitarianism, the contributive value of a life can vary in all respects: A life with positive welfare can have negative contributive value and a life with negative welfare can have positive contributive value. Both theory X' and Average Utilitarianism violate a principle suggested by many theorists:

> *The Mere Addition Principle*: An addition of people with positive welfare does not make a population worse, other things being equal.[3]

Since any addition of people with welfare below the average makes a population worse according to Average Utilitarianism, this also holds for additions of people with positive welfare below the average. Informally, we can show that Ng's theory is not compatible with the Mere Addition Principle with the following diagram:

[3] Cf. Hudson (1987), Ng (1989), and Sider (1991). Cf. fn. 6 below.

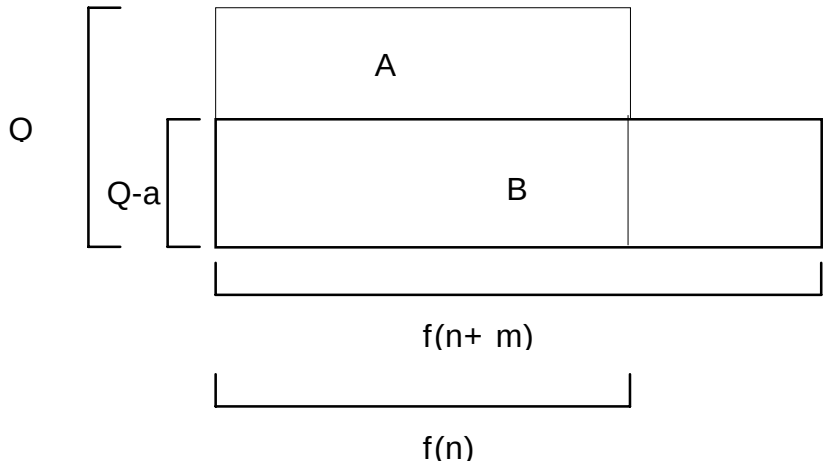

Diagram 4.2.1[4]

In Diagram 4.2.1, 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 f(*n*)*Q* and the value of B is f(*n*+*m*)(*Q*-*a*).

The difference between population A and B is that in B, *m* persons with positive welfare have been added to population A. These added people have a welfare that is below the welfare of the A-people. Hence, they lower the average by *a* units. In Diagram 4.2.1, 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 addition of *m* persons with positive welfare makes population B worse than population A.

Here is an algebraic demonstration. Let A be a population of *n* people with positive welfare *u*. The value of this population according to Ng's theory is $f(n)nu/n = f(n)u$. Let B consist of the A-people and *n* extra persons with positive welfare $v < u$. The value of population B is $f(2n)(nu+nv)/2n = f(2n)(u+v)/2$. Thus the value of population B is less than that of population A if $f(2n)(u+v)/2 < f(n)u$. This will be true if $v < [2f(n)u/f(2n)]-u$. Since *f* is a concave function, $2f(n) > f(2n)$ and $[2f(n)u/f(2n)]-u > 0$. In other words, for any choice of value for the weighing coefficient *k*, and for any positive welfare level *u*, there is a positive welfare level *v*

[4] I owe this diagram to Krister Bykvist.

$< u$ such that an addition of n people with welfare v to a population of n people with welfare u makes the resulting population worse than the original one.

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 one), then the Mere Addition Principle can be preserved for more compelling cases, "cases where the average utility of the added people is not very much lower than those of the pre-existing people, and the number of pre-existing people has not become very large".[5] As the above algebraic proof shows, this is questionable. Let's say that A is a population consisting of only one person with very low positive welfare. For any choice of value of k, there is an addition of one person with lower but positive welfare which would make the resulting population worse than A. And it seems counter-intuitive to claim that there are lives with positive welfare with *much lower* welfare than a life with very low positive welfare. Moreover, Ng's principle yields controversial violations of the Mere Addition Principle when the population is sufficiently large or when the added people's positive welfare is sufficiently low. For example, if $k = 0.99$, then an addition of one billion people to a population of one billion would make the outcome worse, even if the welfare of the added people was as high as 75 percent of the original people.

The Mere Addition Principle might seem compelling but it is controversial – several authors have rejected it. Moreover, one can easily show that the Mere Addition Principle in conjunction with a weak egalitarian condition implies the Repugnant Conclusion (we shall return to this topic in section 6.1 and 10.5, but see also the principles at work in the informal paradox in the introduction). Ng suggests that those who don't accept the Repugnant Conclusion should drop the Mere Addition Principle and Blackorby, Bossert and Donaldson argue similarly that if we have to choose between the Repugnant Conclusion and the Mere Addition

---

[5] Ng (1989), p. 249. This will have as a consequence that theory X' behaves more like Total Utilitarianism even with large populations and yields conclusions analogous to the Repugnant Conclusion. Another way to proceed is to use a function which is more curved towards the end as compared to the beginning. This could reflect an intuition that the value of the total sum of welfare starts to decrease at a certain level. This could be achieved by combining Ng's function with Total Utilitarianism: Let the value of the total sum of welfare 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 total sum of welfare is below the limit. However, it would not satisfy the Mere Addition Principle with large populations and consequently share the problems of theory X' discussed below.

Principle, then the latter must be rejected. Fehige holds that "it's intrinsically wrong to bring people into existence who will have at least one unfulfilled preference", and Parfit thinks that "if the extra people in A+ [a population consisting of the A-people with very high welfare plus the extra people] have lives that are only just worth living, most people find it easy to believe that A+ would be worse than A".[6] Consequently, we shall not adopt the Mere Addition Principle as an adequacy condition.

There is, however, a vexatious relationship between violations of the Mere Addition Principle and a counter-intuitive conclusion. Theories which yield that one can make a population worse by adding people with positive welfare tend to imply that an addition of a few lives with negative welfare decreases the value of a population less than an addition of many lives with positive welfare. Such theories imply what I call the Sadistic Conclusion: It can be *better* to add people with *negative* rather than *positive* welfare, other things being equal. Average Utilitarianism clearly implies this conclusion. Let's say that to a population consisting of one person with welfare 11*u* we can add either nine lives with positive welfare 1*u* or one life with negative welfare -3*u*. The value of the former population, according to Average Utilitarianism, is $(11u + 9u)/10 = 2u$, whereas the value of the latter population is $(11u - 3u)/2 = 4u$. Hence, according to Average Utilitarianism, it is better to add the life with negative welfare than the lives with positive welfare.

Ng's theory yields analogous results. For example, let $k = 0.9$. Assume that we can either add two persons with welfare +1 or one person with welfare -1 to a population consisting of one person with 100 units of welfare. According to Ng's theory, the value of the former population is approximately 92 whereas the value of the latter populations is approximately 94. Consequently, it would be better to add the unhappy life rather than the two happy lives. With large populations, where $f(n)$ is close to its limit and theory X' resembles Average Utilitarianism, Ng's theory

---

[6] Ng (1989), p. 244; Blackorby, Bossert and Donaldson (1995), p. 1305, and (1997), pp. 210-1; Fehige (1998). Ng ascribes to Parfit the view that a population axiology should satisfy the Mere Addition Principle (Ng (1989), p. 238) and one might get that impression from Parfit (1984), pp. 420ff. In personal communication, however, Parfit has expressed doubts about the Mere Addition Principle in cases where the added people are much worse off than the rest of the population, as is indicated in the quote above from his referee report on Arrhenius (2000a). See also Feldman (1997), ch. 10, Kavka (1982), and Carlson (1998a), pp. 288-9. We shall discuss Blackorby et al.'s, Fehige's, and Feldman's views below.

implies highly counter-intuitive implications of this kind for any choice of *k*. By adding many people with very high but *slightly* lower welfare than the original people, the average welfare can decrease more than when adding a few people with *very* negative welfare. In other words, theory X' implies that the addition of the people with very negative welfare would be *better* than the addition of the people with very high welfare. Both Average Utilitarianism and theory X' violate the following reasonable condition:

> *The Non-Sadism Condition*: An addition of any number of people with positive welfare is at least as good as an addition of any number of people with negative welfare, other things being equal.

Erik Carlson has, however, suggested that it sometimes can be better to add people with negative welfare rather than people with positive welfare:

> … [I]t does not seem unreasonable to claim that A+Z is worse than a population, call it A+-1, consisting of the A-people plus one person at a welfare level just below zero. Consider that A+Z is an enormous population where *the vast majority have lives barely worth living*, whereas A+-1 is a large population where *almost everyone has a very high quality of life*, and where there is no great unhappiness.[7]

I have the unsettling feeling that Carlson's argument turns on the tendency to consider lives barely worth living bad for the people living them, which they are not (compare with the discussion of how to understand the Repugnant Conclusion in section 3.1.2). And it still sounds counter-intuitive to me that it could be better to increase the number of unhappy lives than to increase the number of happy lives.[8] But perhaps Carlson has showed us that we shall be suspicious of intuitions of such general nature and that there are reasons to resist the Non-Sadism Condition in

---

[7] Carlson (1998a), p. 302, emphasis in original.

[8] It seems that Carlson finds the corresponding normative version of the Non-Sadism Condition more convincing since he thinks that considering the Sadistic Conclusion (SC) strictly from an axiological perspective "removes some of the repulsiveness of SC". In section 11.5, we shall formulate a normative version of the Non-Sadism Condition.

some particular cases. At any rate, this wouldn't be of much comfort for the average utilitarian or Ng, since, as we saw above, their theories implies very troublesome violations of the Non-Sadism Condition. These theories violate the following logically weaker and, I surmise, unassailable version of the Non-Sadism Condition:

> *The Weak Non-Sadism Condition*: There is a negative welfare level and a number of lives at this level such that an addition of any number of people with positive welfare is at least as good as an addition of the lives with negative welfare, other things being equal.

Let n be any number of lives, let u_1 represent any negative welfare level, and let u_2, u_3 and u_4 represent any three positive welfare levels such that $u_2 < u_3 < u_4$. Now, for any n and u_1, there is an m such that $(nu_1+mu_4)/(n+m) > u_3$, namely $m > n(u_3-u_1)/(u_4-u_3)$, and for any m and u_4, there is a k such that $(mu_4+ku_2)/(m+k) < u_3$, namely $k > m(u_4-u_3)/(u_3-u_2)$. Consequently, Average Utilitarianism implies that for any population with negative welfare, there is a situation where it would be better to add this population rather than a population with positive welfare. Since u_1- u_1 can be any welfare levels fitting the description above, this implication of Average Utilitarianism holds true even if the involved populations have very negative and very high positive welfare. When $f(n)$ is close to its limit and theory X' resembles Average Utilitarianism, it implies this conclusion too.

We have noticed several similarities between Ng's theory X' and Average Utilitarianism. Not surprisingly then, theory X', like Average Utilitarianism, violates the Quality Addition Principle. Again, let A, $N(A)=k$, be any population with very high welfare and let v be the total sum of welfare in this population (cf. Diagram 3.3.2). For any value of v, there is population of n bad lives such that the total sum of their welfare is $-w$, $w > v$. Likewise, for any value of w, there is a population of m lives with very low positive welfare such that the total sum of their welfare is z, $z > w$. According to Ng's theory X', the value of A∪B is $f(k+n)(v-w)/(k+n)$, which is less than zero since $v-w < 0$, whereas the value of C∪B is $f(m+n)(z-w)/(m+n)$, which is greater than zero since $z-w > 0$. Consequently, theory X' yields that for any population with very high welfare, there is a situation where it would be worse to add this population rather than a population with very low welfare. This implication

of Ng's theory is especially noteworthy, since the *raison d'être* of his theory is to avoid repugnant conclusions.

Average Utilitarianism and Ng's principle also have counter-intuitive consequences when applied to populations with general negative welfare. An uncontroversial condition of acceptability is the negative counterpart of the Mere Addition Principle:

> *The Negative Mere Addition Principle*: An addition of people with negative welfare makes a population worse, other things being equal.

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 and "[n]o matter how great is the disutility, it can always be compensated by a sufficiently big amount of utility".[9]

Assume that the average welfare of the A-people is negative in Diagram 4.2.1, that Q is less than zero. In B, we have added persons who will be better off but still have negative welfare. 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 *additional* unhappy people.

The algebraic demonstration of this implication of Ng's theory mirrors the demonstration above of the violation of the Mere Addition Principle. Let A be a population of n people with negative welfare $-u$ and let B consist of the A-people and n extra persons with negative welfare $-v > -u$, that is, the added people have negative welfare but are better off than the A-people. The value of population B is greater than that of population A if $f(2n)(-u-v)/2 > -uf(n)$. This will be true if $-v > [-2f(n)u/f(2n)]+u$. Since f is a concave function, $2f(n) > f(2n)$ and $[-2f(n)u/f(2n)]+u < 0$. In other words, for any choice of value for the weighing coefficient k, and for any negative welfare level $-u$, there is a negative welfare level $-v > -u$ such that an addition of n people with welfare $-v$ to a population of n people with welfare $-u$

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

makes the resulting population *better* than the original one. Consequently, theory X' violates the very compelling Negative Mere Addition Principle.[10] We leave the demonstration of Average Utilitarianism's violation of this principle as an exercise for the reader.

Average Utilitarianism and Theory X' share the feature of giving 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.

4.3 Sider's Principle GV

A second way of constructing a Variable Value Principle is to dampen each person's contributing value. Sider has proposed a theory of this kind:[11]

> Divide the individual welfare profiles of a population into two ordered sets:
>
> (i) $(u_1 . . . u_i . . . u_n)$ - the welfare profiles of the people with positive or zero welfare, in order of *descending welfare* – in case of ties, any order for those tied will suffice.
>
> (ii) $(v_1 . . . v_j . . . v_m)$ - the welfare profiles of the people with negative welfare, in order of *ascending welfare*.

$$GV(X) = \sum_{i=1}^{n} u_i \, k^{i-1} + \sum_{j=1}^{m} v_j \, k^{j-1} \quad 1>k>0$$

---

[10] Ng claims that, disregarding the Mere Addition Principle, theory X' meets all of Parfit's requirements on a population axiology and may be exactly the theory he is after (Ng (1989), p. 245). That is doubtful. Parfit rejects Average Utilitarianism exactly on the ground that it doesn't give enough weight to negative welfare, referring to an example similar to the one used above. Parfit (1984), p. 422, 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". In cases like these involving large populations, theory X' and Average Utilitarianism yield the same result.

[11] See Sider (1991).

Sider's principle first divides a population into two ordered sets: one set with the welfare profiles of the people with positive welfare, in order of *descending* welfare; and another set with the welfare profiles of the people with negative welfare, in order of *ascending* welfare. Sider's principle dampens the value of the welfare of different people to different degrees depending on their place in the orderings of the positive and negative welfare profiles. The higher a person's positive welfare relative to the welfare of others, the less dampening of the value of this person's welfare will take place and, consequently, the more she will contribute to the value of the population. The value of the person with the highest welfare will not be dampened at all. The more negative a person's welfare is relative to the welfare of others, the less dampening of the disvalue of this person's welfare will take place and, consequently, the more she will detract from the value of the population. The disvalue of the person with the most negative welfare will not be dampened at all.

Principle GV satisfies the Quality Condition and avoids the Repugnant Conclusion by being a convergent sum. When there is perfect equality, GV approaches $Q/(1-k)$ which is of finite value; that is, applied to large population sizes, principle GV mimics Average Utilitarianism. With small populations, on the other hand, principle GV mimics Total Utilitarianism.

Like Average Utilitarianism and Ng's theory X', Sider's theory is a context-sensitive theory – the contributive value of a life is dependent on the welfare of the rest of the population – but it differs from the former theories in an important and interesting way: The contributive value of lives with positive welfare is always positive, and the opposite for lives with negative welfare. Consequently, as Sider has shown, this principle doesn't violate the Mere Addition Principle.[12] Let's say that we can add to a population with the welfare profile $(u_1 \ldots u_i \ldots u_n)$, a life with positive utility z, $u_i \geq z \geq u_{i+1}$. Consequently, z will be inserted into the summing sequence between u_i and u_{i+1}. The summing sequence will be the same up to u_i both with and without the extra life. Then we have (the terms representing the welfare of the lives in the new population are on the left hand side):

[12] For a formal proof, see Sider (1991).

$$\xi k^{i+1} \geq u_{i+1} k^{i+1}$$
$$u_{i+1} k^{i+2} \geq u_{i+2} k^{i+2}$$

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, sum A must be greater than sum B.

While this principle may seem promising, it is nevertheless flawed. Suppose that with a population of *m* people enjoying very high welfare, the contributing value of extra people with positive but lower welfare is approximately zero.[13]

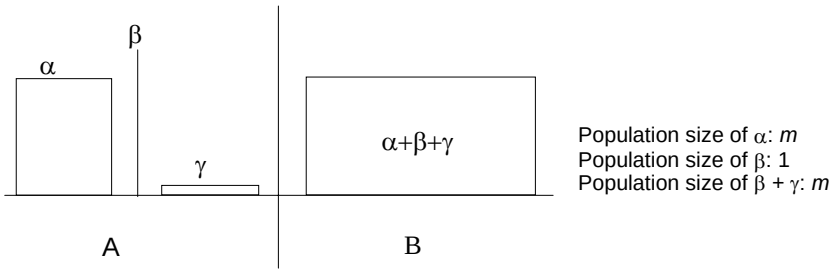

Diagram 4.3.1

In the diagram above, everybody enjoys the same welfare except the γ-people in A who are much worse off and the β-person in A who is slightly better off. 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. In alternative A, β's welfare will not be dampened at all but the welfare of the γ-people will be strongly dampened in both alternatives, that is, the contributive value of the γ-people will be close to zero in both populations in spite of the fact that the γ-people have much higher welfare in population B. Consequently, the small gain for the β-person outweighs the great loss of the γ-people. This evaluation is very anti-egalitarian. Principle GV violates the following plausible principle:

[13] That is, $u_{n+1} k^n \approx 0$, where $u_{n+1} \leq$ very high welfare.

> *The Non-Anti Egalitarianism Principle*: A population with perfect equality is better than a population with the same number of people, inequality, and lower average (and thus lower total) welfare.[14]

Indeed, principle GV's violation of the Non-Anti Egalitarianism Principle is especially serious. It implies the following conclusion:

> *The Very Anti Egalitarian Conclusion*: For any perfectly equal population of at least two persons with positive welfare, there is a population which has the same number of people, lower average (and thus lower total) welfare and inequality, which is better.

Compare the following populations A and B. A contains two persons with welfare $u > 0$. B contains one person with welfare $u+x$ and another person with welfare $u-z >0$, $0 < x < z$. Consequently, there is perfect equality in A as well as a higher total of welfare as compared to B. The values of the two populations according to Sider's principle GV are as follows:

$$GV(\text{A}) = uk^0+uk^1 = u+uk$$
$$GV(\text{B}) = (u+x)k^0+(u-z)k^1 = u+x+uk-zk$$

The difference in population value between B and A is thus $u+x+uk-zk-u-uk = x-zk$. Now, for any k, $1 > k > 0$, there is an x and z such that $zk < x < z$, that is, we can always construct a population B that has higher population value than population A although B is more unequal and has less total welfare. This result can easily be generalised to any perfectly equal population with at least two persons with positive welfare. For example, one can always subject two persons in such a population to the same process as above.[15]

---

[14] See Ng (1989), p. 238. Ng's principle includes a condition to the effect that there is "the same set of individuals" in both outcomes. In his discussion of the principle, however, he appeals to cases where the compared populations consist of different individuals. See especially p. 239, fn. 4.

[15] In fact, Sider doesn't advocate GV because "it generates rather extreme results with respect to distributive justice". See Sider (1991), p. 270, fn. 10.

Since the contributive value of lives with positive welfare is always positive, and the opposite for lives with negative welfare, Sider's theory satisfies the Non-Sadism Condition. It implies, however, conclusions analogous to the Sadistic Conclusion. Assume that the world is crowded by lots of people, all living in the same hell full of illness and pain. Let us ponder whether to add a large number of people. One of these added people will enjoy low positive welfare. The other ones will have the kind of hellish life which is commonplace in this world. Since the number of unhappy lives is already large, the negative contributive value of the extra unhappy lives will be very small - the weight assigned to their life will be very small. The extra happy life will be the only happy life in this world and therefore must be assigned the weight one. Consequently, the negative contributive value of the extra unhappy lives will be outweighed by the positive contributive value of the life with very low positive welfare. According to Sider's principle, it is better to add the life with very low positive welfare and all the hellish lives rather than to refrain from adding them.[16]

[16] This argument was inspired by an analogous argument suggested by Krister Bykvist.

5

Critical Level Theories

5.1 Blackorby, Bossert and Donaldson's Critical-Level Utilitarianism

Blackorby, Bossert and Donaldson's Critical-Level Utilitarianism (CLU) is, in its simplest form, a modified version of Total Utilitarianism.[1] The contributive value of a person's life is her welfare minus a positive critical level. The value of a population is calculated by summing these differences for all individuals in the population. Principle CLU could thus be written in the following form:

$$CLU(X) = \begin{cases} \sum_{i=1}^{n} (u_i - k) & n > 0 \\ 0 & n = 0 \end{cases}$$

In the above formula, n is the population size of X and u_i is the numerical representation of the welfare of the i:th life in population X, and k is the critical level. Blackorby, Bossert and Donaldson assume a positive critical level, that is, the contributive value of lives with positive welfare below the critical level is negative. Consequently, assuming that the critical level is higher than very low welfare, the Repugnant Conclusion is deflected and the Quality Condition is satisfied since the

[1] See Blackorby, Bossert and Donaldson (1997, 1995) and Blackorby and Donaldson (1984). These authors also propose a more refined version of CLU where the contributive value of people's welfare is dampened by a strictly concave function. This modification has no relevance for the arguments made here. Another version of CLU introduces incommensurability among populations and might thus avoid some of the implications pointed out below. We shall discuss incommensurability below.

value of a huge population with positive but very low welfare will be negative. It is easy to see, however, that CLU violates the Non-Sadism Condition:

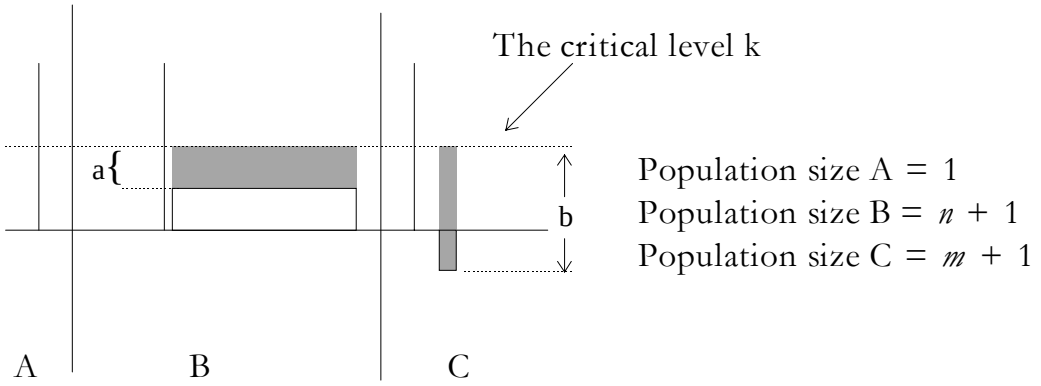

Diagram 5.1.1

In the above diagram, the width of each block shows the number of people, and the height shows their welfare. Outcome A consists of one person with welfare well above the critical level. In outcome B, we have added n people with positive welfare x. Their welfare is a units below the critical level k, as indicated in the diagram. The negative value of this addition is thus $n(x-k) = -na$ which is represented by the grey area in outcome B. In outcome C, m people with negative welfare y have been added. Their welfare is b units below the critical level, as indicated in the diagram. The negative value of this addition is $m(-y-k) = -mb$ which is represented by the grey area in outcome C. Since $mb < na$ (the grey area in outcome C is smaller than the grey area in outcome B), it is better to add the people with negative welfare rather than the people with positive welfare, a clear violation of the Non-Sadism Condition.

CLU implies especially troublesome violations of the Non-Sadism Condition:

> *The Very Sadistic Conclusion*: For any population of lives with negative welfare, there is a population of lives with positive welfare which is *worse*, other things being equal.

There is always a population with sufficiently many people with positive welfare slightly below the critical level such that the total negative value of these people is greater than that of a given population made up of people with negative welfare.

This holds irrespective of how much people suffer and of how many they are. Thus, CLU implies the Very Sadistic Conclusion and violates the Weak Non-Sadism Condition. Since CLU assigns negative contributive value not only to people with negative welfare but also to people with positive welfare, CLU gives less relative weight to negative welfare than Total Utilitarianism.

One of the main problems in finding an acceptable population axiology is to find the right weighting of individual and total welfare.[2] Thus, situations where both the welfare of each individual and the total welfare are increased are unproblematic, and especially so if the increase in welfare results in a perfectly equal population. Consider the following compelling principle:

> *The Extended Egalitarian Dominance Principle*: If population A is a perfectly equal population of greater size than population B, and every person in A has higher positive welfare than every person in B, then A is better than B, other things being equal.[3]

But assume that everybody in a population B has positive welfare below the critical level. In another population A, a number of people with *higher* welfare but below the critical level are added and the welfare of all the other members of the population is raised to the same level as the added people. The average and the total welfare is thus higher in A as compared to B, and there is perfect equality in A – every person in A is better off than any person in B.[4] And A could be worse than B according to CLU, since the negative value of the added people could outweigh the value of the increase in the B-people's welfare.

5.1.1 Incomplete Critical-Level Utilitarianism

A problem for Critical-Level Utilitarianism is how to find an intuitively acceptable critical level. It has to be high enough to avoid the Repugnant Conclusion, and low

---

[2] Cf. our discussion of the Quality Addition Principle in ch. 3.

[3] Cf. the Egalitarian Dominance Condition introduced in section 3.2, which only covers cases involving same sized populations.

[4] An example could be that in population A, people remain childless and in B they have offsprings whose positive welfare is higher than the people's welfare in B. The existence of these children also have a positive effect on the parents' welfare.

enough not to rule out additions of clearly good lives. Blackorby et al. acknowledge this and related problems and suggest an interesting solution.[5] Instead of using one critical level, they propose an interval of critical levels when comparing populations of different size. The interval of critical levels is assumed to be between zero and a positive welfare level α. The idea is that a population A is better than another population B if and only if A is better than B for all critical levels in the interval. Otherwise they are incommensurable, that is, A is neither at least as good as B, nor worse than B. If two populations are of the same size, then they are ranked by Total Utilitarianism. They call this principle Incomplete Critical-Level Utilitarianism:[6]

Incomplete Critical-Level Utilitarianism

(i) If population A and B are of the same size, then A is better than B if and only if $TU(A) > TU(B)$.

(ii) If population A and B are of different size, then A is better than B if and only if $CLU(A) > CLU(B)$ for all k, $0 \leq k \leq \alpha$, where α is the upper bound of the critical interval.

As Blackorby et al. point out, Incomplete Critical-Level Utilitarianism avoids the Repugnant Conclusion and the Sadistic Conclusion.[7] It does this in a questionable manner, however, since it does this by rendering the populations involved incommensurable. For example, let's say that A is a large population with very high welfare and total welfare x and that B, C, D … are populations with very low welfare and with total welfare greater than x. Assume that very low welfare is below the maximal critical level α. If $k = 0$, then Critical-Level Utilitarianism is equivalent with Total Utilitarianism and, consequently, $CLU(A) < CLU(B)$, $CLU(A) < CLU(C)$, and so forth. If $k = \alpha$, on the other hand, then the value of populations B, C, D, … are going to be negative whereas the value of A is going to

---

[5] See Blackorby et al. (1997), pp. 216-9. Variable Value Theories are afflicted by analogous problems regarding how to determine the weighing coefficient in the value functions. Cf. our discussion of Parfit's formulation of the Repugnant Conclusion in ch. 3.

[6] See Blackorby et al. (1997), pp. 216-9 and 226. That the critical levels consists of all numbers between zero and a positive welfare level is not part of Blackorby et al.'s definition of Incomplete Critical-Level Utilitarianism but they assume this in their discussion of it.

[7] See Blackorby et al. (1997), pp. 218-9 and 226.

be positive. Thus, Incomplete Critical-Level Utilitarianism renders all populations B, C, D, …with very low welfare and with total welfare greater than x as incommensurable with A. This is hardly intuitive. Rather, as we pointed out in chapter 3, most people find such populations clearly worse than a large population with very high welfare.

The incommensurability resorted to by Incomplete Critical-Level Utilitarianism in cases involving lives with negative welfare is even more counter-intuitive. For any number n of hellish lives, there is a number $m > n$ of lives with positive welfare just below the highest critical level, such that a population consisting of the hellish lives is incommensurable with the population consisting of the lives with positive welfare. Thus, Incomplete Critical-Level Utilitarianism avoids the Very Sadistic Conclusion but, again, in a disputable manner.

Incomplete Critical-Level Utilitarianism could resort to extensive incommensurability among populations but, I surmise, among the wrong kind of populations. Furthermore, although Incomplete Critical-Level Utilitarianism can avoid the Repugnant Conclusion and the Sadistic Conclusion, it cannot avoid violating the Quality Condition and the Non-Sadism Condition. According to the former condition, there is at least one perfectly equal population with very high welfare which is at least as good as all populations with very low welfare, other things being equal. Incomplete Critical-Level Utilitarianism implies that for any population with very high welfare, there is a population with very low positive welfare which is incommensurable with or better than the former. The Non-Sadism Condition yields that an addition of people with positive welfare is at least as good as an addition of people with negative welfare, other things being equal. According to Incomplete Critical-Level Utilitarianism, for any addition of lives with negative welfare, there is an addition of lives with positive welfare which renders the compared populations incommensurable. One can easily show that Incomplete Critical-Level Utilitarianism also violates the Weak Quality Addition Condition and the Extended Egalitarian Dominance Principle.

Let me conclude this section with some general remarks about incommensurability among populations. Such incommensurability is pretty plausible, I think, if there are other considerations apart from welfarist ones that are relevant for the evaluation of populations. If some kind of pluralism is true and there are other values than welfare, then it wouldn't be remarkable if some populations turn out to be incommensurable. For example, it might be that both

liberty (of some kind) and welfare should count but that there is no method of weighing gains in welfare against losses in liberty and vice versa. If one population is better than another population in respect to welfare but the other is better in respect to liberty, then these two populations would be incommensurable if the above pluralism were true.

It's important to remember, however, that we are discussing cases where other things are equal: Roughly, the populations that we are comparing only differ in respect to the welfare levels of their constituent lives and size. Moreover, Incomplete Critical-Level Utilitarianism is a welfarist principle. In general, for an appeal to incommensurability to have any credibility as an argument against the adequacy condition we have proposed, and in particular for welfarists such as Blackorby et al., one must produce a good *welfarist* reason for incommensurability.

There are, I think, three plausible sources of incommensurability among populations which are relevant in respect to the adequacy conditions that we are proposing.[8] One of these concerns a condition that we have not yet introduced, so we shall postpone our discussion of it until later. The first apparent source of incommensurability from a welfarist perspective has to do with comparisons of different people's welfare: One can reject interpersonal comparability of welfare. This move certainly yields extensive incommensurability among populations, but it would be, I surmise, too extensive to be plausible and, as we pointed out in chapter 2, rejecting interpersonal comparability of welfare leads to Arrowian impossibility theorems. At any rate, Blackorby et al. are obviously not denying interpersonal comparisons of welfare since their theories presuppose the meaningfulness of such comparisons.

The second welfarist source of incommensurability can be found in the orderings of lives. It seems possible that there are pairs of lives such that we cannot say whether one is better than the other, nor can we say whether they are equally good. In real life, such cases are, of course, numerous, because of epistemological problems. But it also seems possible that there are lives whose welfare is incomparable in principle. This kind of incommensurability would carry over to population axiology. Let's say that we have two populations of the same size

---

[8] Notice that we are not suggesting that the discussion below covers all the possible welfarist sources for incommensurability among populations.

consisting of lives whose welfare is incommensurable, that is, we cannot determine whether the lives in one of the populations have at least as high welfare as the lives in the other populations, and *vice versa*. Other things being equal, these populations are certainly incommensurable.

If there are lives whose welfare is incommensurable, then the relation "has at least as high welfare" is not complete over the set of all possible lives and we will only have a quasi-ordering of possible lives. As we pointed out in our discussion of this subject in chapter 2, our adequacy conditions only presuppose a quasi-ordering of lives. Blackorby et al., on the other hand, presuppose completeness, since they assume that welfare can be measured on a ratio-scale and measurement on this scale, in turn, presupposes the completeness of the relation "has at least as high welfare as" over the set of lives whose welfare is measured. In other words, incompleteness in the ordering of lives in regard to welfare is not available for Blackorby et al. as a source of incommensurability among populations. More importantly, incompleteness in the ordering of lives would hardly yield the kind of incommensurability among populations that Incomplete Critical-Level Utilitarianism implies. It would be bizarre to claim that lives with very high welfare are incomparable in regard to welfare with lives enjoying very low positive welfare, or, for that matter, that hellish lives are incomparable with lives enjoying positive welfare. In other words, the plausible incommensurability among lives that may exist can hardly be wielded as an argument against the adequacy conditions that we have proposed so far.

5.2 Fehige's Antifrustrationism

Christoph Fehige has proposed an axiology called *antifrustrationism*. Roughly, only frustrated preferences count, and they count negatively, whereas satisfaction of preferences has no value in itself: "What matters about preferences is not that they have a satisfied existence, but that they don't have a frustrated existence."[9]

The value carriers in Fehige's axiology are material conditionals of the following type:

(i) If individual *a* at time *t* wants with strength *s* that *p*, then *p*

[9] Fehige (1998), p. 518.

In Fehige's terminology, any material conditional of type (i) is called a "Good Sentence".[10] It has negative value if it is false and neutral value if it is true. Such a conditional is false, of course, if the antecedent is true and the consequent is false. It can be true, however, in two ways: Either the antecedent and the consequent are both true, or the antecedent is false. A key assumption in Fehige's axiology is that it doesn't matter in which of these two ways a Good Sentence is true: "… [T]he two options – a satisfied preference and no preference – are equally good …"[11] If a person doesn't exist in a world, then all Good Sentences referring to her are true in that world since the antecedent is false – if *x* doesn't exist, then *x* doesn't want *p*.

How does this carry over to population axiology? A person's welfare is determined by the Good Sentences that are true or false of her. To determine the relative value of two populations, we compare the truth-values of the Good Sentences in the two worlds in which these populations occur. For example, to determine the value of adding a person, we compare the truth-values of the Good Sentences in the world where she exists with the truth-values of the Good Sentences in the world where she doesn't exist.

The principle that Fehige suggests for comparing Good Sentences is a combination of two dominance principles. First he gives two conditions for comparisons of individuals' welfare:[12]

The Principle of Antifrustrationism (PAF)

(i) If the Good Sentences true of individual *a* form a proper subset of those true of individual *b*, then *b* is better off than *a*.

(ii) If the Good Sentences true of individual *a* form a subset of those true of individual *b*, then *b* is at least as well off as *a*.

This principle looks like it yields *interpersonal* comparisons of welfare and it seems that Fehige intends it to yield such comparisons (as we shall soon see). But since any Good Sentence involves a reference to a specific individual, it is hard to

---

10 Fehige (1998), p. 509.
11 Fehige (1998), p. 508.
12 Fehige (1998), p. 524.

see how the Principle of Antifrustrationism can compare different individuals' welfare. The set of Good Sentences true about me *cannot* be a subset of the Good Sentences true about you since these two sets of sentences refer to different people – you and me. Thus, the principle of Antifrustrationism can only make *intrapersonal* comparisons of welfare, that is, compare the welfare of the *same person* in different worlds. A less misleading formulation of this principle would run as follows:

(i) If the Good Sentences true of *a* in world A is a proper subset of those true of *a* in world B, then the welfare of *a*'s life is higher in B than in A.
(ii) If the Good Sentences true of *a* in world A is a subset of those true of *a* in world B, then the welfare of *a*'s life is at least as high in B as in A.

Notice also that the Principle of Antifrustrationism yields a very incomplete intrapersonal ordering of lives. Let's say that the only difference between your life in world A and B is that in A you have at *t* and *t+1* preferences for chocolate muffins but only the former is satisfied, and in B you have the same muffin-preferences but only the latter is satisfied. This seems like a minute difference that doesn't affect your welfare in any relevant sense but it is enough for silencing Fehige's Principle of Antifrustrationism: There won't be any subset of the Good Sentences true of you in world A which is a subset of those true of you in world B and vice versa, since there is one Good Sentence which is true in A which is not true in B and vice versa. One can construct analogous examples involving the strength parameter.

Fehige's second dominance principle determines the comparative value of worlds:[13]

[13] See Fehige (1998), p. 529. I have reformulated Fehige's principles in a terminology akin to the one used elsewhere in this essay.

The Format of a General Universal Pareto Principle (FGUPP)

(i) If there exists a mapping from the set of lives in world A to the set of lives in world B such that for every pair (a_i, b_j) of lives from A and B, a_i has at least as high welfare as b_j, and for at least one pair, a_i has higher welfare than b_j, then A is better than B.

(ii) If there exists a mapping from the set of lives in world A to the set of lives in world B such that for every pair (a_i, b_j) of lives from A and B, a_i has at least as high welfare as b_j, then A is at least as good as B.[14]

The motivating idea behind FGUPP is, says Fehige, "universalizability … the ideal, widely accepted in ethics, that it must not matter who plays which part".[15] The diagram below provides an illustration:

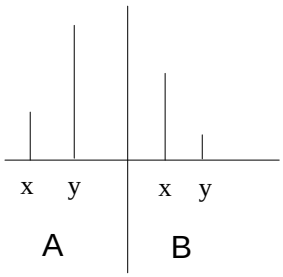

Diagram 5.2.1

Although *x* is worse off in A as compared to B, FGUPP would rank A as better than B since there is a mapping such that every person in A is better off than the person in B to whom she is compared: *x*-in-A is better off than *y*-in-B and *y*-in-A is better off than *x*-in-B. Notice that this principle presupposes that interpersonal comparisons of welfare are meaningful – otherwise it would not make sense to compare the welfare of lives led by different people. FGUPP differs in this important respect from the classical Pareto Principle which doesn't employ interpersonal comparisons.[16] On the other hand, FGUPP is much less information-

---

[14] To avoid contradictory results when comparing populations of infinite size, some additional restrictions are needed which I haven't included above. See Fehige (1998), pp. 528-9.

[15] Fehige (1998), p. 527.

[16] It is therefore misleading to say that in choices which involve the same people, FGUPP plus antifrustrationism "… *is compatible with, and can be plugged into, practically every social welfare function* …

demanding than the other theories we have discussed so far. FGUPP only presupposes that possible lives can be ordered by the relation "has at least as high welfare as".

The population axiology that Fehige proposes is FGUPP in conjunction with the Principle of Antifrustrationism. He calls this combined principle "the General Universal Pareto Principle" (GUPP). As we saw above, the Principle of Antifrustrationism only yields an intrapersonal quasi-ordering of possible lives – it only directs us how to order some possible lives of the same person. Hence, there is a pressing gap in Fehige's theory: The Principle of Antifrustrationism yields an ordering which doesn't contain enough information for FGUPP to do the work Fehige intended it to do. As a matter of fact, GUPP turns out to be just a restatement of the classical Pareto principle formulated in terms of welfare.

Let's, for the sake of argument, reformulate the Principle of Antifrustrationism so that it yields a partial interpersonal ordering of lives.[17] At the same time, let us get rid of the counter-intuitive restriction to "same time" wants. Let a *Good Sentential Formula* be any material conditional (If x at time t wants with strength s that p, then p). The Good Sentential Formula (If x at time t wants with strength s that p, then p) is *true of individual a* if and only if the Good Sentence (If a at time t wants with strength s that p, then p) is true. For any times t_1 and t_2, the formulas (If x at time t_1 wants with strength s that p, then p) and (If x at time t_2 wants with strength s that p, then p) are *corresponding* Good Sentential Formulas. We can now reformulate the Principle of Antifrustrationism in terms of Good Sentential Formulas:

---

that anybody would ever dream of defending" (Fehige 1998, p. 537, emphasis in original). Avoidance of interpersonal comparisons of preference satisfaction is a defining character of the classical Pareto Principle. The appeal of the classical principle trades on unanimity and thus avoids all interpersonal comparisons of preference satisfaction. See Mongin (1997) for a lucid discussion of different (mis)understandings of the classical Pareto Principle. Fehige's Pareto principle also differs from the classical formulation in another important respect. The latter is formulated in terms of people's preferences over alternatives, not in terms of people's welfare. Cf. Mongin (1997), p. 5.

[17] This refomulation draws on a suggestion from Howard Sobel.

The Principle of Antifrustrationism 2 (PAF2)

(i) If the Good Sentential Formulas true of individual *a* correspond one-to-one to the members of a proper subset of those true of individual *b*, then *b* has higher welfare than *a*.

(ii) If the Good Sentential Formulas true of individual *a* correspond one-to-one to the members of a subset of those true of individual *b*, then *b* has at least as high welfare as *a*.

This principle yields an interpersonal quasi-ordering of lives and avoids the "muffin" counter-example above which we directed against Fehige's formulation of antifrustrationism.[18] Henceforth, we shall assume that GUPP is a combination of the above principle and FGUPP.

The implications for population axiology look pretty straightforward: An addition of people cannot make a population better but it can make it worse.[19] Since it is possible that a population with very high welfare only involves lives with complete preference satisfaction, GUPP avoids the Repugnant Conclusion and implies the Quality Condition since any population with complete preference satisfaction is at least as good as any other population. Likewise, it entails the Weak Quality Addition Condition.

More troubling, however, is that GUPP implies a strong version of the Reversed Repugnant Conclusion: A population with very high positive welfare can be worse than an empty population. Since most lives with very high welfare can be assumed to have at least one frustrated preference, such lives are worse than non-

---

[18] It is only going to be a quasi-ordering since lives involving preferences of different strength will in many cases not be ordered by PAF2. For example, let's say that the only difference between your life in world A and B is that in A you have at t_1 a preference with *strength ten* for a muffin and at t_2 a preference with *strength eleven* for a muffin, but only the former is satisfied, whereas in B you have the same muffin-preferences but only the latter satisfied. In this case, there won't be any subset of the Good Sentential Formulas true of you in world A which is a subset of those true of you in world B and vice versa, since there is one Good Sentential Formula which is true in A which is not true in B and vice versa.

[19] Fehige (1998), p. 537, claims that GUPP entails "… that it is ceteris paribus wrong to bring people into existence who will have an unfulfilled preference". In fact, contrary to Fehige, this doesn't follow from GUPP. It only follows that an addition of a person with an unfulfilled preference makes a population worse, other things being equal. Fehige's easy moves between deontic and axiological terms suggests that he presupposes the truth of some form of act-consequentialism. Nothing to that effect is, however, stated in his paper.

existence according to Fehige's theory. Consequently, GUPP yields that an empty population can be better than a population consisting of lives with very high positive welfare.

Fehige is probably aware of this implication since he states that: "Nothing can be better than an empty world (a world without preferences, that is)." But he is clearly wrong when he claims "that the only alternative to [the claim above] is obligations to procreate – and now who's being counter-intuitive?"[20] Fehige conflates an axiological statement with a deontic statement: One can consistently hold the view that a world with very satisfied people is better than an empty world but deny any obligation to procreate. It is possible even inside a consequentialist framework to favour some worlds with very high welfare over an empty world without implying counter-intuitive obligations, since one can appeal to other values such as (parental) autonomy.

One might object here that we've misrepresented Fehige's position and that his theory doesn't imply the conclusions which we attributed to it above. Rather, his position should be construed as a "Schopenhauerian" theory of well-being: There are no possible lives with positive welfare. On the contrary, most lives have negative welfare and the best possible lives only have neutral welfare.[21]

If this is Fehige's position, then his theory violates one of the preconditions of this study: We have assumed that there are possible lives with positive welfare. Of course, a theory about welfare that denies this is highly counter-intuitive. Antifrustrationism in this interpretation implies, to take just one of its many odd implications, that a life of one year with complete preference satisfaction has the same welfare as a completely fulfilled life of a hundred years, and has higher welfare than a life of a hundred years with all preferences but one satisfied.

Notice also that if Fehige's theory implies that there cannot be any people with positive welfare, then it follows in a trivial way that his theory implies the Repugnant Conclusion, since this conclusion is formulated in terms of people with positive welfare (cf. section 3.1.2). His long discussion of this matter would thus be superfluous.

---

[20] Fehige (1998), pp. 521-2.

[21] This interpretation of Fehige's theory is put forward by Ryberg (1996), p. 140-1. Ryberg rejects antifrustrationism as an "implausible theory of well-being".

Fehige isn't very clear on this point, so I'm not sure which is his position. There are plenty of indications, however, that he believes that there are possible lives with positive or negative welfare. At some moments, he seems to mean that if a person considers her life worth living, then she has positive welfare.[22] At other moments, he talks about lives which are "very very happy" as compared to lives which are "dreadful" and "terrible".[23]

Understood in this manner, Fehige's theory is a population axiology analogous to Blackorby et al.'s Critical Level theory but with a maximally high positive critical level. This interpretation is also supported by Fehige's comment on Critical Level theories: "[T]hough it is a move in the right direction, it does not go far enough."[24]

We have already observed that Fehige's theory seen as a population axiology has one perplexing implication. How does GUPP fare in regard to the Non-Sadism Condition? As a matter of fact, Fehige's theory is neutral in regard to this condition and to most of the conditions that we shall discuss. His theory neither implies any of those conditions, nor their negations. This is so because GUPP is deeply incomplete – it is silent on most of the important issues in population axiology. The relation between GUPP and the Non-Sadism Condition provides a good illustration. Let's say that we have to choose between adding a few people with negative welfare or several people with positive welfare:

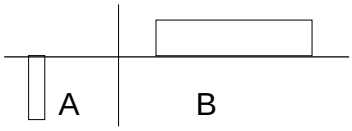

Diagram 5.2.2

Given that the people with positive welfare have at least one preference frustrated, GUPP has nothing to say about this case, which seems especially awkward from the perspective of Antifrustrationism. GUPP's impotence follows from its lack of any specification of how to weigh some people's preference

---

[22] Fehige (1998), p. 527. Fehige talks about dashes which represent individuals' welfare, and "if a dash is above the horizontal line, then the corresponding person considers her life worth living".
[23] Fehige (1998), p. 531-5.
[24] Fehige (1992), p. 16.

frustrations against other people's frustrations. Although every individual in A is much more frustrated than any individual in B, there are more people that have frustrated preferences in B. Since A and B differ in population size, there is, for example, no mapping such that for every pair (a_i, b_j) of lives from A and B, the B-person is better off than the A-person. The A-people can be mapped onto a subset of the B-population such that for every pair, the B-people are better off. But the rest of the B-population has to be mapped against the "welfare" of their ghostly non-existence in A and since non-existence implies maximal preference satisfaction (minimal preference frustration), the rest of the B-people are "worse off" in B than in A.

The case above can be made more extreme: Consider a large population with n tormented people and a population of $n+1$ persons with blissful lives – Fehige's theory is silent on the relative values of these two populations. This is clearly unsatisfactory. The more surprising then, that Fehige claims that GUPP "… entails the Narveson type of slogan that we have obligations to make people happy (preferrers satisfied), but no obligations to make happy people (satisfied preferrers)", and "that it is obligatory not to bring into existence an unhappy person".[25]

[25] Fehige (1998), p. 538. Narveson's own theory implies that B rather than A ought to be the case since his "… concern is that whatever people there are be as happy as possible …". See Narveson (1978), p. 55. Cf. section 8.3.

6

Discontinuity and Lexical Levels

6.1 Griffin's Discontinuity

In a discussion of aggregation of individual welfare, James Griffin has proposed that there can be a "discontinuity" among prudential values (welfare) of the form "enough of A outranks any amount of B". Discontinuity entails, he explains:

> … the suspension of addition; … we have a positive value that, no matter how often a certain amount is added to itself, cannot become greater than another positive value, and cannot, not because with piling up we get diminishing value or even disvalue … , but because they are the sort of value that, even remaining constant, cannot add up to some other value. - - - … it is more plausible that, say, fifty years at a very high level of well-being – say, the level which makes possible satisfying personal relations, some understanding of what makes life worth while, appreciation of great beauty, the chance to accomplish something with one's life – outranks any number of years at the level just barely living – say, the level at which none of the former values are possible and one is left with just enough surplus of simple pleasure over pain to go on with it.[1]

In a comment on Parfit's discussion of the Repugnant Conclusion, Griffin illustrates how this reasoning could carry over to population axiology:

> … there is another possibility confined entirely to the reasoning about beneficence. Parfit's argument seems implicitly to employ a totting-up conception of measuring well-being; it treats well-being as measurable on

[1] Griffin (1986), p. 85-6.

> a single continuous additive scale, where lower numbers, if added to themselves often enough, must become larger than any initial, larger number. But this seems not true in prudential cases, and it would seem likely that this [discontinuity…] in prudential values would get transferred to interpersonal calculation. Perhaps it is better to have a certain number of people at a certain high level than a very much larger number at a level where life is just worth living. Then we might wish to stop the slide [to the repugnant conclusion …] at that point along the line where people's capacity to appreciate beauty, to form deep loving relationships, to accomplish something with their lives beyond just staying alive … all disappear.[2]

One might think that the principle that Griffin has in mind is something like the following:

> *The Lexical View*: There is no limit to the positive value of the total sum of welfare; but the contributive value of a sufficiently large number of lives *n* with very high welfare ("a certain high level") is higher than the contributive value of any number of lives with very low welfare ("a level where life is just worth living").[3]

As we understand this view, welfare is measurable on a ratio scale but there are two quality levels – a high and a low level – defined by a certain amount of welfare in a life. This principle is supposed to be a modification of Total Utilitarianism and it will yield the same result in all cases where the number of high quality lives aren't "sufficiently large", that is, less than *n*. There is, however, a kind of lexical superiority of the high quality lives over the low quality ones: If we have *n* or more people enjoying welfare above the high level, then the contributive value of these

---

[2] Griffin (1986), en. 27, p. 340.

[3] There are other versions of the Lexical Principle but they all share the problems discussed above. Some of these versions are discussed in Arrhenius and Bykvist (1995), p. 75. Parfit (1984), p. 414 formulates a similar principle but in terms of "mediocre" and "blissful" lives and "no amount of Mediocre lives could have as much value as one Blissful life".

lives is higher than the contributive value of any number of lives with welfare below the low quality level.

This view has been proposed by Roger Crisp as an interpretation of Griffin's idea of discontinuity of value and as a plausible way to avoid the Repugnant Conclusion.[4] It is not easy to understand exactly what this principle amounts to and, as I shall shortly argue, I don't think it is Griffin's view. At any rate, while it seems pretty clear that the Lexical View respects the Quality Condition, it is in other respects an incomplete population axiology. It doesn't specify, for example, whether for any number of high quality lives, there is a number of lives with welfare in between the two quality levels, let's call them "middling" lives, which taken together have higher contributive value. Similarly, for any number of middling lives, is there a number of low quality lives which taken together have higher contributive value? Since the Lexical View is supposed to be a modification of Total Utilitarianism, and the only modification is the lexical ordering of (a certain number of) high quality lives over low quality ones, one would guess that the answer is yes in both cases. But then this view is inconsistent. If population A consists of *n* people with high quality of life, then there exists a population B with middling lives which is better, and a population C with low quality lives which is better than B, and, by transitivity, better than A. But according to the last clause of the Lexical View, A is better than C.

Even if there is an intuitive way of rendering the Lexical View consistent (I doubt it), it would still violate the Non-Anti Egalitarianism Principle.

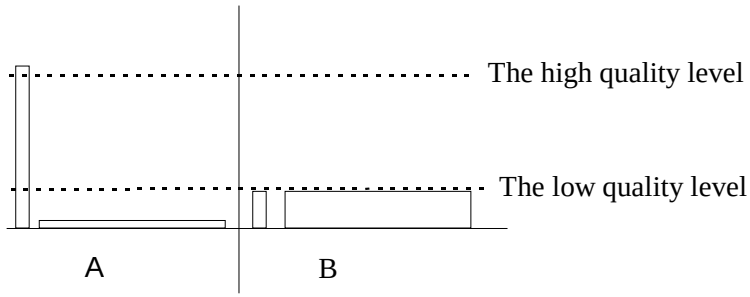

Diagram 6.1.1

[4] See Crisp (1988), p. 188, and Crisp (1992), p. 151. Cf. Klint Jensen (1996), pp. 90-1.

Consider the above case where in outcome A we have a number of people slightly above the high quality level and a much greater number of people with positive welfare below the low quality level. In outcome B we have decreased the welfare of the best off people but increased the lot of the worst off. The average and total utility is greater in B than in A and there is perfect equality in B. If the number of the best off people in A is sufficiently large, then A is better than B according to the Lexical View, since no increase in the total welfare in lives below the low quality level can outweigh a loss in the total welfare in lives above the high quality level.

As I said above, I don't think that Griffin has the Lexical View in mind. Rather, I think that he simply suggests that the welfare of lives is not measurable on a scale that makes talk about average and total welfare meaningful, or at least that holds for the kind of lives we have been discussing here. In other words, the ordering of all possible lives in regard to welfare cannot be represented on a ratio or an interval scale but only on some weaker scale, such as an ordinal scale. There might be orderings of subsets of all possible lives which can be represented by an interval or a ratio scale – "pockets of cardinality", so to speak.[5] Subsets including both lives with very high and very low welfare are not, however, of this kind.

In a discussion of whether the appreciation of the beauty of Rembrandt's paintings can be substituted by the appreciation of the rest of the Dutch school or kitsch, he writes:

> We have simply to *rank a life with beauty against a life with only lots of kitsch*. - - - What goes on in comparing a few Rembrandts with all the rest of the Dutch School is not arithmetical addition of a larger number of slightly smaller values to a great overall sum. We have to decide how we value greater number and more variety against a few supreme, less varied examples. And this is itself a *basic preference*.[6]

---

[5] Griffin (1986), pp. 98-105, uses this expression in his discussion of measurement of welfare to point out, correctly, that even if welfare cannot always be measured on an interval or ratio scale, it doesn't follow that welfare can never be measured on such scales. As obvious as this may seem, I think that Griffin is making an important point which has been largely overlooked in the discussion of the measurement of welfare.

[6] Griffin (1986), p. 88, emphasis added.

Similarly, Griffin's statement quoted earlier – that fifty years at a very high level of well-being outranks any number of years at the level of just barely living – should just be understood to state, I surmise, that a life with fifty years at a very high level of well-being has *higher welfare* than a life made up of years at the level of just barely living. In other words, this is a *basic ranking*, not a ranking derived from some kind of intrapersonal aggregation (summing, averaging, and the like) of the welfare of the different parts of such lives.[7] Since, according to Griffin, this discontinuity "in prudential values would get transferred to interpersonal calculation", his conjecture that "[p]erhaps it is better to have a certain number of people at a certain high level than a very much larger number at a level where life is just worth living", should, I think, be understood in the similar manner: When determining which population is the best in regard to people's welfare, we cannot arrive at this judgement by some kind of *interpersonal* aggregation (summing, averaging, and the like) of the welfare of the members of the compared populations. Rather, as in the case of comparing welfare of lives, this judgement also has to be a basic ranking.

We agree, of course with the intuition that Griffin expresses regarding the basic ranking of a large population with high quality of life and a population with low quality of life, since it is exactly that intuition we have imputed in the Quality Condition. But it seems clear that Griffin thinks he is doing something more than stating an intuition – he thinks that his theory is supposed to solve the Mere

---

[7] Parfit has, in Parfit (1986), pp. 161-4, sketched a theory which looks very much like Griffin's. He writes: "Suppose that I can choose between two futures. I could live for another 100 years, all of an extremely high quality. Call this *the Century of Ecstasy*. I could instead live for ever, with a life that would always be barely worth living … the only good things would be muzak and potatoes. Call this *the Drab Eternity*. - - - I claim that, though each day of the Drab Eternity would be worth living, the Century of Ecstasy would give me a better life. This is like Mill's claim about the 'difference in quality' between human and pig-like pleasures. It is often said that Mill's 'higher pleasures' are merely *greater* pleasures … As Sidgwick wrote, 'all qualitative comparisons of pleasures must really resolve itself in quantitative [comparisons]'. - - - But this is what I have just denied. The Century of Ecstasy would be better for me in an essential qualitative way. Though each day of the Drab Eternity would have some value for me, *no* amount of this value could be as good for me as the Century of Ecstasy." (emphasis in original). Like Griffin, Parfit suggests that this feature of individual welfare aggregation would carry over to interpersonal aggregation of welfare. I think Parfit's suggestion is apt for the same analysis that we have given Griffin's theory and that it shares the positive and negative features that I ascribe to Griffin's theory below.

Addition Paradox. How could that be? Let's see how Griffin's theory might handle the Mere Addition Paradox.[8]

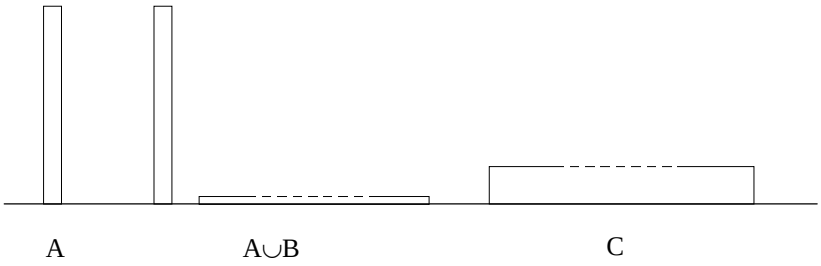

Diagram 6.1.2

In the above diagram, A is a population of people with very high welfare, B is a much larger population than A but consisting of people with very low welfare. C is a population of the same size as A∪B. Everybody in C has very low welfare but they are all better off than the people in B. Moreover, there is perfect equality in C and the total welfare in C is higher than in A∪B.

Assume that Griffin's theory, in accordance with his intuition quoted above, yields that A is better than C. What would Griffin say about the ranking of A and A∪B? It seems that his theory at least has a maximising character, other things being equal.[9] Consequently, we can assume that his theory implicitly complies with the Mere Addition Principle, that is, A∪B is at least as good as A. To avoid a contradiction, Griffin's theory must rank A∪B as better than, or incommensurable with C. In other words, it looks like Griffin's theory violates the Non-Anti Egalitarianism Principle. This is, however, not true. The Non-Anti Egalitarianism Principle is formulated in terms of "lower average and total welfare". Since Griffin holds that the welfare of lives in sets involving both lives with very high and very

---

[8] In all essentials, this version of the Mere Addition Paradox is the same as the one presented in Ng (1989). Cf. ch. 1, especially fn. 1.

[9] Griffin (1986), p. 247, writes "On the deepest level of moral theory, in what earlier I called the general characterization of the right and wrong, … the maximizing principle applies. It may not be the only principle that applies there, but that anyway is where it applies." Moreover, in his discussion of one version of Parfit's Mere Addition Paradox, his objection is not directed against Parfit's premise that an addition of people with positive welfare doesn't render an outcome worse, but that "Parfit's argument seems implicitly to employ a totting-up conception of measuring well-being …". See Griffin (1986), en. 27, p. 340.

low welfare cannot be compared on at least an interval scale, statements about total and average welfare are meaningless in regard to comparisons of populations such as A∪B and C. Thus, Griffin's theory doesn't violate the Non-Anti Egalitarianism Principle but neutralises it in regard to cases like the above. And it does so in a completely legitimate way – there is nothing mysterious about denying that the welfare of all lives can be measured on at least an interval scale.[10]

So far, Griffin's theory looks promising. It shows us how a theory can retain a maximising character but avoid the Repugnant Conclusion by a plausible assumption about the measurement of welfare: Not all lives can be compared on a scale that makes sense of talk about average and total welfare. It is, of course, sketchy and one wonders how it could be spelled out to handle, for example, alternatives that involve negative welfare. Anyway, it has given us a reasonable solution to, or rather evasion of, the Mere Addition Paradox. But there are problems ahead: Griffin's theory violates an egalitarian condition which is intuitively more compelling and logically weaker than the Non-Anti Egalitarianism Principle:

> *The Inequality Aversion Condition*: For any triplet of welfare levels **𝔸**, **𝔹**, and **ℂ**, **𝔸** higher than **𝔹**, and **𝔹** higher than **ℂ**, and for any population A with welfare **𝔸**, there is a larger population C with welfare **ℂ** such that a perfectly equal population B of the same size as A∪C and with welfare **𝔹** is at least as good as A∪C, other things being equal.

Another way of stating the Inequality Aversion Condition is to say for any welfare level of the best off and worst off, and for any number of best off lives, there is a greater number of worst off lives such that it would be at least as good to

---

[10] Similar egalitarian conditions, which thus also presuppose measurement on at least an interval scale, play a crucial role in the paradoxes/theorems of Ng (1989), Blackorby and Donaldson (1991), Arrhenius (1995, 2000a), and Parfit (1984), pp. 419-30. Parfit (1984), pp. 433-41 and (1986), pp. 156-60, presents two more versions of the Mere Addition Paradox. These paradoxes seem to involve measurement on at least a ratio scale, since they involves claims to the effect that "… a worse-off group would gain *several times as much* …" than a better off group would lose (Parfit (1984), p. 435, emphasis in original). Thus, since all these paradoxes/theorems involve measurement on a scale at least as strong as an interval scale, Griffin's theory evades these paradoxes too. Cf. fn. above.

have an equal distribution of welfare on any level higher than the worst off, other things being equal.

The Inequality Aversion Condition is applicable even if only ordinal measurement of welfare is possible. Here's an example of a principle which only presupposes ordinal measurement of welfare and satisfies this condition: If the worst off make up at least 99.99% of a population, then it would be better to have an equal distribution of welfare on a level higher than the worst off. An ordinal principle that violates the Inequality Aversion Condition is "Maximax": Maximise the welfare of the best off.

Non-Anti Egalitarianism is logically stronger than the Inequality Aversion Condition. If an axiology implies the former, then it implies the latter, but not vice versa. The Inequality Aversion Condition is, for example, satisfied by a theory which demands that the total welfare must be, say, ten times higher for a population with perfect equality to be better than an unequal population of the same size. Moreover, the Inequality Aversion Condition only presupposes that lives can be ordered by the relation "has at least as high welfare as" whereas the Non-Anti Egalitarianism Principle presupposes measurement on a scale at least as strong as an interval scale.

If Griffin's population axiology is consistent, then it is going to violate the Inequality Aversion Condition. Let, as before, population A in Diagram 6.1.2 be a population of people with very high welfare such that A is better than any population consisting of people with very low welfare. Assume that Griffin's theory complies with the Inequality Aversion Condition. Then there is a population C with perfect equality and very low but higher welfare than the B-people, which is at least as good as A∪B. The Mere Addition Principle yields that A∪B is at least as good as A. It follows that C is at least as good as A. But A is better than any population made up of people with very low welfare. Thus, we have derived a contradiction: C is at least as good as A and C is worse than A. Consequently, it cannot be the case that Griffin's population axiology satisfies the plausible Inequality Aversion Condition.

6.2 A Possible Solution to the Mere Addition Paradox

It seems to be unanimously agreed in the literature that inequality aversion of some kind is a prerequisite for an acceptable population axiology. As we mentioned above, Sider's theory has anti-egalitarian implications but he rejects his own theory

just because it favours unequal distributions of welfare. Ng states that "Non-Antiegalitarianism is extremely compelling" and Carlson claims that "[r]ejecting NAE [the Non-Anti Egalitarianism Principle] is … a very unattractive option". Blackorby et al. hold that "weak inequality aversion is satisfied by all ethically attractive … principles" and Fehige rhetorically asks "… if one world has more utility than the other and distributes it equally, whereas the other doesn't, then how can it fail to be better?"[11] However, assuming that we can distinguish between small and large welfare differences, there is a possible objection to the Inequality Aversion Condition which also can be directed against the Non-Anti Egalitarianism Principle. Assume as before that the C-people in Diagram 6.1.2 have higher welfare than the B-people. The only difference in the welfare between these two groups of people, however, is that the latter people experience a mildly painful pin-prick in their right thumb on their fifth birthday. Still, according to the Inequality Aversion Condition and the Non-Anti Egalitarianism Principle, if there is a great enough number of B-people, C is at least as good as A∪B. One could reasonably claim that one less mildly painful pin-prick seems insignificant from a moral point of view, and that such a minute increase in the welfare for each of the worst off individuals, irrespective of how many they are, cannot balance the great loss in the welfare for each of the best off individuals. Rather, we might find that in this case, A∪B is better than C or that these populations are incomparable. As we said above, there is a (somewhat) plausible source of incommensurability from a welfarist point of view to which we were going to return. This is it: We might find it impossible to weigh a very huge number of minute gains in welfare against a smaller number of great losses. And by claiming that A∪B is better than C or that these populations are incomparable; we can escape the Mere Addition Paradox.

There are problems with this solution, however. Firstly, let me point out that talk about "gains" and "losses" might be quite misleading in the present context – it sounds like we are "taking" welfare from some well-off people and "giving" it to some worse-off people, that we are considering "moving" from an existing unequal population to an equal population. This need not be the case, however. For example, the compared populations might be two future populations consisting of

---

[11] See Sider (1991), p. 270, fn. 10; Ng (1989), p. 239, fn. 4; Carlson (1998a), p. 288; Blackorby, Bossert and Donaldson (1997), p. 210; and Fehige (1998), p. 12.

different people. And it might be that in regard to cases involving the same people and a move from an existing unequal population to a possible equal population, some people find other considerations apart from welfare relevant which explains their misgivings about the Inequality Aversion Condition. One might, for example, believe in some kind of "negative right" of not having welfare components taken away, but no corresponding "positive right" to receive welfare components. A believer in strict property rights might find a change from the unequal to the equal population unacceptable if it involves taking property from the rich and giving it to the poor. Likewise, such a theorist would not accept, at least not in theory, a move from the equal to the unequal population if it involves taking property from the poor. But if other considerations are relevant, such as violations of people's rights, then the Inequality Aversion Condition is not applicable since other things are *not* equal. Consequently, this kind of criticism of the Inequality Aversion Condition misses the mark. It is important to remember that talk of "gains" and "losses" in the present context is quite metaphorical since the cases considered don't need to involve any particular people who gain or lose. Keeping this in mind, however, I don't think these expressions should lead to any misunderstandings, and for the sake of convenience of expression, we shall continue to use them.

A more serious problem for the above solution to the Mere Addition Paradox is the following. There must be at least some increase in welfare for the worst off people in Diagram 6.1.2 that can compensate for the loss in welfare for the best off people. Let's call such an increase a "significant increase in welfare". Now, if we assume that the people in C have significantly higher welfare than the people in B, then we will regain the paradox by reasoning analogous to our discussion in the previous section.

There is a possible rejoinder, however. We could claim that if C has significantly higher welfare than B, then C is better than A. This is how we could reason. If a person with very low welfare gets a significant increase in her welfare, then she no longer has very low welfare. Consequently, population C doesn't consist of people with very low welfare and it would not be an instance of the Repugnant Conclusion to claim that C is better than A. In other words, there is a way of escaping the Mere Addition Paradox without committing oneself to some kind of anti-egalitarianism. We reject the Non-Anti Egalitarianism Principle and the Inequality Aversion Condition in cases involving very small increases in welfare of the worst off but keep them in cases involving significant increases in welfare. Moreover, we claim

that a significant increase in the welfare of someone with very low positive welfare always lifts that person's welfare above the threshold of very low positive welfare.

Is this convincing? It all depends on the crucial assumption regarding significant increases in the welfare of people with very low welfare. Is it believable that for any person with very low welfare, if she gets a significant increase in her welfare, then she no longer has very low welfare? It is hard, of course, to specify exactly what a significant increase in welfare is and there will certainly be a grey zone between insignificant and significant increases in welfare. But there are increases in welfare for the greater number of worst off which taken together can outweigh a decrease in welfare for the fewer best off, albeit a greater decrease for each of the best off. Assume that a child has some incurable disease which will cause her death before she reaches the age of ten. The disease doesn't affect her day to day life until the last month when she will die pretty painlessly. Since she dies so young, however, it is reasonable to say that her lifetime welfare is very low. Now, there is a medicine which would delay the progress of the disease by a couple of years. Nevertheless, since her life would still be very short, her lifetime welfare would still be very low. It is easy to see how this example could be magnified. Assume that the B-people in Diagram 6.1.2 have very short lives. In C, people's lives are increased by a couple of years. It is hard to deny that such an improvement for a very huge number of worst off people cannot balance a decrease in welfare, albeit a greater decrease for each person, in a much smaller number of best off people. Hence, I'm sceptical regarding the viability of a solution to the Mere Addition Paradox based on the claim that there are great losses for the best off that cannot be outweighed by any number of small gains for the worst off. Moreover, in spite of the criticism of the Inequality Aversion Condition above, there is a compelling reason not to abandon it. As we shall see in the next section, it is implied by a condition that is very hard to reject.

6.3 Non-Elitism

The objection to the Inequality Aversion Condition in the beginning of the preceding section was based on a concern about how to weigh a great number of small increases in welfare against a smaller number of greater decreases in welfare. As we said, one might find it impossible to weigh a very huge number of minute gains in welfare against a smaller number of great losses, or one might think that great enough losses cannot be outweighed by any number of very small gains.

However, the Inequality Aversion Condition can be derived from a condition which doesn't involve such comparisons. Consider the following condition:

> *The Non-Elitism Condition*: For any triplet of welfare levels $\mathbb{A}$, $\mathbb{B}$, and $\mathbb{C}$, $\mathbb{A}$ slightly higher than $\mathbb{B}$, and $\mathbb{B}$ higher than $\mathbb{C}$, and for any one-life population A with welfare $\mathbb{A}$, there is a population C with welfare $\mathbb{C}$, and a population B of the same size as A∪C and with welfare $\mathbb{B}$, such that for any population X consisting of lives with welfare ranging from $\mathbb{C}$ to $\mathbb{A}$, B∪X is at least as good as A∪C∪X, other things being equal.

Roughly, the intuition which the Non-Elitism Condition tries to capture is that there is at least some very small decrease in the welfare of *one* of the best off lives which can be compensated for by an increase in welfare for at least some number of worst off people. Consequently, the application of the Non-Elitism Condition doesn't need to involve any comparisons of great losses and small gains. The gains and losses for each involved individual might be of the same size or the gain for each worst off individual might be greater than the loss for each best off individual.

In chapter 10, we shall formally prove that the Non-Elitism Condition implies the Inequality Aversion Condition.[12] Here, we shall only give an intuitive argument.

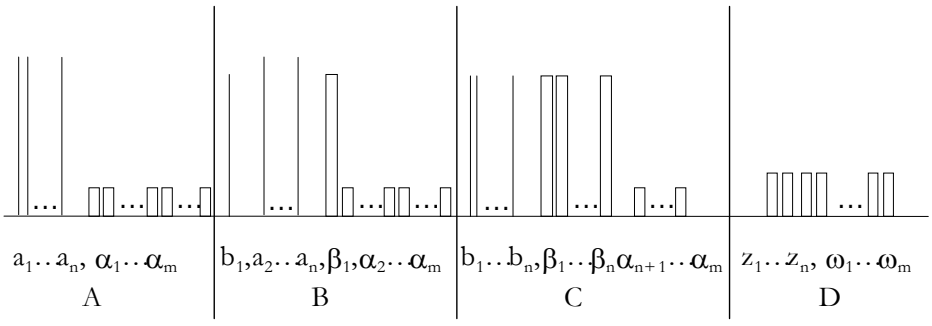

Diagram 6.3.1

[12] As in the demonstration in section 3.2, we are assuming here that the differences between any two welfare levels consists of a finite number of "slight welfare differences" which seems intuitively correct. We shall make this assumption more precise in chapter 10.

The diagram above shows four populations of the same size. Population A consists of a number of best off lives, a_1, a_2, …, a_n, and a number of worst off *groups* of lives, α_1, α_2, …, α_m. In population B, one of the best off lives (a_1) has been replaced with a life (b_1) enjoying welfare just lower than the welfare of the best off. Moreover, one of the worst off groups of lives (α_1) has been replaced by a same sized group of lives (β_1) with the same welfare as life b_1. Assume that this decrease in welfare for one of the best off lives is compensated for by the increase in the welfare for one of the group of worst off lives, that is, population B is at least as good as A according to the Non-Elitism Condition (we can assume, if we so choose, that the gains for each of the worst off individuals are much greater than the loss for each of the best off individuals). By repeating the same procedure n-1 times, we will reach population C in which all the best off lives (b_1, b_2, …, b_n, β_1, β_2, …, β_n) enjoy the same welfare as life b_1. By transitivity, C is at least as good as A. Moreover, as long as there are enough worst off groups in population A (and thus in C), that is, as long as m is great enough, we can repeat this process to reach population D where everybody enjoys welfare at some given level higher than the welfare of the worst off people in A. Consequently, Non-Elitism implies the Inequality Aversion Condition. This puts those theorists, who want to reject the Inequality Aversion Condition because its application might involve comparisons of a small number of great losses against a greater number of small gains, in an awkward spot: They also have to reject the Non-Elitism Condition, but the application of this condition doesn't involve the kind of comparisons that these theorists worry about.

One might argue that we shall reject the Non-Elitism Condition just because it together with transitivity implies the Inequality Aversion Condition. But I think such a move would be disingenuous. To find a convincing reason for rejecting the Non-Elitism Condition, we must find a pairwise comparison to which it gives the wrong answer, or at least an answer for which we could give some plausible reason why it might not be the right answer. And I'm pretty sure that we cannot find such a case: How could it be that there is no slight loss for one best off individual that cannot be balanced by some number of gains for the worst off? To claim this is tantamount to claiming that there are cases where only the welfare of the best off counts.

I think there are only two options here: Accept the Inequality Aversion Condition or reject transitivity of the relation "is at least as good as". Since we have

understood a population axiology to be at least a quasi-ordering of populations, the latter move amounts to giving up the project of finding an acceptable population axiology (we shall return to this issue in chapter 11). And since the Inequality Aversion Condition is quite plausible in itself, I don't think the fact that it follows from any population axiology which satisfies the Non-Elitism Condition gives us enough reason to abandon this project – it is all too early to give up our hopes. Rather, what the above discussion shows, I think, is that we have to jettison the idea that there are great decreases in the individual welfare of the best off which cannot be balanced by a sufficient number of small increases in the individual welfare of the worst off.

6.4 Extreme Negativism, Maximin and Leximin

According to Negative Total Utilitarianism the value of a population is calculated by summing the welfare of all lives with *negative* welfare in the population. Lives with positive welfare neither add to nor detract from the value of a population. Consequently, Negative Total Utilitarianism violates the following plausible condition:

> *The Non-Extreme Priority Condition*: There is a number *n* of lives such that for any population X, a population consisting of the X-lives, *n* lives with very high welfare, and a single life with slightly negative welfare is at least as good as a population consisting of the X-lives and *n*+1 lives with very low positive welfare, other things being equal.

According to Negative Total Utilitarianism, if one population contains a life with negative welfare, and another doesn't, then the latter population is always better and the difference in positive welfare doesn't matter at all. Negativist theories that violate the Non-Extreme Priority Condition give too much weight to negative welfare since they don't allow for any trade-offs between negative and positive welfare.

It is not only extremely negativist theories that violate the Non-Extreme Priority Condition, however. A well-known principle is the Maximin Principle. Maximin ranks populations according to the welfare of the worst off: The lower the welfare of the worst off, the worse the population, and if the worst off enjoy the same welfare in two populations, then these populations are equally good. In other

words, Maximin gives priority to the welfare of the worst off. Maximin clearly satisfies the adequacy conditions that we have suggested earlier (we leave it to the reader to check this). The same holds true for Maximin's cousin, Leximin. According to Leximin, if the worst off in A are better off than the worst off in B, then A is better than B. If the worst off in A and B have the same welfare, then A is better than B if the second worst off in A are better off than the second worst off in B, and so forth.[13] Both of these principles violate the Non-Extreme Priority Condition in cases where the worst off in the compared populations is a person with slightly negative welfare. Maximin and Leximin don't only rule out trade-offs in such cases, but in all cases where a gain for the worst off is at stake. Consequently, these principles also violate a generalised version of the Non-Extreme Priority Condition:

> *The General Non-Extreme Priority Condition*: There is a number *n* of lives such that for any population X, and any welfare level 𝔸, a population consisting of the X-lives, *n* lives with very high welfare, and one life with welfare 𝔸, is at least as good as a population consisting of the X-lives, *n* lives with very low positive welfare, and one life with welfare slightly above 𝔸, other things being equal.

According to Maximin and Leximin, if the worst off life has higher welfare in one population as compared to another one, then the former population is always better and the differences in the welfare of the other lives in the compared populations don't matter at all. In other words, the slightest gain in welfare for one person outweighs a very large loss for any number of people. Of course, Maximin and Leximin imply conclusions that are even more extreme than their violation of the General Non-Extreme Priority Condition. Assume that a population A consists of a very large number of people with blissful lives and one person suffering terrible pain. In another population B, everybody suffers terrible pain but slightly

---

[13] As we have stated Leximin, it's unclear how it could compare populations of different size. One can reformulate Leximin in different ways to widen its scope, but for reasons which will soon be apparent, we shall not pursue this question further.

less than the poor person in A. According to Maximin and Leximin, B is better than A.

One could say that Maximin and Leximin imposes a dictatorship of the worst off. In general, I think that principles that violate the General Non-Extreme Priority Condition give too much weight to the welfare of the worst off since they don't allow for any trade-offs between gains in the welfare of the worst off and losses in the welfare of those who are better off.

In conjunction with the transitivity of the relation "at least as good as", however, the General Non-Extreme Priority Condition has an implication that some theorists with a negativist inclination might find bothersome. It implies, as we shall show in chapter 10, that for any given number of lives with very negative welfare, there is a (much) greater number of lives with very high welfare such that a population consisting of these two groups of lives is at least as good as a same sized population consisting of lives with slightly positive welfare. Let's call this implication *bad lives for very good lives*.

As a matter of fact, I'm not sure that *bad lives for very good lives* is counter-intuitive. Consider the two following possible futures. In A, the vast majority of people have very high welfare and only a tiny tiny fraction of all these billions of billions of people have very bad lives. In B, all of these billions of billions of people have lives barely worth living. In the choice between these two futures, I don't find A clearly worse than B. As Alastair Norcross points out, most of us seem to be ready to accept even the much stronger claim that small benefits to a great enough number of people can outweigh great harms for a given number of people:

> If there were a national speed limit of 50 mph [in USA], it is overwhelmingly likely that many lives would be saved each year, as compared with the current situation. One of the costs of the failure to impose such a speed limit is a significant number of deaths. The benefits of higher speed limits are increased convenience for many. Despite this, it is far from obvious that failure to impose a 50 mph speed limit is wrong.[14]

---

[14] Norcross (1997), p. 159. There has been quite a lively discussion in Norcross (1997, 1998), Temkin (1996), Rachels (1998) and Carlson (1998b) on how to weigh a great number of small

Those that agree with Norcross, shouldn't have any problem in accepting the comparatively much weaker conclusion *bad lives for very good lives*. At any rate, I don't think we are facing a choice between accepting or rejecting the General Non-Extreme Priority Condition. As we said in connection with our discussion of the Non-Elitism Condition, to find a convincing reason for rejecting a principle, we should find a pairwise comparison to which it gives the wrong answer, or at least an answer for which we could give some plausible reason why it might not be the right answer. And for every pairwise comparison, the General Non-Extreme Priority Condition seems extremely compelling: How could it be that there is no number of great losses that cannot outweigh a slight gain for one person? Again, we are faced with two options: Accept *bad lives for very good lives* or abandon the project of finding a plausible population axiology. And since *bad lives for very good lives* is not clearly counter-intuitive – most people seem to be ready to accept much more radical trade-offs – I think it is all too early to conclude that we have an impossibility theorem for an acceptable population axiology on our hands.

It is instructive to compare the Non-Extreme Priority and the General Non-Extreme Priority Condition with the Inequality Aversion and the Non-Elitism Conditions. Roughly, according to the latter conditions, it is not the case that only the welfare of the best off matters, whereas according to the former condition, it is not the case that only the welfare of the worst off matters. A reasonable population axiology should yield a trade-off between the welfare of the worst off and best off that avoids the extreme solution of prioritising the welfare of just one of these groups.

---

benefits against a smaller number of great harms. This discussion has its source in Quinn's (1990) classical paper. Much more could be said about this interesting topic than we have limited ourselves to above, and I think that the main issues in this discussion could be considerably clarified by adopting the kind of formalism that we shall present in ch. 10. Regrettably, we don't have the space to pursue this discussion further here.

7

Welfarist Egalitarianism and the Priority View

7.1 Introduction

Equality plays a fundamental role in moral and political reasoning. Views about equality can differ immensely, however, depending on a number of factors: What kind of equality one is seeking (political, legal, moral, and so forth); the "currency" of equality (welfare, opportunity, rights, and so forth); among what kind of objects equality is supposed to hold (citizens, human beings, sentient beings, possible beings, groups, and so forth). It goes without saying that a full treatment of this subject is far beyond the reach of the present essay.[1] We shall take a look at one kind of equality: equality of welfare among people. Still, this is such a complex idea that we cannot give it the full treatment it deserves. Our project is further limited. Our main task is to consider some possible egalitarian objections against the adequacy conditions that we have suggested. Of course, in doing so we cannot avoid discussing general questions about the value of equality of welfare. However, our answers to those questions will be tentative.

7.2 Welfarist Egalitarians

One can distinguish two kinds of Welfarist Egalitarians: monists and pluralists. The former think that equality of welfare is the sole consideration when ranking populations, whereas the latter think that equality of welfare is one among other relevant factors in ranking populations. Probably, no one has ever held the position of the Monist Welfarist Egalitarian since it implies, to say the least, clearly unpalatable conclusions. For example, it has the absurd implication that a population with very high welfare and some inequality is worse than a population

[1] See, for example, Broome (1991), Roemer (1996), Sen (1992), and Temkin (1993a) for a discussion of many of the intricate aspects of equality.

of equally tormented people. Strict Welfarist Egalitarianism violates the following very plausible condition:

> *The Dominance Principle*: If population A contains the same number of people as population B, and every person in A has higher welfare than every person in B, then A is better than B, other things being equal.

Monist Welfarist Egalitarianism has a number of other counter-intuitive implications in same-number cases, but let's leave those aside and turn to Pluralist Welfarist Egalitarianism. A reasonable pluralist has to downplay the importance of equality of welfare such that her theory satisfies the Dominance Principle. Thus, it is fair to ask: In which cases can equality of welfare make a difference in the ranking of populations? This is a tricky question that the Pluralist Welfarist Egalitarian has to answer. As a matter of fact, I doubt that many people, on reflection, really believe that equality of welfare has a value in itself. This might seem surprising since equality is such an entrenched value in moral and political reasoning. Most of us believe in some kind of equality (I certainly do), such as equality before the law, equal rights, political equality, similar cases should be treated equally, everyone's interests matter and matter equally, and so forth.[2] These ideas of equality are very important but different from the idea of equality of welfare. One reason why appeals to equality of welfare look attractive at first sight is, I think, that these other kinds of equality are important and reasonable considerations. There is nothing inconsistent, however, in endorsing those kinds of equality and rejecting appeals to equality of welfare.

It is also important to remember that to reject the idea that equality of welfare has value in itself is not to deny that equality of welfare may have good effects and that inequality may have bad effects. Inequality of welfare can undermine people's self-respect, cause envy and thus undermine the cohesion of society, be bad for the economy, and so forth. Consequently, inequality of welfare can diminish the general welfare in a population. As true as this might be, this is beside the point of the matter since if any such factors are at play, then the effects are already included

---

[2] Kymlicka (1990), p. 4, suggests that all modern moral and political theories are based on some conception of equality.

in the specification of people's welfare. What we are considering is whether equality of welfare has a value in itself, apart from any instrumental side effects it might have.

7.3 The Priority View

Many might still find my claim about the role of equality of welfare in our moral reasoning perplexing since there are so many cases where we clearly appeal to exactly such considerations. Typically, they would point to cases like the following:

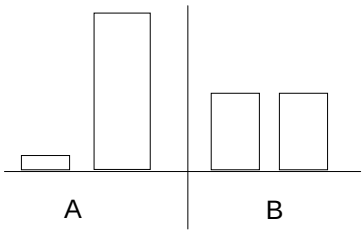

Diagram 7.3.1

The two populations A and B in Diagram 7.3.1 are equally large and have the same average utility. The only difference is that there is inequality in A whereas B is perfectly equal. Is it not obvious that B is better than A and does that not show that equality of welfare is a value in itself? Why would we otherwise rank B as better than A?

I certainly agree that B is better than A but this is not because I value equality of welfare as such, but because the worst off are better off in B than in A and because I think that the loss of the best off is more than compensated for by the gain of the worst off. In other words, I think that we mistake intuitions about the value of equality of welfare with intuitions about priority of the welfare of the worst off. Roughly, the idea is that we should maximise welfare, but gains in welfare matter more, the worse off people are, and losses in welfare matter less, the better off people are. Let us call this idea, following Parfit, the Priority View.[3] Another way to express this intuition is to say that the marginal value of welfare is diminishing: If John has higher welfare than Wlodek, then an extra unit of welfare in Wlodek's life

[3] Parfit's formulation of the Priority view is, however, different from mine: "Benefiting people matters more the worse off people are." See Parfit (1993), p. 57.

increases the value more than an extra unit of welfare in John's life. One achieves this result by applying a strictly concave transformation to the numerical representation of people's welfare.[4] This description of the Priority View is not very exact but precise enough to explain cases, such as the one depicted in Diagram 7.3.1, where the gain of the worst off equals the loss of the best off. Since, according to the Priority View, the marginal value of the gain of the worst off is higher than the marginal value of the loss of the best off, the value of population B is higher than population A. In general, if we are to distribute a fixed amount of welfare among a fixed number of people, the Priority View opts for a completely equal distribution. Consequently, in such cases our beliefs are equally well explained by the Priority View as by appeals to equality of welfare. Moreover, the Priority View implies the Dominance Principle.

One of the adequacy conditions that we have suggested, the Inequality Aversion Condition, seems to involve an appeal to equality. Certainly, people who believe in equality of welfare would endorse this condition. But so would those who believe in the Priority View (they would also, of course, endorse the Non-Elitism Condition). We could have called it the "Non-Anti-Priority Condition" or "Non-Priority to the Best Off Condition". It is, of course, a much weaker condition than the Priority View: The latter implies the former, but not vice versa. The Inequality Aversion Condition is compatible with principles that give much greater weight to the welfare of the best off as compared to the welfare of the worst off. The plausibility of the Inequality Aversion Condition trades, however, on intuitions in the same vein as expressed in the Priority View: It seems implausible that whatever the number of people with very low welfare that we can benefit, this can never outweigh a given decrease in welfare for a given number of people with very high welfare. If this was true, then, in one sense, the welfare of the best off would trump the welfare of the worst off.

Although it is reasonable to give extra weight to the welfare of the worst off, it is, as we pointed out in the previous chapter, unreasonable to prioritise a small increase in the welfare of one slightly bad off person at any cost to the general

---

[4] See Broome (1991), ch. 9. This way of expressing the Priority View 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.

welfare. An acceptable version of the Priority View should satisfy the Non-Extreme Priority and the General Non-Extreme Priority Condition. Unsurprisingly, Monist Welfarist Egalitarianism violates these conditions. Assume that A consists of a very large number of people with very high welfare and one person with slightly negative welfare. In B, we have a small increase in the welfare of the worst off in A but a large decrease in welfare of all the best off in A: everybody has the same very low positive welfare in B. According to Monist Welfarist Egalitarianism, B is better than A. Of course, this would hold even if there was no increase in the welfare of the worst off and everybody else's welfare was decreased to slightly negative welfare. Again, a reasonable Pluralist Welfarist Egalitarian downplays the importance of equality to avoid this kind of conclusion.

I think that the Priority View can explain our beliefs about distribution of welfare in same-number cases better than an appeal to equality. We have not properly shown this, however, since we have only looked at cases which involve comparisons of perfectly equal populations with unequal populations. One also has to consider cases where both of the compared populations involve inequality of welfare. The Priority View can be applied to such cases as it stands, and yields, I think, fairly reasonable answers, whereas the Welfarist Egalitarian has to devise some method of measuring degrees of inequality. There are a number of different suggestions, more or less convincing (mostly less), but it would take us too far away from the main topic to consider them all. Let us instead turn to the last three adequacy conditions to which a Welfarist Egalitarian possibly might object: the Weak Quality Addition, the Non-Sadism, and the Weak the Non-Sadism Condition. These three conditions may involve comparisons of populations of different size. Thus, we have to ask: Could appeals to equality of welfare be decisive in a different number context? Since our intuitions in same-number cases seem to be better explained by the Priority View, it would, I think, be very surprising if appeals to equality of welfare would be decisive in different-number cases. Let's take a look at a case where an egalitarian might object to the Weak Quality Addition Condition:

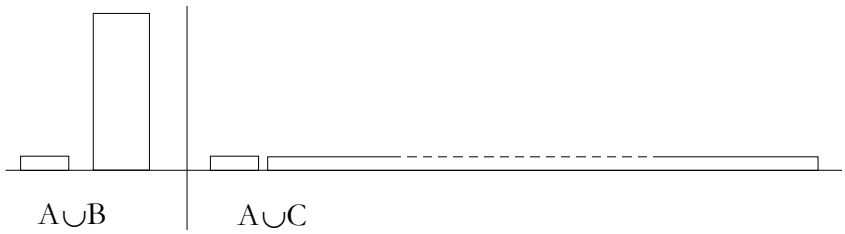

Diagram 7.3.2

According to the Weak Quality Addition Condition, for any population X, there is at least one perfectly equal population with very high positive welfare such that its addition to X is better than any addition of a population with very low positive welfare to X. Assume that population B in Diagram 7.3.2 is such a high welfare population in relation to population A. In population A∪B, B has been added to population A with very low positive welfare. In A∪C, instead of population B, the very large population C with the same very low positive welfare as A has been added. According to the Weak Quality Addition Condition, A∪B is at least as good as A∪C. One might object to this valuation and hold that A∪C is better than A∪B, since there is inequality in the latter population whereas there is perfect equality in the former population.

A Monist Welfarist Egalitarian would rank A∪C as better than A∪B. As we have seen, this view has highly counter-intuitive implications in same-number cases. Its implications in different number choices are no different. Assume that the A-people and the C-people in Diagram 7.3.2 don't enjoy positive welfare but very negative welfare – they all have terrible lives. According to the Monist Welfarist Egalitarian, it would be better to add the terrible C-lives rather than the excellent B-lives since the resulting population of the former addition would be perfectly equal.

A Pluralist Welfarist Egalitarian could avoid this conclusion by, for example, also assigning importance to the total welfare. Can the pluralist give us a reason to discard the Weak Quality Addition Condition? Hardly. Indeed, one can hold that A∪C is in one respect better than A∪B since there is perfect equality in the former population but not in the latter one, but it seems clear that this aspect is outweighed by the greater quality of life in the B-population as compared to the C-population. If a Pluralist Welfarist Egalitarian theory put such a value on equality of welfare that it implied that A∪C is better than A∪B, then that would constitute a good argument against such a theory. An example of the case described in Diagram 7.3.2

could be that the A-people either have children who enjoy very high welfare or that they have more children with the same poor welfare (perhaps because of lack of resources) as themselves. It seems indeed odd that the prospective parents should opt for the latter alternative for reasons of equality. Consequently, a reasonable Pluralist Welfarist Egalitarian would agree with the Weak Quality Addition Condition.

What implications would the Priority View have in regard to cases such as the one depicted in Diagram 7.3.2? As a matter of fact, as we have defined the Priority View, it violates both the Weak Quality Addition Condition and the Quality Condition and implies the Repugnant Conclusion since it ranks populations according to the total sum of people's transformed welfare. But of course, using summing as an aggregation method is as contentious with transformation of individual welfare as without it. We assumed it above just for reasons of simplicity. The core idea of the Priority View – that gains in welfare matter more, the worse off people are, and losses in welfare matter less, the better off people are – can be combined with other aggregation methods, such as, for example, the one used in Average Utilitarianism. Combined with this aggregation method, the Priority View would yield the same results in same-number cases as the ones described earlier, whereas its results in different-number cases would be pretty much the same as those of Average Utilitarianism. The same holds for a combination of the Priority View with the other aggregation methods discussed in the previous chapters. No specific method for aggregating the (transformed) welfare of different lives seems to follow from the core idea of the Priority View and, hence, it is hard to see how this idea could affect our evaluation of different-number cases such as the one discussed here. It seems that the Priority View, like Welfarist Egalitarianism, is an idea mainly about how to distribute welfare among a fixed number of people.

Might a Welfarist Egalitarian object to the Non-Sadism and the Weak Non-Sadism Condition? Let's say that A consists of one person with very high welfare, B consists of one person with negative welfare, and C consists of a large number of people with very low positive welfare. The difference between population A∪B and A∪C is thus that in the former population, a person with negative welfare has been added to A, whereas in the latter population, a large number of people with very low positive welfare have been added to A. Consequently, according to the Non-Sadism Condition, A∪C is at least as good as A∪B. Now, somebody might claim that A∪B is better than A∪C, since it is better in regard to equality of welfare. It is,

of course, by no means apparent how this could be the case. Here we have to compare two populations that both involve inequality. Hence, how to evaluate these populations from a Welfarist Egalitarian perspective all depends on how to measure degrees of inequality. On one measure – the difference in welfare between the best off and worst off – A∪B is worse than A∪C in regard to inequality. But according to another view, entertained by Larry Temkin, what matters, among other things, is the number of the worst off: the greater the number of worst off, the worse the inequality.[5] According to this view, A∪B is better than A∪C in regard to inequality. Curiously, on still another view, proposed by Parfit, the reverse holds true: If the proportion of worst off increases, then the inequality decreases.[6] These examples shows how indecisive appeals to equality are in different-number cases such as these. This is further underscored by the fact that Temkin worries about implications of his theory analogous to (albeit more extreme than) the one discussed here.[7] In other words, if a theory implies that A∪B is worse than A∪C in regard to equality, then that might even be considered as an argument against that particular theory as a theory of equality.

A complete treatment of this issue would include a discussion of all the different methods of measuring inequality and consider whether any of them yield acceptable answers in different-number cases. I'm pretty convinced that none of these methods would stand the test since all of them were originally devised for same-number cases and thus were not intended to be applicable in different-

---

[5] See Temkin (1993a), p. 200-2.

[6] Parfit compares two populations, A+ and Alpha. A+ consists of two groups of people of the same size, one with 100 units of welfare per person, and one with 50 units of welfare per person. Alpha consists of one group of the same size as A+ but with 105 units of welfare per person and a very large group of people with 45 units of welfare per person. He writes: "The inequality in Alpha is in one way worse than the inequality in A+, since the gap between the better-off and the worse-off people is slightly greater. But in another way the inequality is less bad. This is a matter of the relative numbers of, or the *ratio* between, those who are better-off and those who are worse-off. Half of the people in A+ are better off than the other half. This is a worse inequality than a situation in which almost everyone is equally well off, and those who are better off are only a fraction of one per cent. - - - All things considered, the natural inequality in Alpha is not worse than the natural inequality in A+." Parfit (1986), p. 156. Needless to say, I find Parfit's argument indecisive.

[7] Temkin (1993a), pp. 218-27.

number cases.[8] Such an exercise would be pretty tiresome and I don't think it is necessary for our present task. It is hard, if not impossible, to decide which one of populations A∪B and A∪C is better in regard to equality since different egalitarian considerations pull in different directions: There is a bigger gap between the best off and the worst off, and the worst off are worse off in A∪B as compared to A∪C; on the other hand, there is a greater number of worst off in A∪C. Our intuitive *all things considered* ranking of these two populations is, however, pretty robust – intuitively, it seems clear that an addition of lives with negative welfare cannot be better than an addition of people with positive welfare. An argument to the effect that we should give up this intuitive judgement must be very convincing. As we have seen, egalitarian concerns are pulling in different directions and are thus very indecisive in cases such as these. Consequently, egalitarian concerns can hardly give us any reason to change our all things considered ranking of A∪B and A∪C.[9] Clearly, this is even less probable in regard to the Weak Non-Sadism Condition.

It is still an open question whether appeals to equality of welfare are applicable in any interesting sense in different-number cases; we haven't decisively shown that that is not the case. But we have shown that this idea cannot yield convincing arguments against the adequacy conditions which we have proposed. Moreover, we have shown that in many instances our "egalitarian" intuitions in same-number cases can be better explained by the Priority View, and that this view is compatible with, and in some cases implies, the adequacy conditions that we have suggested for

---

[8] See Temkin (1993a) for a detailed discussion of the drawbacks of these different methods. Temkin is, to the best of my knowledge, the only theorist who has made a serious effort to develop a method for comparing equality of welfare in different-number cases.

[9] Temkin would probably not consider A∪B better than A∪C *all things considered*, since he's not a Monist Welfarist Egalitarian. For example, he considers and rejects the following argument directed against his view of equality (Temkin (1993a), p. 217): "…[One] may object that if proportional increases worsen inequality, then proportional decreases should improve it. Thus the egalitarian should favor a *Shrinking World*. More particularly, for any pattern of inequality, the *best* world will be the one with the *smallest* number of people in the better- and worse-off groups consistent with that pattern. This, it may be contended, is absurd. - - - Surely, it *would* be absurd to claim that a two-person world with the same pattern of equality as A and B [two worlds with much more people]…would be better than A and B *all things considered*. - - - Put simply, I am unpersuaded that this objection seriously challenges [my arguments]… - - - Why shouldn't the egalitarian insist that the former worlds are better than the latter one[s] regarding inequality, but admit that they are worse *all things considered*?"

same-number cases. As we shall see in chapter 8, the Priority View can also explain some of our beliefs about different-number cases.

8

Non-Neutral Axiologies

8.1 Introduction

All the theories that we have considered so far satisfy what we could call *Neutrality*:

> *Neutrality*: If there is a one-to-one mapping from population A to population B such that every person in A has the same welfare as their counterpart in B, then A and B are equally good.

A number of theorists have suggested that the crux of the problems in population axiology resides in an all too "impersonal" axiology and that these problems can be solved by a shift to a so-called "person-affecting" axiology. What exactly this distinction amounts to has not been spelled out in the literature. The different theories that have been proposed under the banner of a person affecting axiology, however, are welfarist theories which share the feature that they violate Neutrality. These theories count people's welfare differently depending on the temporal location or the modal features of their lives: *presentists* draw a distinction between presently existing people and non-existing people; *necessitarians* distinguish between people that exist or will exist irrespective of how we act and people whose existence is contingent on our choices; *actualists* differentiate people that have existed, exist or who are going to exist in the actual world, on the one hand, and people who haven't, don't, and won't exist, on the other; and *comparativists* draw a distinction between people that are *uniquely realisable*, that is, people that only exist in one out of two compared outcomes, and those that exist in both of the compared outcomes.[1] These distinctions don't amount to the same thing but there are

[1] The concepts of necessary and contingent persons are from Österberg (1992, 1996), although my definition differs slightly from Österberg. We shall discuss Österberg's theory below. My discussion of Actualism draws on Bykvist (1998), from whom I also got the term "uniquely realisable person".

relations among them. A presently existing person is also a necessary and actual person but not the other way around since necessary and actual people may be located in the past and the future. A necessary person is also an actual person but a future actual person may be contingent on our choice. Assume, for example that a couple is deliberating about whether to have a child and, as a matter of fact, they do decide to have the child (but they could have chosen otherwise). A uniquely realisable person is also a contingent person, but a contingent person is not necessarily uniquely realisable in respect to all pairs of outcomes in a choice situation since she can exist, for instance, in two out of three outcomes.

A strict presentist, necessitarian, actualist, or comparativist, only counts the welfare of present, necessary, actual, or non-uniquely realisable people respectively. Some of the positions advocated in the literature are not of this kind. Rather, according to these theorists, we should only count the positive welfare of present, necessary, actual, or uniquely realisable people, but count the negative welfare of *all* people. In other words, these theorists respect Neutrality in regard to populations with negative welfare. Their reason behind this move is that they try to incorporate an idea called *Asymmetry*: We have no moral reasons to create people with positive welfare, other things being equal, but we have reasons not to create people with negative welfare, other things being equal.

The above distinctions are, regrettably, seldom made explicit in the literature.[2] Rather, these different views are often conflated or mixed in a confusing fashion such that it is hard to determine what position a theorist really holds. Consequently, in most cases my exegetical ambitions, in regard to the theorists to which I ascribe one or another of the above views, are modest. The exposition will be analytical rather than exegetical. We shall state the simple version of each view first, and then proceed to the more complex versions.

In most cases, the motivation behind drawing one or the other of the above distinctions is an idea which goes under the name of the *Person Affecting Restriction*.[3]

---

[2] For an enlightening exception, see Bykvist (1998).

[3] Temkin (1993a, b) claims that this restriction, which he dubs "the Slogan", is presupposed in many arguments in moral philosophy, political theory, and welfare economics. The term "Person Affecting Restriction", introduced by Glover (1977), p. 66 (but see also Narveson (1967)), might be misleading since many theorists would, sensibly I think, lessen the restriction to also include sentient beings. Cf. Holtug (1996). Below, I shall only discuss applications of the Person Affecting Restriction on human populations. Consequently, whenever I claim that a certain

In its slogan form, this view states that an outcome can only be better (or worse) than another if it is better (or worse) for people. From some of the contributions in the literature, one can get the impression that this restriction is supposed to entail one or another of the above distinctions. How this entailment is supposed to work is by no means clear and depends, of course, on how one understands the Person Affecting Restriction. This is what we shall now look at.

8.2 The Person Affecting Restriction

In its slogan form – an outcome can only be better (worse) than another if it is better (worse) for people – the Person Affecting Restriction appears reasonable. It is terribly vague, however, and open to several interpretations. It could be understood as an idea about which kind of objects have moral value, for example, that all moral values are essentially related to the interests of human beings. All moral claims would thus necessarily involve a reference to humans: Outcome A is better than outcome B since people have higher welfare in the former as compared to the latter outcome, or since in the former but not in the latter outcome people's rights are fulfilled, or in the former but not in the latter people have equal opportunities, and so forth. Examples of putative moral claims which are ruled out by this restriction would thus be: Outcome A is better than outcome B since the scenery is beautiful in the former but ugly in the latter outcome, or since the ecosystem is in balance in the former but not in the latter outcome, and so forth. Roughly, this interpretation of the Person Affecting Restriction, which we could call the Human Good Restriction, claims that two outcomes can only differ in value if they differ in regard to some aspect of human goods.[4] This restriction is pretty reasonable and I think that much of the appeal of the Person Affecting Restriction derives from the Human Good Restriction.[5] It is, however, clearly insufficient to yield any kind of distinction between the contributive value of present, necessary,

---

interpretation of the Person Affecting Restriction is reasonable, this claim only holds for human populations.

[4] Perhaps it is this restriction which is at stake in Moore's criticism of Sidgwick at the turn of the century. It can be seen as a denial of Moore's idea in *Principia Ethica* that an unpopulated beautiful world is intrinsically better than an unpopulated ugly world, and a reaffirmation of Sidgwick's view that all moral goods must be of "Human Existence" See Moore (1903), section 50, and Sidgwick (1907), Bk. I, ch. IX, section 4.

[5] Cf., however, fn. 3 above.

or actual people, or people that exist in more than one outcome, on the one hand, and future, contingent, non-actual, or uniquely realisable people, on the other hand. Nor does it imply any kind of asymmetry between lives with positive or negative welfare.

One can give a stronger interpretation of the Person Affecting Restriction than the one given above. One can stress an individualist aspect of value: All moral goods are *personal goods* which, roughly, are non-relational goods, "belonging to" or "located in" individuals. Another way to put it is to say that personal goods are intrinsic properties of individuals.

Consider the following two outcomes: In A, Krister and Erik are equally happy. In B, they are both happier than in A but Krister is happier than Erik. As we noticed in the previous chapter, an egalitarian might argue that B is worse, or at least in one respect worse, than A, since although both Erik and Krister are better off in A than in B, B involves inequality whereas there is perfect equality in A. One might say that B is worse in regard to one aspect of human goods, namely its distribution. "Worse for whom?" some theorists ask rhetorically. Perhaps they endorse a reading of the Person Affecting Restriction, which we could call the Personal Good Restriction, to the effect that an outcome cannot be worse than another, if it isn't worse in regard to personal goods.[6]

The egalitarian concern above is grounded in a relational good: What is bad about outcome B is that one person is worse off than another person. Consequently, this concern is ruled out by the Personal Good Restriction. Since B is not worse than A in respect to personal goods, B cannot be worse than A. In other words, if we find this restriction plausible, then we have another reason for rejecting Welfarist Egalitarianism.[7] The Personal Good Restriction, however, neither implies any value distinctions based on temporal or modal properties of lives, nor implies the Asymmetry. It is compatible with such distinctions: One might decide, perhaps on purely intuitive grounds, that only personal goods

---

[6] I have taken the term "personal good" from Broome (1991), ch. 8. The Personal Good Restriction is not, however, equivalent to his principle of personal good.

[7] As we pointed out in the previous chapter, many intuitions that on the surface look like egalitarian concerns can be captured by the Priority View which is compatible with the Personal Good Restriction. Broome (1991), pp. 180-1, suggests a way of understanding the goodness of equality that turns it into a personal good.

belonging to actual people count. It is, however, equally compatible with principles which don't distinguish between actual and possible people. Total Utilitarianism, for example, entails the Personal Good Restriction.

The next step to take is to stress the individualist aspect of value even more by claiming that morality is essentially *person comparative*: If an outcome is better (worse) than another, then it is better (worse) for at least one person. We shall formulate this view with a little bit more content:

> *The Person Affecting Restriction*
> (a) If outcome A is better (worse, equally as good) than (as) B, then A is better (worse, equally as good) than (as) B for at least one individual.
> (b) If outcome A is better (worse) than B for someone but worse (better) for no one, and B is better (worse) than A for no one, then A is better (worse) than B.

This is the principle that I shall henceforth refer to as the Person Affecting Restriction. In cases involving only necessary people, this view is not very controversial. In cases involving contingent people, however, this restriction is ambiguous. An outcome A is better than B for Peter if Peter has, for example, higher welfare in A as compared to B (we are, of course, assuming that if a person has higher welfare in one population as compared to another, then the former population is better for that person, other things being equal). But what if Peter exists in outcome A but not in outcome B? Is outcome A then better than outcome B *for Peter*? This is the crux of the matter. Depending on the answer to this question, different versions of the Person Affecting Restriction result. We shall soon look at some possible answers, but let us first show that irrespective of which answer one gives to this question, the Person Affecting Restriction neither entails any version of Actualism, nor Necessitarianism, nor Presentism.

Assume that in population A Peter enjoys 10 units of welfare whereas in B he enjoys 5 units of welfare, and that these populations are equal in all other respects. The Person Affecting Restriction would rank A as better than B since Peter is better off in A as compared to B. This holds irrespective of whether Peter is an actual, non-actual, necessary, contingent, present, or future person. Compare with, for example, a strict presentist. Since such a theorist only counts the welfare of present people, she would rank A and B as equally good, or perhaps as

incommensurable, if Peter was a future person. Consequently, the Person Affecting Restriction doesn't imply that the welfare of non-actual, contingent, and future people counts differently as compared to the welfare of actual, necessary, and present people respectively.

What about Comparativism? One possible answer to the question whether existence can be better or worse for a person is to claim that non-existence is neither better, nor worse, nor equally good as existence for a person: Non-existence and existence are incomparable in value for a person. This answer in combination with the Person Affecting Restriction yields a version of the *Comparativist* view: We should disregard the welfare of *uniquely realisable* people, that is, people that only exist in one out of two compared outcomes. At times, this appears to be David Heyd's view. He holds that "… the very comparison of the welfare of two possible children is based on the fallacious notion of an abstract, impersonal quantity of happiness in the world which should be maximized" and argues against the Asymmetry by claiming that it "is inconsistent with a person-affecting theory as it presupposes the comparability of non-existence with life of a certain quality". He thinks that we can solve the problems in population axiology "… by simply rejecting the logical legitimacy of comparisons between the welfare of a possible population A and a possible population B (when they consist of *different* people)".[8]

This version of the Person Affecting Restriction, taken as a population axiology, is inconsistent. Assume that the x- and y-people exist in outcome A, the y- and z-people exist in B, and the z- and x-people exist in C. Assume that all of these people have positive welfare, but that the y-people are better off in B as compared to A, the z-people are better off in C as compared to B, and the x-people are better off in A as compared to C.[9]

---

[8] Heyd (1988), pp. 159 - 61, emphasis in original. The logic of Heyd's reasoning is not completely clear to me. He claims that his view is "grounded in an 'anthropocentric' conception of value according to which value is necessarily related to human interests, welfare, expectations, desires and wishes – that is to say to human volitions" (p. 164). How this "volitional concept of value" is supposed to generate the conclusion that "[e]xcluding the welfare and interest of future merely possible person … is a necessary consequence of a coherent person-regarding theory of value" (p. 161) is not spelled out in clear fashion by Heyd. As I pointed out above in the discussion of the Human Good Restriction, I'm sceptically inclined towards the validity of such deductions.

[9] A similar example is used by Temkin (1987), pp. 168-9, to illustrate the intransitivity of the Person Affecting Restriction.

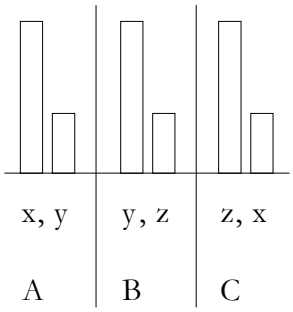

Diagram 8.2.1

Since the x-people don't exist in B, B is neither worse nor better than A for them. Similarly, since the z-people don't exist in A, A is neither worse nor better than B for them. However, B is better than A for the y-people. Consequently, B is better than A according to the second clause of the Person Affecting Restriction. The same reasoning yields that C is better than B, and A is better than C. But if B is better than A, and C is better than B, then transitivity yields that C is better than A. Consequently, C is both better and worse than A.

Perhaps an adherent of this version of the Person Affecting Restriction could argue that we should abandon transitivity of the relation "is better than". Apart from the counter-intuitive implications of this move, it wouldn't help much since there are other problems ahead. Consider the following case:

> The Energy Policy Case: A country is facing a choice between implementing a certain energy policy (alternative A) or not (alternative B). Were this country to implement this policy, then there would be a marginal increase in the welfare of the present people of this country (the x-people). On the other hand, this increase would be greatly outweighed by the misery the waste from this energy system will cause in the lives of people in the future (the y-people). The existence of these future people is contingent upon the implementation of this energy policy. If the country doesn't implement this energy policy, other people will exist in the future with very good lives (the z-people). The advantages and disadvantages of other effects of this policy balance out.

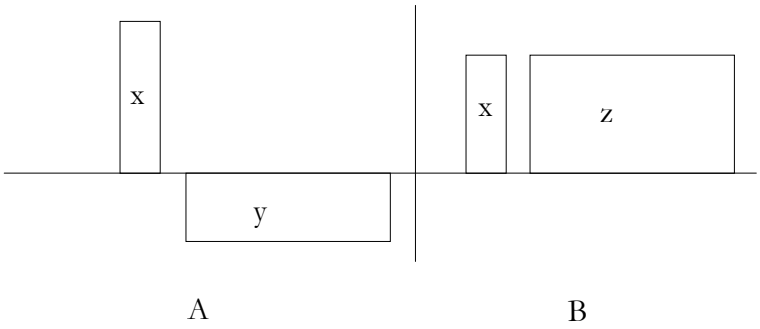

Diagram 8.2.2

Most of us, I presume, would consider outcome B clearly superior to outcome A and, since the cost to present people is marginal, we ought to realise B rather than A. Moreover, to rank A better than B would be a flagrant violation of the Inequality Aversion Condition (given the assumption that A and B are of the same size).[10]

According to the Comparativist version of the Person Affecting Restriction, A is incomparable in value to B for all the y- and z-people, since they are uniquely realisable people. Consequently, outcome A is neither better nor worse for the y- and z-people as compared to B. Outcome A is slightly better for the x-people, however, and consequently, this version of the Person Affecting Restriction ranks A as better than B. But that is clearly the wrong answer to the Energy Policy Case. As a matter of fact, Heyd's view has the intriguing feature of yielding results that violate *all* the adequacy conditions that we have proposed since it renders all pairs of populations consisting of different people incommensurable.

Incorporating some kind of asymmetry into this version of the Person Affecting Restriction wouldn't help much since we can just restate the Energy Policy Case such that it only involves people with positive welfare: Just assume that the y-people have very low positive welfare.

Another possible answer to the question whether existence can be better or worse for a person is to claim that non-existence is equally good for a person as existence. A slightly less implausible answer introduces an asymmetry in "good for"

---

[10] If we assume that B is only slightly worse for *one* of the x-people, then it would be a clear violation of the Non-Elitism Condition to rank A better than B.

and claims that non-existence is equally good for a person as existence with positive welfare, whereas non-existence is better than existence with negative welfare. Neither of these are very tempting answers and they run into the same problems as the ones discussed above, so I shall say no more about it. Let me instead return to a topic that we discussed in section 2.2.3. As we noticed there, the reason why theorists have been inclined to deny that it can be better or worse for a person to exist than not to exist is that they believe that this position implies that it can be better or worse for a "person" not to exist than to exist and they find this implication nonsensical or even absurd. As we discussed, one can deny this implication, and hold that it can be better or worse for a person to exist than not to exist, but it cannot be better or worse for a "person" not to exist than to exist. Given this position, one can claim that it is better (worse, equally good) for a person to exist with positive (negative, neutral) welfare than (as) not to exist without implying any absurdities. This answer to the question whether existence can be better or worse for a person yields a version of the Person Affecting Restriction which doesn't have any of the disagreeable implications of the versions discussed above. This version won't, however, have any force. If we, for example, combine this trivial version of the Person Affecting Restriction with Total Utilitarianism, we will just get a restatement of this principle in a person affecting form which is extensionally equivalent with the original formulation:

> *A Person Affecting Version of Total Utilitarianism*: An outcome A is better (worse, equally good) for people than another outcome B if and only if the total sum of people's benefits is higher (lower, the same) in A as compared to B.

Since this principle is extensionally equivalent to Total Utilitarianism, it implies the Repugnant Conclusion: The B-people in Diagram 3.1.1 receive *together* a greater benefit than the A-people, although each individual in B receives a smaller benefit than each individual in A.[11] In other words, with this version of the Person Affecting Restriction we haven't gained any ground in our search for a reasonable population axiology. Let's turn to Presentism.

---

[11] See Parfit (1984), pp. 394-96 for the same argument.

8.3 Presentism

In his early pioneering work in population ethics, Jan Narveson sounds like a presentist. He claims that "moral questions *presuppose* the existence of people" and that "[w]e are in favor of making people happy, but neutral about making happy people."[12] Indeed, since Heyd says that his approach "allows … only for considerations relating to the welfare and interests of existing persons …", perhaps he's a presentist after all.[13]

A strict presentist only counts the welfare of presently existing people. We shall render this position precise as follows:

> *Strict Presentism*
>
> (a) If population A is better (worse) than B in regard to the welfare of present people, then A is better (worse) than B; and if A and B do not differ in regard to the welfare of present people, then A and B are equally good.
>
> (b) If all the present people are equally well off in A, and better off in A as compared to B, then A is better than B in regard to the welfare of present people.

The first clause (a) in the above formulation expresses the core idea of Strict Presentism. Notice that our formulation is compatible with a wide range of methods for determining better, worse, and equally good in regard to the welfare of present people, such as summing, averaging, and so forth. We shall assume, however, that all these methods yield results that respect Neutrality in cases involving only present people. We also included a weak dominance condition (b) in our formulation. It is a pretty innocuous condition and I'm sure that all presentists would agree with this condition. At any rate, if Presentism didn't entail this condition, then we could dismiss this view out of hand for it would violate the Egalitarian Dominance Condition in cases involving only presently existing people

---

12 Narveson (1973), pp. 73, 80, emphasis in original. As we shall see below, Narveson isn't a strict presentist.

13 Heyd (1988), p. 161.

(the same reasoning holds for our formulation of Actualism and Necessitarianism below).

Can Presentism solve the problems in population axiology? At first glance, it looks like this view cannot escape any of the problems facing neutral theories. Assume that A is a population of present people (for example, the present people in country *x*) enjoying very high welfare and that B is a larger population of present people (for example, the present population in country *y*) with very low welfare but with higher total welfare than A. Which population is the better one according to Presentism? It all depends on which method the presentist selects for determining better, worse and equally good in regard to the welfare of present people. Since Presentism respects Neutrality in regard to populations of present people, all the problems discussed in the previous chapter will reappear in the search of this method. Consequently, Presentism is not an advance towards a satisfactory population axiology.

Obviously, the proponents of Presentism think they have added something to the discussion. Their idea might be that axiological evaluations are only important insofar as they are relevant for evaluations of actions, or that moral evaluations per definition are about outcomes of actions only (in contrast to, for example, aesthetic evaluations).[14] In other words, it seems that presentists are not interested in orderings of all possible populations but only in orderings of populations that are outcomes or parts of outcomes of alternative actions. And of course, the present populations of two countries are not outcomes of alternative actions since these populations belong to the *same* outcome of the *same* past action(s) or event(s). Analogous reasoning holds for Actualism and Necessitarianism. Seen as population axiologies proper, these theories cannot avoid the problems discussed in the previous chapters. This is, of course, a serious drawback of these theories. But if we only consider populations that are outcomes of possible alternative actions, then

[14] Although I agree that an acceptable axiology should be action-guiding in some sense, I'm sceptical about a moral/non-moral distinction between evaluations of populations that are outcomes of alternative actions and those that are not. If evaluations of the goodness or badness of historical events, or the present state of different countries, or of possible natural catastrophes, and so forth, are not moral evaluations, what are they then? Aesthetic? Or do they belong to a category of their own? And wouldn't it be a decisive case against an axiology if it ranked a past population consisting of miserable people as better than another past population consisting of people with very good lives?

these theories might have something to add to the discussion. This is how we shall understand the axiological presentist, actualist, and necessitarian's project: They try to develop a *partial* axiology which covers axiological evaluations that are normatively relevant, namely the evaluations of populations which are relevant for determining the normative status of actions. Their goal is to develop an axiology that orders populations that are outcomes, or parts of outcomes, of alternative actions in possible choice situations. This ordering is used to determine the normative status of actions in conjunction with a bridging principle, such as some form of consequentialism. Although this project cannot solve all the problems in population axiology, it is an important task, or, as some people might argue, *the* important task in population axiology, since inconsistency in normatively relevant evaluations is, arguably, more disturbing than in other areas of population axiology.[15]

It might be that Presentism, Actualism, and Necessitarianism are normative theories in an axiological disguise.[16] Most contributions in this area involve a mix of axiological and normative statements (in a not very clear fashion), so both interpretations are possible. Interpreted as normative theories, the restriction to outcomes of alternative actions is perfectly legitimate. Since a normative theory concerns what we ought to do, and since "ought" implies "can" (in some sense), a normative theory is not in the business of ordering populations that are not outcomes of alternative actions in some choice situation. However, since there are some theories that are clearly axiologies (for example, Österberg's theory below), and since the contributions that involve normative statements are set in a consequentialist framework (for example, Narveson's and Singer's Utilitarianism), we have a subject for the present chapter. We shall postpone the discussion of the putative differences between axiological and normative population theories until chapter 11.

Let's return to our discussion of Presentism. How does Strict Presentism fare in regards to the Energy Policy Case? Not very well, I'm afraid. Since the ranking of the two outcomes is determined only by the welfare of the present people, Strict

---

15 Whether or not this is true involves some interesting and tricky meta-ethical questions which unfortunately fall outside the scope of this essay.

16 See Bykvist (1998) for a discussion of preferentialist versions of these views formulated as normative principles.

Presentism selects outcome A, that is, implementing the Energy Policy which yields a slight benefit for the present people at the expense of future people's misery. Because this holds irrespective of the number of the *y*-people, Strict Presentism entails violations of the Inequality Aversion Condition even in cases where the gain for the worst off is great and the loss for the best off is negligible.

What about cases involving only future people with positive welfare? A presentist ranks all pairs of populations consisting of only future people with positive welfare as equally good.[17] Our choice is just, as Narveson seems to mean, "a matter of taste":

> … [W]e 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.[18]

Unfortunately, Strict Presentism not only ranks different sized populations with positive welfare as equally good. It also ranks any populations of future people with *very negative* welfare as equally good as a same sized population of future people with

---

[17] Alternatively, a strict presentist might claim that such populations are incommensurable in regard to moral value. In the cases that we shall discuss below, this version of Strict Presentism has the same problematic implications as the one we have stated above. The same remark is true for the statements of Actualism and Necessitarianism below.

[18] Narveson (1973), pp. 72, 80 (emphasis added).

the same *very high positive* welfare. In other words, Presentism violates the unassailable Egalitarian Dominance Condition.

One might object here by pointing out that Strict Presentism doesn't violate the Egalitarian Dominance Condition if we further restrict its scope. One could claim that a presentist, as well as not being interested in rankings of populations which do not belong to alternative outcomes, is not interested in ranking populations which are only parts of alternative outcomes. Rather, the presentist is only interested in ordering populations that consist of all the people that will ever live or all the present and future people that will ever live, that is, total outcomes or total future outcomes of actions. Since the present people are going to be part of any population understood in this manner, Strict Presentism doesn't violate the Egalitarian Dominance Condition. If all the (present and future) people that will ever live are better off in A as compared to B, then it follows that the present people also are better off in A than in B, and Strict Presentism ranks A as better than B.

Of course, Strict Presentism would still violate the Egalitarian Dominance Condition in what Partha Dasgupta calls "genesis problems", that is, in cases where no people yet exist and thus all people are future people.[19] One might find such counter-examples to a theory moot, however. More to the point, it is unclear why we should, even if we are only concerned with normatively relevant evaluations, *only* be interested in rankings of whole world histories or futures. Assume that the performance of some action would exclusively affect people who will live 300 years from now. If this action was performed, then these future people would enjoy a much higher welfare as compared to the welfare they would enjoy if it wasn't performed. It seems clear that the only relevant fact in determining the goodness of this action is the ranking of these two alternative future populations. In other words, we would only compare these two future populations when deciding whether to perform the action, not the two world histories of which these populations are a part. Moreover, Presentism would still fall prey to a slightly reformulated version of the Egalitarian Dominance Condition which seems equally as plausible as the original version: If population A is a perfectly equal population of the same size as population B, and every person whose welfare differs in A and

[19] Dasgupta (1988), p. 110.

B has higher welfare in A than in B, then A is better than B, other things being equal.[20]

An asymmetrical presentist only counts the positive welfare of present people but counts the negative welfare of *all* people. Narveson sounds like he is an asymmetrical presentist since he claims that "[i]f you bring people into existence, then of course you must treat them in accordance with their moral status as human beings. And if you can foresee … that no matter how much you or anyone tries, you won't be able to succeed in enabling them to live a worthwhile life, then that is a reason for not starting on the project in the first place."[21]

Asymmetrical Presentism avoids some of the most absurd implications of Strict Presentism. The former but not the latter position implies violations of the Egalitarian Dominance Condition only in cases involving comparisons of populations with positive welfare. On the other hand, Asymmetrical Presentism violates the Non-Extreme Priority Condition in more cases than Strict Presentism. Since no weight is given to the positive welfare of future people, any improvement in the welfare of a person with negative welfare can outweigh any decrease in positive welfare among any number of future people. In cases where the person with negative welfare is presently existing, Strict Presentism violates this condition too.

One doesn't need to be a strict or asymmetrical presentist. It is possible to hold a view that assigns some weight to the positive welfare of future people but greater weight to the welfare of present people. Since this is a more reasonable position than Strict Presentism, perhaps it is something like this that the presentists in the literature had in mind. Let's call this view Soft Presentism (again, this view can be combined with some kind of asymmetry, but such a modification makes no difference for the present discussion). For example, one can give lexical priority to the welfare of present people: If A is better (worse) than B in regard to the welfare of present people, then A is better (worse) than B; but if A is equal to B in regard

---

[20] Analogous reasoning also holds for our discussion of Actualism and Necessitarianism in connection with the Egalitarian Dominance Condition. Notice also that the above narrowing of Presentism's scope doesn't save it from the violations of the Minimal Inequality Aversion Condition discussed earlier, and the violations of the Non-Extreme Priority Condition discussed below.

[21] Narveson (1973), p. 76.

to the welfare of present people, then A is better (worse, equally as good) than (as) B if and only if A is better (worse, equally as good) than (as) B in regard to the welfare of non-present people. In one of his later papers on population ethics, Narveson seems to have an asymmetrical version of the lexical view in mind, although he is expressing it in deontic terms:

> (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. 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.[22]

If we explicate "better in regard to the welfare of future people" in a reasonable manner, the lexical version of Soft Presentism, both in its symmetrical and asymmetrical guise, wouldn't violate the Egalitarian Dominance Condition. Because of its lexical properties, it would violate the Inequality Aversion Condition and the Non-Extreme Priority Condition in cases such as the ones described above. One could formulate non-lexical versions of Soft Presentism which satisfy these conditions too, but we shall not pursue this matter further since there is a problem shared by all versions of Soft Presentism. In cases where the compared populations are equally good in respect to the welfare of present people, Soft Presentism's ranking of the involved populations will be completely determined by how good they are in respect to the welfare of future people. Consequently, all the problems discussed in the previous chapter will reappear in the specification of the method for determining better, worse and equally good in regard to the welfare of future people: Summing doesn't satisfy the Quality Condition; averaging doesn't satisfy the Non-Sadism Condition, and so forth. Partha Dasgupta's proposal of a

[22] Narveson (1978), p. 55-56, emphasis in original.

"generation dependent morality" is a case in point. He suggests that the "goodness of states of affairs is conditional upon who exists" and that we should give greater weight to the welfare of existing people: "Suppose for concreteness that the living standard of actual lives count for twice that of potential living standards and the evaluation of alternative social states is based upon the *weighted* sum of individual living standards…".[23] Now, in all cases where the welfare of the present people is not affected, Dasgupta's theory determines the ranking by the total sum of the future people's welfare.[24] Consequently, like Total Utilitarianism, it violates the Weak Quality Addition Condition. Indeed, it implies the Repugnant Conclusion even in cases that involve great losses in the welfare of present people. Assume that k is a positive finite number that represents the extra weight given to the welfare of present people (if it weren't a finite number, Dasgupta's theory would be extensionally equivalent to the lexical version of Soft Presentism discussed above). For any population of n present people with very high welfare u_1, there is a mixed population of m present and future people with very low positive welfare u_2 such that $nku_1 < mu_2$, namely a mixed population consisting of $m > nku_1/u_2$ people with welfare u_2. In other words, Soft Presentism in general, and Dasgupta's theory in particular, doesn't constitute any kind of advance towards a satisfactory population axiology.

8.4 Actualism

A strict actualist only counts the welfare of people that have existed, exist, or will exist. For example, John Bigelow and Robert Pargetter suggest that we should "… bring about the outcome which is of the greatest value for the totality of *actual* (past, present and future) moral agents. But we do not need to bring about the outcome which would be of greatest value for the totality of moral agents that there

---

[23] Dasgupta (1988), p. 120, emphasis in original. Although Dasgupta uses the term "actual lives", I think it is correct to describe him as a presentist since he writes (p. 117) that "[i]n an Actual Problem there are actual people – existing persons whom I shall call the current generation here – who deliberate over future population sizes and future living standards".

[24] To be fair to Dasgupta, his theory is not a simple as I have described it above. It also involves a two-step procedure for handling alternatives that involve more than two options, and a backward induction procedure for handling inconsistencies over time in the ordering of outcomes. Since none of these features of come into play in the cases discussed above, I shall not dwell on them here.

would be if we brought it about." Moreover, from this idea "…we should conclude that there is no basis for the protection of merely potential persons, in a host of situations that have worried utilitarians, such as contraception and early abortion".[25] Mary Warren claims that "… the prima facie aim of morality should be to maximise the extent to which each *actual* – present or future – person's interests are promoted. - - - Each person's interest must be given prima facie equal weight; but it is only those who do or will exist, who can possibly have interests to be weighed." She concludes "that in most cases no moral justification at all is required for the decision to remain celibate, use contraceptives, or to have an abortion" and that her theory "… has an important bearing upon our long-term population policies …".[26] We shall define Strict Actualism as follows:

Strict Actualism

(a) If population A is better (worse) than B in regard to the welfare of actual people, then A is better (worse) than B; and if A and B do not differ in regard to the welfare of actual people, then A and B are equally good.

(b) If all the actual people are equally well off in A, and better off in A as compared to B, then A is better than B in regard to the welfare of actual people.

The implications of Actualism in population axiology are analogous to Presentism. Consider the following version of the Energy Policy Case. Assume that there is a third outcome C consisting of the x-people and some other people, different from the y- and z-people, and that outcome C is what actually will be the case. Consequently, the y- and z-people are non-actual people. Since the *x*-people –

---

[25] Bigelow and Pargetter (1988), pp. 180-1, emphasis in original. Bykvist (1998), pp. 94-5 ascribes the actualist position to Bigelow and Pargetter. Although their theory certainly includes actualist considerations, it sometimes looks like they are entertaining a view which is a combination of Actualism, Presentism, and Necessitarianism (p. 180): "In deciding what act to perform, morality requires that we consider the value the resulting world has for all present, actual agents. But we do not need to consider the value of that world for non-actual agents, or for agents whose existence depends on whether we perform the action or not."

[26] Warren (1978), pp. 24, 16, emphasis in original. Since Warren seems to exclude the interests of past people, strictly speaking she is not a strict actualist according to our definition. This difference has no relevance for the discussion below.

the actual people – are slightly worse off in B as compared to A, B is worse than A according to Strict Actualism although the z-people are much better off in B than the y-people in A. Since this holds irrespective of the number of the *y*-people, Strict Actualism entails violations of the Inequality Aversion Condition even in cases where the gain for the worst off is great and the loss for the best off is negligible. Again, we have a principle that gives the wrong answer in the Energy Policy Case.

But perhaps an Actualist can answer this objections in the following way: Why should we at all care about rankings of populations involving non-actual people? These are, after all, people and populations that will never exist. True as this is, it still seems reasonable that an axiology should order not only possible outcomes consisting only of actual people but also possible outcomes involving non-actual people. More to the point, if we don't need to rank possible outcomes involving non-actual people, why should we care about ranking possible non-actual outcomes consisting of actual people? After all, those outcomes are never going to be actual.

A particular problem for Actualism is that it, in combination with some form of consequentialism, makes the normative status of an action dependent on whether the action itself is actually performed. Consider the Energy Policy again with a slight variation: Imagine an outcome B without the z-people. Now, if we actually implemented the Energy policy (alternative A), then we ought not to have implemented it, since the negative welfare of the actual people in the future outweighs our slight increase in welfare. But if we didn't implement it, then we ought to have implemented it since that action would have increased our welfare and thus the welfare of the only actual people in this case. But the choice situation in these two possible worlds do not differ in any relevant respects apart from the actuality of us implementing the energy policy – we face the same choice in both of these situations.[27]

Let's consider populations which only contain non-actual people. Analogously to Strict Presentism, Strict Actualism ranks all pairs of population consisting of non-actual people with positive welfare as equally good. Consequently, Actualism violates the Egalitarian Dominance Condition. Strict Actualism violates the Non-Extreme Priority Condition too, in a manner analogous to Strict Presentism – just replace the present people with actual people in the case discussed in connection

[27] For the same argument, see Bykvist (1998), pp. 103-4. Cf. Carlson (1995), chs. 5 and 6.

with Presentism. Since the cases above don't involve people with negative welfare, the same conclusions hold true for Asymmetrical Actualism. As with Presentism, there is a soft version of Actualism, that is, one which gives weight to the welfare of both actual and non-actual people but greater weight to the former. Again, with this version of Actualism, all the problems discussed in the previous chapter will reappear in the specification of the method for determining better, worse and equally good in regard to the welfare of non-actual people.

8.5 Necessitarianism

In his famous book *Practical Ethics*, Peter Singer suggests that when we are deliberating over a decision, we shall only "count … beings who already exist, or at least will exist independently of that decision" and he "denies that there is value in increasing pleasure by creating additional beings".[28] This sounds like Necessitarianism: We should give priority to the welfare of people who will exist irrespective of how we act. Let's call a person a *necessary person,* relative to a set of all alternative populations in a choice situation*,* exactly if she exists in all alternative populations. A person is a *contingent person* exactly if she exists in some but not all alternative populations. We shall define the strict version of Necessitarianism as follows:

Strict Necessitarianism

(a) If population A is better (worse) than B in regard to the welfare of necessary people, then A is better (worse) than B; and if A and B do not differ in regard to the welfare of necessary people, then A and B are equally good.

(b) If all the necessary people are equally well off in A, and better off in A as compared to B, then A is better than B in regard to the welfare of necessary people.

One can distinguish between two kinds of social choice or policy options: choices that involve only necessary people and choices that also involve contingent

[28] Singer (1993), pp. 103-4.

people.[29] The first kind of choice does not affect the identity of people – the same people exist in all the possible populations. In other words, all the people involved in this kind of choice are necessary people. The typical problem of this kind is how to distribute goods among a given group of people. A number of decisions also affect the identity of people. Most obviously, decisions about having children affect the identity of the people that will exist: If one has a child at a certain point, then a person will exist who wouldn't have otherwise existed. The population which results from this decision contains one person who is not part of the population that would result if one decided not to have a child. Major social decisions which affect the welfare of future generations also affect the identity of the people who are going to exist.[30] It follows that if a social policy is put into effect, there will exist people who would not have existed had the policy not been adopted. After several generations, it is likely that no one alive would have existed otherwise. Consequently, the future populations which are at stake when deciding whether to implement a major social policy are made up of different people. In other words, all of these future people are contingent.

In like manner to the views discussed above, Strict Necessitarianism ranks all pairs of population consisting of contingent people with positive welfare as equally good. Consequently, Strict Necessitarianism violates the Egalitarian Dominance Condition. Strict Necessitarianism violates the Inequality Aversion Condition and the Non-Extreme Priority Condition too, in a manner analogous to Strict Presentism – just replace the present people with necessary people and the future people with contingent people in the case discussed in connection to Presentism. Likewise for the asymmetrical version of Necessitarianism.

One can also be a Soft Necessitarian. The same objection pertains to this view as to Soft Presentism and Soft Actualism: the problems afflicting weakly anonymous theories reappear. Here we have an actual example since Jan Österberg

---

[29] Cf. Parfit (1984), p. 356, for a similar distinction.

[30] Two very plausible claims support this conclusion. Firstly, the identity of a person is dependent on her genetic make-up which in turn depends on who conceived her and the timing of her conception. Secondly, the implementation of a social policy will affect when and by whom a person is conceived. This could happen through a number of perhaps minor but widespread and cumulative effects on people's lives and in a purely accidental way (I recommend any reader who doubts this to inquire whether she or he would have been around to doubt this had the First World War never occurred).

has proposed such a theory which he calls Pessimism Utilitarianism. It is based on the following principles:[31]

(a) 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.
(b) 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.
(c) The negative intrinsic value of a world V is the sum of the unhappiness of the unhappy beings.
(d) The intrinsic value of a world is the positive intrinsic value minus the negative intrinsic value.

As the concern for contingent people's welfare is very limited in this theory, it violates the Egalitarian Dominance Principle. Assume that a population A consists of people with very low positive welfare. Population B consists of the same people as in A but with very high positive welfare, equally shared by all. Population C is empty. Since C contains no contingent people with positive welfare, $n = 0$, and since C is empty the people in A and B are also contingent people relative to the set of alternatives. Consequently, Pessimism Utilitarianism ranks A and B as equally good.

Not surprisingly, Österberg's theory violates the Non-Extreme Priority Condition. Assume that population A consists of a large number of people with very high positive welfare and one person with slightly negative welfare. Population B consists of the same people as in A but all with slightly positive welfare. Population C consists of one person with negative welfare. Since C contains no contingent people with positive welfare, $n = 0$. Consequently, the value of A is determined only by its negative welfare and A will be ranked as worse than B.

Perhaps Österberg would find this acceptable since his theory is supposed to be a negativist theory, that is, a theory that gives more weight to unhappiness than to happiness. His theory has, however, implications that are especially odd from a

[31] Österberg (1992).

negativist point of view. It implies what we might call the Very Repugnant Conclusion: Assume that we have a large population of people with very 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.

Österberg has also presented a new but very incomplete version of his Pessimism Utilitarianism.[32] Since it is an incomplete theory, it is hard to derive any conclusions from it. It says enough, however, for us to conclude that it has at least two counter-intuitive implications.

The restricted Pessimism Utilitarianism states that when we consider "mere additions" of happy people, only the average welfare of necessary people matters.[33] Consequently, Österberg's restricted theory violates the Inequality Aversion Condition. Assume that in A, we have a number of very happy necessary people and a large number of contingent people that are slightly happy. In B, the necessary people are *slightly* less happy whereas the contingent people are *much* happier than in A since they are now equally as happy as the necessary people. In C, only the necessary people exist and they enjoy a low level of happiness. The restricted Pessimism Utilitarianism yields that A is better than B since the average happiness of the necessary people is higher in A than in B although the gain for the worst off is great and the loss for the best off is negligible in the move from A to B.

Österberg applies the "Average View" to populations that only involve contingent happy people.[34] Consequently, like Average Utilitarianism, his new theory violates the Quality Addition Principle. Assume that we have two populations of contingent people: A∪B and A∪C, $N(A)=n$, $N(B)=k$, $N(C)=m$. Assume that $u_1 < u_2 < u_3$ represents three very low positive welfare levels. Let the welfare level of all the lives in A be u_1 and in C u_3. For any population B consisting of k lives with very high welfare, there is a an n such that $AU(A\cup B) < u_2$. Moreover, for any n, there is an m such that $AU(A\cup C) > u_2$. In such cases, the addition of the population with very low welfare (C) is better than the addition of the population with very high welfare (B) according to the "Average View".

---

[32] Österberg (1996).
[33] Österberg (1996), p. 100.
[34] Österberg (1996), p. 104.

8.6 Asymmetry

According to *Asymmetry,* we have no moral reasons to create people with positive welfare, other things being equal, but we have reasons not to create people with negative welfare, other things being equal. On the axiological level, this amounts to claiming that additional lives with positive welfare have neutral contributive value or introduce incomparability among populations, whereas additional lives with negative welfare have negative contributive value. Österberg, for example, states that he "… cannot see … that it would be better that one happy person existed than that no sentient beings existed. - - - I find it equally obvious … that it would be worse that one unhappy person existed than that no sentient beings existed at all."[35] We can formulate the axiological version of Asymmetry as follows:

> *The Asymmetry Principle*: Adding a life with positive welfare neither makes a population better nor worse, other things being equal. Adding a life with negative welfare makes a population worse, other things being equal.

Is this a convincing condition? I'm sceptical. Is it counter-intuitive to claim that, other things being equal, we make a population better by creating an extra person with very high welfare? It is important to keep in mind that other things are equal, that is, we are comparing two populations which only differ in respect to the one person with very high welfare. In other words, we are not considering cases where the creation of extra people would have detrimental effects on the welfare of already existing people, nor are we considering cases where we could relieve suffering of already existing people instead of using our scarce resources on new people. And isn't it such cases that we have in mind when we are questioning whether extra people make a population better?[36]

At any rate, there is a clear-cut reason for abandoning the Asymmetry Principle. Consider the following two populations: A consists of a number of people with very low positive welfare and B is a population of the same size as A but made up

---

[35] Österberg (1996), p. 97.
[36] See Glover (1977), p. 70, and Bykvist (1998), p. 123, for the same point.

of people with the same very high welfare. If we so fancy, we can assume that the A- and B-people are future, contingent and non-actual people. In other words, we have a choice of either adding the A-people or the B-people. According to the Asymmetry Principle, A and B are equally good or incomparable. According to the Egalitarian Dominance Principle, B is better than A. Hence, any theory that implies the Asymmetry Principle is going to violate the Egalitarian Dominance Principle. One can show that the Asymmetry Principle also violates the Non-Extreme Priority Condition.

We have to jettison this principle. I think that one of the motivating ideas underlying Asymmetry has to do with the weight of suffering: It is more important to relieve suffering than to increase (already happy people's) happiness. We can retain this important intuition underlying Asymmetry (perhaps the main intuition underlying it) by giving more weight to negative welfare than to positive welfare by, for example, incorporating some version of the Priority View in our axiology. This move yields that in general, we have a stronger moral reason to refrain from creating people with negative welfare, or to increase the welfare of existing suffering people, than to create people with positive welfare, but it avoids the disagreeable implications of the Asymmetry Principle.

9

The Appeal to Desert

9.1 Introduction

A common objection to Total Utilitarianism is that it is insensitive to matters of distributive justice. Fred Feldman has developed a desert-adjusted version of Total Utilitarianism, Justicism, which he thinks fares better in this respect.[1] Moreover, Feldman claims that as a "happy by-product, justicism also generates a plausible answer to Parfit's awesome question": How many people should there ever be?[2] As a theory of distributive justice, Feldman's theory has been criticised elsewhere.[3] We shall focus on Justicism's implications in population axiology.

9.2 Feldman's Desert-Adjusted Utilitarianism

In hedonism, the value of an episode of pleasure or pain is a function of its hedonic level. In Justicism, the value of such an episode is determined not only by the hedonic level but also by the recipient's desert level: "… the intrinsic value of an episode of pleasure or pain is a function of two variables: (i) the amount of pleasure or pain the recipient *receives* in that episode, and (ii) the amount of pleasure or pain the recipient *deserves* in that episode."[4] A person's desert level is determined by factors such as her excessive or deficient past receipt of pleasure or pain, her moral worthiness, her rights and legitimate claims, her past conscientious efforts, and so forth.[5] A person is said to have "positive desert" if she deserves some pleasure, "negative desert" if she deserves some pain, and "neutral desert" if she neither

---

[1] Feldman (1995a), (1995b), reprinted in Feldman (1997).

[2] Feldman (1997), p. 195.

[3] See Carlson (1997), Persson (1997), and Vallentyne (1995).

[4] Feldman (1997), pp. 162-3, emphasis in original. Feldman couches Justicism as a version of classical hedonism mainly for pedagogical reasons. It could equally well have been stated in terms of Feldman's propositional theory of pleasure or in terms of some other theory of welfare. See Feldman (1997), p. 152.

[5] Feldman (1997), pp. 161-2, 202-3.

deserves pleasure nor pain. Feldman partly describes the relationship between pleasure, pain, desert and intrinsic value with the following six principles:[6]

M1. Positive desert enhances the intrinsic goodness of pleasure.
M2. Negative desert mitigates the intrinsic goodness of pleasure.
M3. Neutral desert neither enhances nor mitigates the intrinsic goodness of pleasure.
M4. Positive desert enhances the intrinsic badness of pain.
M5. Negative desert mitigates the intrinsic badness of pain.
M6. Neutral desert neither enhances nor mitigates the intrinsic badness of pain.

An important aspect of Feldman's theory, as we shall see, is that in some cases of negative desert, mitigations or enhancements can yield that pleasure is intrinsically bad and pain is intrinsically good. He calls this the "transvaluation" of the evil of pains and the goodness of pleasure.[7]

Unfortunately, in his discussion Feldman doesn't consistently abide by his own principles. He claims that "receipt of much less [good] than you deserve is not good for the world"; that the intrinsic value of a life led by person who deserves 100 units of pleasure but receives only one unit is -49; and "as a person begins to receive more than she deserves, additional increments of pleasure have decreasing marginal intrinsic value".[8] These claims are clearly inconsistent with M1.[9] Moreover, the first two of these claims are crucial for Feldman's results in population axiology. As Ingmar Persson has pointed out, Feldman oscillates between two ideas: the Merit-idea and the Fit-idea.[10] According to the former idea, the higher the desert level, the higher the value of pleasure. The latter idea, on the other hand, focuses on the degree of fit between desert and receipt. The Merit-idea corresponds pretty well with M1-6 above, whereas the Fit-idea does the work in Feldman's discussion of population axiology. We shall therefore replace M1-6 with

---

6 Feldman (1997), pp. 163-9.
7 Feldman (1997), pp. 165, 167.
8 Feldman (1997), pp. 206, 163, 209.
9 Carlson (1997), p. 315, makes the same point.
10 Persson (1997).

some new principles that better accord with Feldman's intuitions in this area. Furthermore, we shall incorporate his idea of transvaluation in the principles. Call a person's pleasure "deserved" if it roughly corresponds to her desert level, that is, if she receives exactly what she deserves or close to what she deserves. If a person's pleasure doesn't roughly correspond with her desert level and it is more (less) than she deserves, then this pleasure is "under-deserved" ("over-deserved"). The following principles probably capture Feldman's intuitions about desert and pleasure better than M1-6:

F1. Positive desert enhances the intrinsic goodness of deserved pleasure.
F2. Positive desert mitigates the intrinsic goodness of under-deserved pleasure.
F3. Positive desert mitigates and might transvaluate the intrinsic goodness of over-deserved pleasure.
F4. Negative desert mitigates and might transvaluate the intrinsic goodness of pleasure.
F5. Neutral desert neither enhances nor mitigates the intrinsic goodness of pleasure.
F6. Positive desert enhances the intrinsic badness of pain.
F7. Negative desert mitigates and might transvaluate the intrinsic badness of pain.
F8. Neutral desert neither enhances nor mitigates the intrinsic badness of pain.

As we noticed above, if pleasure is over-deserved, then we might get transvaluation of the intrinsic goodness of pleasure – hence the formulation of F3. What about under-deserved pleasure? Feldman isn't very clear on this point, but we shall interpret his talk about "decreasing marginal intrinsic value" such that positive desert can mitigate but not transvaluate the intrinsic goodness of under-deserved pleasure. Consequently, F2 doesn't say anything about transvaluations.[11]

[11] Feldman (1997), p. 168, also claims that "it is not so good for a person who deserves pain to get either more or less pain than he deserves. This corresponds to the intuition that punishment must be proportional to the crime." This idea is compatible with M5 (F7) but I would suggest reformulating this principle too since it seems odd, from the perspective of proportional justice,

Finally, according to Justicism, the intrinsic value of a person's life is the sum of the desert-adjusted intrinsic value of the episodes of pleasure and pain that occur in her life. The value of a population is the sum of the values of all the lives in the population.[12]

9.3 Justicism and the Repugnant Conclusion

Feldman has not given us any exact formula for calculating the desert-adjusted value of a life. He says, in his discussion of the Repugnant Conclusion, that a person who deserves 100 units of pleasure and receives exactly that amount of pleasure, has a contributive value of 200. As we noticed above, if a person deserving 100 units only receives one unit of pleasure, then the contributive value of her life is -49.[13]

How is Justicism supposed to avoid the Repugnant Conclusion? In an interesting reversal of the Christian doctrine of the original sin, Feldman assumes that there is "some modest level of happiness that people deserve merely in virtue of being people".[14] Furthermore, he assumes that this modest level corresponds to 100 units of pleasure and that people with very low welfare enjoy only one unit of pleasure. Since such lives have a negative contributive value of -49, any population consisting of people with very low welfare and desert level 100 has negative value, whereas any population with very high welfare has positive value.[15] Consequently, it seems like Justicism avoids the Repugnant Conclusion and satisfies the Quality Condition.

The intuition behind Feldman's explanation of the unacceptability of the Repugnant Conclusion – that there is some level of welfare that people deserve merely in virtue of being people – is compelling and probably shared by many

---

that negative desert mitigates the intrinsic badness of very under-deserved pain, that is, pain that goes far beyond the deserved pain. To fully capture the Fit-idea, M3 and M6 (F5 an F8) also need to be reformulated, but I shall not pursue this matter further here.

12 Feldman (1997), p. 169, writes: "The intrinsic value of a whole consequence is the sum of the justice-adjusted intrinsic value of the episodes of pleasure and pain that occur in that consequence." On p. 208, he says that "… the relevant … value of a world … is the sum of the values of the lives lived there, adjusted for desert …".

13 Feldman (1997), pp. 206, 209.

14 Feldman (1997), p. 194.

15 Given the assumption that positive desert cannot yield transvaluation of the intrinsic goodness of under-deserved pleasure.

people. Moreover, Feldman's theory can explain ideas such as Blackorby et al.'s critical level. But I'm not sure it really delivers what it promises. Feldman's reasoning involves a questionable interpretation of the *ceteris paribus* clause in the Repugnant Conclusion. He implicitly assumes that the *ceteris paribus* clause is satisfied whenever the people in the compared populations have the same desert level. This interpretation – let's call it the "Same Merit Interpretation" – is questionable for two reasons. Firstly, given the Fit-idea, which is crucial for Feldman's "solution" to the Repugnant Conclusion, the Same Merit Interpretation seems out of place. Rather, closer at hand is the view that the *ceteris paribus* clause is satisfied if there is the same fit between what people deserve and what they receive in compared populations. Again, Feldman oscillates between the Merit- and the Fit-idea. More importantly, it is not at all clear why we should focus on each individual's desert level. As we said in section 3.1.1, the *ceteris paribus* clause is satisfied if and only if the compared populations are (roughly) equally good in regard to other axiologically relevant aspects apart from welfare. Consequently, what we are looking for are cases where the compared populations are, in some sense, equally good in regard to desert.

Admittedly, it is not completely clear how we should understand the *ceteris paribus* clause in relation to Justicism since the two axiologically relevant aspects are entangled in a complex manner. As we saw above, the intrinsic value of an episode of pleasure or pain depends on the amount of pleasure or pain the recipient deserves in that episode. Strictly speaking, this is not compatible with Feldman's own idea of intrinsic value. As he writes in another context, "[s]urely, if something is intrinsically good, it must be good in virtue of the way it is in itself, not merely because of some extrinsic relation it happens to bear to some other thing".[16] But in Justicism, the intrinsic value of an episode of pleasure or pain depends on contingent facts regarding the desert level of the recipient. More in line with Feldman's idea of intrinsic value would be to consider the fit between desert and receipt another intrinsic value apart from pleasure and pain. Let's call the value of the fit between desert and receipt in a life that life's *desert value*. For example, Feldman says that the value of a life enjoying a deserved one unit of pleasure is

16 Feldman (1997), p. 138.

two.[17] On our suggested revision of Justicism, this means that the intrinsic value of the pleasure in this life is one unit, and the intrinsic value of the fit between desert and receipt in this life, its desert value, is also one unit. These two values taken together yield that the intrinsic value of this life is two units.

Given this revision of Justicism, there is a straightforward interpretation of the *ceteris paribus* clause: The compared populations should be equally good in regard to desert value, we shouldn't have a reason to chose one or the other of the compared populations because of the fit between desert and receipt. Now, to be able to determine how good a population is in regard to desert value, we need a method of aggregating this value. Given Feldman's framework, it is natural to assume that we should determine this value by adding up the measure of fit between desert and receipt (recall that according to Justicism, the value of a population is the sum of the desert-adjusted intrinsic value of the episodes of pleasure and pain that occur in the population). Consequently, two populations are equally good in regard to desert if and only if the total sum of desert value is the same in the compared populations. Let's call this the "Same Desert Value Interpretation" of the *ceteris paribus* clause. Given Feldman's ideas about desert, I find this a much more plausible reading of the *ceteris paribus* clause than the Same Merit Interpretation. If the only relevant axiological aspects of a population are people's welfare and their desert value, and two populations differ in regard to people's welfare but not in regard to their desert value, then it is plausible to say that compared populations are equally good in regard to all other axiologically relevant aspects apart from welfare.

Of course, Feldman might not accept our revision of his theory and instead opt for revising his conception of intrinsic value.[18] But the above discussion has given us a natural way to understand the *ceteris paribus* clause also in relation to Feldman's original version of Justicism. Again, it seems reasonable to claim that the *ceteris paribus* clause is satisfied if the compared populations are equally good in regard to desert value. Given the original version of Justicism, we need to slightly adjust our definition of desert value, however. In relation to that theory, we shall define a life's desert value in terms of the difference between the value of that life and the value it

---

[17] Feldman (1997), p. 212.

[18] He could also claim that pleasure doesn't have value in itself and that the only carriers of intrinsic value are compound states of affairs consisting of a person's experience of pleasure or pain *and* their desert level.

would have had if it had neutral desert, that is, in terms of how much the fit between desert and receipt contributes or detracts from the intrinsic value of the welfare of a life. We are, so to say, factoring out the desert component of the intrinsic value of a life. And this way of understanding a life's desert value yields, of course, the same result as the definition we suggested in connection with the revised version of Justicism. For example, the desert value of a life enjoying one unit of deserved pleasure equals (2-1)=1 unit.

For the arguments below, it doesn't matter whether we take Justicism in one or the other version that we have discussed above. The interesting question is whether Justicism avoids the Repugnant Conclusion if we read the *ceteris paribus* clause according to the Same Desert Value Interpretation. I don't think so. Let's first consider a population A with very high deserved welfare. Assume that the total desert value of this population is x units. Again, the desert value of a life enjoying a deserved one unit of pleasure is one. Consequently, a population B consisting of x lives enjoying a deserved one unit of pleasure will have the same total desert value as population A (we are here simplifying our reasoning by assuming that the desert value of a population can always be represented by an integer, but it should be clear how the argument could proceed without this assumption). If the total welfare of population B is less than the total welfare of population A, then just add a sufficient number of lives with neutral desert and one unit of pleasure. The resulting population consists only of people with very low positive welfare, and is equally as good as A in regard to desert, but better than A according to Justicism since the total welfare is greater. We can proceed similarly with populations of people with under-deserved very high welfare, populations of people with over-deserved very high welfare, and populations of people with very high welfare but with varying desert value. Admittedly, we cannot demonstrate this in an exact manner since such an exercise has to involve two factors not clearly defined by Feldman: How to measure the fit when the desert factor and receipt don't match, and how to calculate transvaluation of pleasure. In all likelihood, however, for any population of the above mentioned type with very high welfare, we can find a population with very low positive welfare and with the same desert value. And by adding a sufficient number of lives with neutral desert and very low positive welfare, we will get populations that have the same desert value as the populations with very high welfare, but which are better according to Justicism since the total welfare is

greater. Consequently, contrary to Feldman, it looks like Justicism implies the Repugnant Conclusion.

As we showed above, in cases involving deserved welfare Justicism implies repugnant conclusions. Feldman discusses a case like this, although, since he reads the *ceteris paribus* clause according to the Same Merit Interpretation, he doesn't think it exemplifies the Repugnant Conclusion. Nevertheless, he considers whether this "variant of the original example will prove just as repugnant".[19] He writes:

> … it is not entirely clear that Z' [a population of two billion billion people with deserved pleasure one] ought to be considered horrible. Note that the residents of Z' are not like us. They deserve far less than we deserve. Each of them deserves just +1 and each of them gets exactly what he or she deserves. Since Z' is so incredibly populous, and since the total amount of good enjoyed by the residents is so huge, and since everything in Z' is said to be just as it ought to be, it is not clear that we should find Z' repugnant.[20]

I don't find this answer convincing. It rests on the kind of misunderstanding of the Repugnant Conclusion we discussed in section 3.1.2. Again, the counter-intuitiveness of the Repugnant Conclusion doesn't essentially rest on categorical properties of populations with very low positive welfare, for example, that such populations are repugnant or very bad in themselves. The unacceptability of the Repugnant Conclusion arises from the fact that any population with very high welfare is *worse* than some population with very low welfare. It is this *comparative aspect* of the Repugnant Conclusion that we find hard to accept. And this counter-intuitiveness is not ameliorated by Feldman's appeal to desert. On the contrary, I suggest that those who embrace Feldman's explication of desert should have an even more firm belief about the unacceptability of the Repugnant Conclusion than those who don't: How could the fact that the people with very high welfare also have a *very high desert level*, and the fact that the people with very low welfare also have a *very low desert level*, reverse our intuitive judgement about the Repugnant

---

[19] Feldman (1997), p. 212.
[20] Feldman (1997), p. 212.

Conclusion? Since the people with very low desert level must be less morally worthy, made less conscientious efforts, and so forth, than the people with very high welfare, these differences in desert only serve to strengthen the dreadful character of the Repugnant Conclusion.

Feldman also claims that "it is not clear that the description of Z' [the population of two billion billion people with deserved pleasure one] is coherent":

> I stipulated that merely in virtue of being a person, each of us deserves +100. - - - The people in Z' allegedly deserve much less. But why do they deserve much less? It must be because they did something wrong. If they did something wrong, that would make Z' worse. The description of Z' is therefore incomplete. These people must have done some evil deeds. Yet the evil of those deeds is neither described nor included in the calculations.[21]

I find this claim surprising. It seems clear to me that the "evil of those deeds" is included in the calculations made when we evaluate worlds according to Justicism. These deeds are reflected in the low desert-level of the Z'-people and in their low welfare. For example, the population with very high welfare could be inhabitants of a world where people cooperate and help each other. Their high desert level is a reflection of their cooperative and helpful characters whereas their high welfare is due to the fruits of cooperation. In Z', on the other hand, people don't cooperate but only look out for themselves. They live in a Hobbesian state of nature: "war of every one against every one" which makes life "nasty, brutish and short".[22] The low desert level in this world reflects the mean character of its inhabitants whereas their low welfare is caused by the lack of cooperation.

For all that we can say, Justicism in conjunction with the Same Desert Value Interpretation of the *ceteris paribus* clause implies the Repugnant Conclusion. Assume, implausibly, that Feldman could muster some decisive argument to the effect that although Justicism is based on the Fit-idea, the Same Merit Interpretation is the most plausible way to understand the *ceteris paribus* clause.

---

[21] Feldman (1997), p. 212.
[22] Hobbes (1962), p. 100.

Although Justicism avoids the Repugnant Conclusion, given the Same Merit Interpretation, it would imply analogous conclusions. If all the people involved have neutral desert, then Justicism yields the same ranking as Total Utilitarianism, since neutral desert neither enhances nor mitigates the intrinsic goodness (badness) of pleasure (pain). Consequently, for any population with very high welfare and neutral desert, there is a population with very low welfare and neutral desert which is better.[23] But worse is yet to come.

9.4 Justicism and the Non-Sadism Condition

Justicism implies that an addition of lives with negative welfare can make a population better. At one point, Feldman claims that "it is slightly good (+ 2.5) for a person to receive 10 units of pain when this is precisely what he deserves".[24] Consequently, we can make the world much better by adding a large number of lives with deserved negative welfare. Since we can assume that the people in the compared populations have the same desert level, Justicism violates the Negative Mere Addition Principle given the Same Merit Interpretation. According to Feldman, the positive value of lives with deserved negative welfare "expresses the retributivist axiological intuition that sometimes it is good for bad people to be punished".[25] It is rather, I surmise, a distortion of the retributivist's axiological intuition. An axiological retributivist thinks that *if* people commit crimes, *then* it is good that they are punished. In other words, it is worse if crimes are committed with impunity than if they are committed and punished, but neither of these states of affairs are good. Retributivists don't think that it is good that people commit crimes and are punished.

Since lives with positive welfare might have negative contributive value and lives with negative welfare might have positive contributive value, it should come as no surprise that Justicism violates the Non-Sadism Condition. As we noticed above, Feldman's theory yields that the intrinsic value of a person who deserves 100 units of pleasure but receives only one unit is -49. What is the intrinsic value of

---

[23] As we pointed out in fn. 11, to fully capture the Fit-idea, M3 (F5) also need to be reformulated. From the perspective of the Fit-idea, it seems reasonable to claim that neutral desert mitigates the intrinsic goodness of pleasure. This refomulation wouldn't save Justicism from the conclusion above, however.

[24] Feldman (1997), pp. 167-8.

[25] Feldman (1997), p. 167.

a person who deserves 100 units of pleasure but receives terrible pain? As it is presented, Feldman's theory doesn't give an exact answer to this question. At any rate, it follows from F6 that such a life is also going to have negative intrinsic value, presumably much below -49.[26] Assume that the value of such a life is $-k$. Now, for any number n of people who suffer terrible pain but who deserve 100 units of pleasure, there is a number m of people with over-deserved positive welfare but with lower total contributive value, namely any number $m > kn/49$ of such lives. From the perspective of desert, I find this violation of the Non-Sadism Condition perplexing. Surely, if everybody deserves positive welfare and one has a choice of adding people suffering terrible pain or people with positive (albeit over-deserved) welfare, then the latter addition must be the better one.

To be fair to Feldman, it's not completely clear whether he subscribes to transvaluation of the evil of pain and the goodness of pleasure by negative desert. At some points, he suggests that the intrinsic value of a life with deserved negative welfare is zero.[27] Still, this version of Justicism doesn't capture the view of the axiological retributivist since lives with deserved negative welfare have neutral rather than negative contributive value. Moreover, since the contributive value of people with over-deserved positive welfare can be negative, this version of Justicism, in conjunction with the Same Merit Interpretation, still violates the Non-Sadism Condition.

Given the Same Desert Value Interpretation, Justicism doesn't violate the Non-Sadism Condition and the Negative Mere Addition Principle since if the desert value is the same in the compared populations, the ranking is going to be determined solely by the total sum of people's welfare. This is not much of a comfort, however, since Justicism will still have the implications pointed out above, although, since other things won't be equal, these implications won't formally count as violations of the Non-Sadism Condition and the Negative Mere Addition Principle. Moreover, with this interpretation of the *ceteris paribus* clause, Justicism

---

[26] Feldman's choice of numerical representation is, according to himself, "somewhat arbitrary" but since the value of a person who deserves 100 units of pleasure and receives nothing is -50, and positive desert enhances the intrinsic badness of pain (see M4 and F6 above), a person who deserves 100 units of pleasure and receives pain must have negative intrinsic value below -50. See Feldman (1997), p. 206.

[27] Feldman (1997), p. 167. At p. 165, Feldman also suggests that the value of a person with negative desert but positive welfare is zero rather than negative.

implies, in all likelihood, the Repugnant Conclusion and violates the Quality Condition. Again, we haven't gained any ground in our search for a reasonable population axiology. As we shall see in the next chapter, the prospect for such an axiology is indeed bleak.

10

Four Axiological Impossibility Theorems

10.1 Introduction

As we have seen, all the population axiologies presented in the literature have counter-intuitive implications. We shall now challenge the idea that there is a satisfactory population axiology. One can do this by proving that no population axiology satisfies a set of adequacy conditions, that is, by proving an impossibility theorem. We shall prove four theorems. We have put the theorems in order of the intuitive plausibility of the conditions involved. The first two involve conditions that we have found some reasons to dispute. The third and especially the fourth theorem, however, involve conditions that we find hard to reject, and thus challenge the very idea that there is a satisfactory population axiology.

We shall start by formally proving two results which we discussed informally in chapter 3 and 6. The first theorem illustrates the tension between the Quality and the Quantity Condition. The second theorem is a version of the Mere Addition Paradox but with logically weaker and intuitively more plausible conditions than those used elsewhere in the literature.

The third theorem involves less controversial assumptions than the first two since it doesn't involve any version of the controversial Mere Addition Condition nor the Quantity Condition. We shall show that no population axiology can fulfil the Egalitarian Dominance Condition, the Inequality Aversion Condition, the Non-Extreme Priority Condition, the Non-Sadism Condition, and the Weak Quality Addition Condition. This theorem shows that the on-going project of constructing an acceptable population axiology has gloomy prospects.

The fourth and last theorem is an extension of the third. As we discussed in section 4.2, some theorists have objected to the Non-Sadism Condition. We shall show that one can replace this condition with the Weak Non-Sadism Condition given that one substitutes two logically stronger but intuitively at least as convincing conditions for the Inequality Aversion and the Non-Extreme Priority Condition,

namely the Non-Elitism Condition and the General Non-Extreme Priority Condition. This theorem involves the intuitively most compelling conditions and thus provides the most serious challenge to the very idea that there is an acceptable population axiology.

Some readers might find this chapter a difficult read. As before, however, we shall use diagrams as an expository device to make it easier to follow the steps in the arguments. If one so chooses, one can get an intuitive grasp of the theorems by following the diagrammatic presentation of the populations involved in the application of a certain condition in a certain step of a theorem. It is important to remember, however, that these diagrams have limited significance. The blocks in the diagrams only represent possible pairs of populations that fit the description of some condition. For example, the area of the blocks that we draw cannot properly be said to represent the total welfare of a population since we haven't assumed that welfare can be measured on at least a ratio scale. Only if welfare were measurable on a ratio scale, would it make sense to say that the area of a block represented the total welfare of a population.

10.2 The Basic Structure

For the purpose of proving the theorems, it will be useful to briefly repeat the definitions and basic assumptions from earlier chapters and introduce some new notation. A *life* is individuated by the person whose life it is and the kind of life it is. A *population* is a finite set of lives in a possible world (for other constraints on possible populations, see section 2.3). We shall assume that for any natural number n and any welfare level $\mathbf{X}$, there is a possible population of n people with welfare $\mathbf{X}$ (this is a generalisation of the *No-Limit Assumption* from section 3.1.3). Two populations are identical if and only if they consist of the same lives. Since the same person can exist (be instantiated) and lead the same kind of life in many different possible worlds, the same life can exist in many possible worlds. Moreover, since two populations are identical exactly if they consist of the same lives, the same population can exist in many possible worlds. A *population axiology* is an "at least as good as" quasi-ordering of all possible populations, that is, a reflexive, transitive, but not necessarily complete ordering of populations in regard to their goodness (see chapter 2).

A, B, C,… A_1, A_2,…, A_n, $A \cup B$, and so on, denote populations of finite size. The number of lives in a population X (X's population size) is given by the

function $N(X)$. We shall adopt the convention that populations represented by different letters, or the same letter but different indexes, are pairwise disjoint.

The relation "*has at least as high welfare as*" quasi-orders (reflexive, transitive, but not necessarily complete) the set $\mathcal{L}$ of all possible lives. A life p_1 has higher welfare than another life p_2 if and only if p_1 has at least as high welfare as p_2 and it is not the case that p_2 has at least as high welfare as p_1. p_1 has the same welfare as p_2 if and only if p_1 has at least as high welfare as p_2 and p_2 has at least as high welfare as p_1. As we discussed in chapter 2, we assume that there are possible lives with positive or negative welfare. We shall say that a life has *neutral welfare* if and only if it has the same welfare as a life without any good or bad welfare-components, and that a life has *positive* (*negative*) welfare if and only if it has higher (lower) welfare than a life with neutral welfare.

By a *welfare level* $\mathbb{A}$ we shall mean a set such that if a life a is in $\mathbb{A}$, then a life b is in $\mathbb{A}$ if and only if b has the same welfare as a. In other words, a welfare level is an equivalence class on $\mathcal{L}$. Let a^* be a life which is representative of the welfare level $\mathbb{A}$. We shall say that a welfare level $\mathbb{A}$ is higher (lower, the same) than (as) a level $\mathbb{B}$ if and only if a^* has higher (lower, the same) welfare than (as) b^*; that a welfare level $\mathbb{A}$ is positive (negative, neutral) if and only if a^* has positive (negative, neutral) welfare; and that a life b has welfare below (above, at) $\mathbb{A}$ if and only if b has higher (lower, the same) welfare than (as) a^*.

We shall assume that *Discreteness* is true of the set of all possible lives $\mathcal{L}$ or some subset of $\mathcal{L}$:

> *Discreteness*: For any pair of welfare levels $\mathbb{X}$ and $\mathbb{Y}$, $\mathbb{X}$ higher than $\mathbb{Y}$, the set consisting of all welfare levels $\mathbb{Z}$ such that $\mathbb{X}$ is higher than $\mathbb{Z}$, and $\mathbb{Z}$ is higher than $\mathbb{Y}$, has a finite number of members.

Some of the adequacy conditions that we have discussed, for example the Quantity Condition, involve the not so exact relation "slightly higher welfare than". In the exact statements of those adequacy conditions, we shall instead make use of two consecutive welfare levels, that is, two welfare levels such that there is no welfare level in between them. Discreteness ensures that there are such welfare levels. Intuitively speaking, if $\mathbb{A}$ and $\mathbb{B}$ are two consecutive welfare levels, $\mathbb{A}$ higher than $\mathbb{B}$, then $\mathbb{A}$ is just slightly higher than $\mathbb{B}$. More importantly, the intuitive plausibility of the adequacy conditions is preserved. Of course, this presupposes

that the order of welfare levels is fine-grained, which is exactly what is suggested by expressions such as "Peter is slightly better off than Gert" and the like.

Discreteness can be contrasted with Denseness:

> *Denseness*: There is a welfare level in between any pair of distinct welfare levels.

My own inclination is that Discreteness rather than Denseness is true. If the latter is true, then for any two lives p_1 and p_2, p_1 with higher welfare than p_2, there is a life p_3 with welfare in between p_1 and p_2, and a life p_4 with welfare in between p_3 and p_2, and so on *ad infinitum*. It is improbable, I think, that we can make such fine discrimination between the welfare of lives, even in principle.[1] Rather, what we will find at the end of such a sequence of lives is a pair of lives in between which we cannot find any life or only lives with roughly the same welfare as both of them. One might think otherwise, and a complete treatment of this topic would involve a detailed examination of the features of different welfarist axiologies. We shall not engage in such a discussion here. The important question is whether the validity and plausibility of the theorems below depend on whether Denseness or Discreteness is true. But that is not the case (indeed, it would have been an interesting result if the existence of a plausible axiology hinged on whether Denseness or Discreteness is true). If Denseness is true of the set of all possible lives $\mathscr{L}$, then we can form a subset $\mathscr{L}_1$ of $\mathscr{L}$ such that Discreteness is true of $\mathscr{L}_1$, and such that all the conditions which are intuitively plausible in regard to populations which are subsets of $\mathscr{L}$ also are intuitively plausible in regard to populations which are subsets of $\mathscr{L}_1$. Given that Denseness is true of $\mathscr{L}$, one cannot plausibly deny that there is such a subset $\mathscr{L}_1$ since the order of the welfare levels in $\mathscr{L}_1$ could be arbitrarily fine-grained even though Discreteness is true of $\mathscr{L}_1$. Now, since all the populations which are subsets of $\mathscr{L}_1$ also are subsets of $\mathscr{L}$, if we can show that there is no population axiology satisfying the adequacy conditions in regard to the

[1] In ch. 2, I expressed scepticism about the reliability of our intuitions concerning cases which involve populations of infinite size or "infinite welfare". I think the same concern may hold true for infinitesimal differences in welfare. Consequently, even if Denseness is true, we may have epistemological reasons to be sceptical about arguments that essentially rely on this feature of Denseness.

populations which are subsets of $\mathcal{L}_1$, then it follows that there is no population axiology satisfying the adequacy conditions in regard to the populations which are subsets of $\mathcal{L}$.

Notice that Discreteness doesn't exclude the view that for any welfare level, there is a higher and a lower welfare level (compare with the natural numbers).

Given Discreteness, we can index welfare levels with integers in a natural manner. Discreteness in conjunction with the existence of a neutral welfare level and a quasi-ordering of lives implies that there is at least one positive welfare level in $\mathcal{L}$ such that there is no lower positive welfare level.[2] Let $\mathbb{W}_1$, $\mathbb{W}_2$, $\mathbb{W}_3$,… and so forth represent positive welfare levels, starting with one of the positive welfare level for which there is no lower positive one, such that for any pair of welfare levels $\mathbb{W}_n$ and $\mathbb{W}_{n+1}$, $\mathbb{W}_{n+1}$ is higher than $\mathbb{W}_n$, and there is no welfare level $\mathbb{X}$ such that $\mathbb{W}_{n+1}$ is higher than $\mathbb{X}$, and $\mathbb{X}$ is higher than $\mathbb{W}_n$. Analogously, let $\mathbb{W}_{-1}$, $\mathbb{W}_{-2}$, $\mathbb{W}_{-3}$,… and so on represent negative welfare levels. The neutral welfare level is represented by $\mathbb{W}_0$.

A *welfare range* $\mathbb{R}$(x, y) is a union of at least *three* welfare levels defined by two welfare levels $\mathbb{W}_x$ and $\mathbb{W}_y$, $x < y$, such that for any welfare level $\mathbb{W}_z$, $\mathbb{W}_z$ is a subset of $\mathbb{R}$(x, y) if and only if $x \leq z \leq y$.[3] We shall say that a welfare range $\mathbb{R}$(x, y) is higher (lower) than another range $\mathbb{R}$(z, w) if and only if $x > w$ ($y < z$); that a welfare range $\mathbb{R}$(x, y) is positive (negative) if and only if $x > 0$ ($y < 0$); and that a life p has welfare above (below, in) $\mathbb{R}$(x, y) if and only if p is in some $\mathbb{W}_z$ such that $z > y$ (z < x, $y \geq z \geq x$).

10.3 Adequacy Conditions for the First Theorem

In section 3.2, we claimed that an axiology cannot satisfy the Quality Condition if it satisfies the Egalitarian Dominance Condition and the following condition:

> *The Quantity Condition*: For any pair of positive welfare levels $\mathbb{A}$ and $\mathbb{B}$, such that $\mathbb{B}$ is slightly lower than $\mathbb{A}$, and for any number of lives n, there is a greater number of lives m, such that a population of m people at level

---

[2] There might be more than one since we only have an quasi-ordering of lives, that is, there might be lives and thus welfare levels which are incommensurable in regard to welfare.

[3] The reason for restricting welfare ranges to unions of at least three welfare levels, as opposed to at least two welfare levels, is that this restriction allows us to simplify the exact statements of the adequacy conditions.

$\mathbb{B}$ is at least as good as a population of n people at level $\mathbb{A}$, other things being equal.

The Quantity Condition (exact formulation): For any two positive welfare levels $\mathbb{W}_x$ and $\mathbb{W}_y$, such that $x=y+1$, and for any population size n, there is a population size $m > n$, such that **if** $N(A)=n$, $N(B)=m$, $A \subset \mathbb{W}_x$, and $B \subset \mathbb{W}_y$, **then** B is at least as good as A, other things being equal.

Notice that the exact formulation of this condition doesn't involve the not so precise concept "slightly lower welfare". Instead, we have formulated it in terms of two consecutive welfare levels.

The two other conditions that we shall employ in this theorem are the following:

The Egalitarian Dominance Condition: If population A is a perfectly equal population of the same size as population B, and every person in A has higher welfare than every person in B, then A is better than B, other things being equal.

The Egalitarian Dominance Condition (exact formulation): For any populations A and B, $N(A)=N(B)$, and any welfare level $\mathbb{W}_x$, **if** all members of B have welfare below $\mathbb{W}_x$, and $A \subset \mathbb{W}_x$, **then** A is better than B, other things being equal.

The Quality Condition: There is at least one perfectly equal population with very high positive welfare which is at least as good as any population with very low positive welfare, other things being equal.

The Quality Condition (exact formulation): There are two positive welfare ranges $\mathbb{R}(u, v)$ and $\mathbb{R}(1, y)$, $u > y$, and a population size $n > 0$, such that **if** $\mathbb{W}_z \subset \mathbb{R}(u, v)$, $A \subset \mathbb{W}_z$, $N(A)=n$, and $B \subset \mathbb{R}(1, y)$, **then** A is at least as good as B, other things being equal.

In chapter 2, we claimed that concepts such as "very high positive welfare", "very low positive welfare", "slightly negative welfare", and the like are not essential

for our discussion and results. It is now time to show that this is true. Thus, in the exact formulation of the Quality Condition, we have eliminated the concepts "very low positive welfare" and "very high positive welfare" and replaced them with two non-fixed positive welfare ranges, one starting at the lowest positive welfare level, and the other one starting anywhere above the first range.[4]

10.4 The First Impossibility Theorem

> ***The First Impossibility Theorem***: There is no population axiology which satisfies the Quality Condition, the Egalitarian Dominance Condition, and the Quantity Condition.

Proof: We show that the contrary assumption leads to a contradiction. Let

(1) $\mathbb{R}(u, v)$ and $\mathbb{R}(1, y)$, $u > y$, be two positive welfare ranges, and n_1 a number, that satisfy the Quality Condition;
(2) $r = u-1$;
(3) $n_{i+1} > n_i$ be a number which satisfies the Quantity Condition for $\mathbb{W}_{u-(i-1)}$, $\mathbb{W}_{u-i}$, and n_i for all i, $1 \leq i \leq r$;
(4) $A_i \subset \mathbb{W}_{u-(i-1)}$, $N(A_i) = n_i$ for all i, $1 \leq i \leq r+1$.

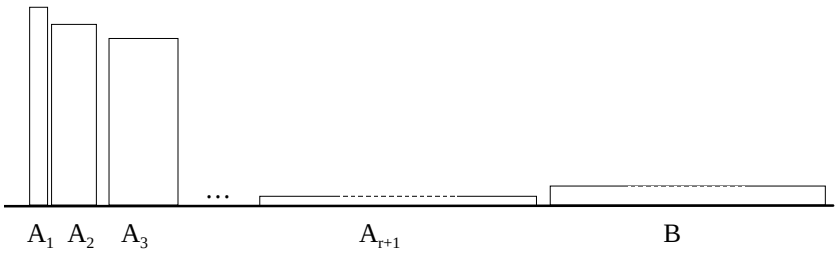

Diagram 10.4.1

From (3), (4), and the Quantity Condition, it follows that (see Diagram 10.4.1)

---

[4] The exact formulation of the Quality Condition is in one sense stronger than the informal formulation, but in an intuitively trivial way. Since a welfare range consists of at least three welfare levels, the former implies that there are populations on at least three "very high" welfare levels which are at least as good as all populations with "very low positive welfare".

(5) A_{i+1} is at least as good as A_i.

Transitivity and (5) yield that

(6) A_{r+1} is at least as good as A_1.

Let

(7) $B \subset \mathbb{W}_y$, $N(B)=N(A_{r+1})$ (see Diagram 10.4.1).

Since $A_{r+1} \subset \mathbb{W}_{u-r}=\mathbb{W}_1$ (2, 4) and $y > 1$ (1), the Egalitarian Dominance Condition implies that

(8) B is better than A_{r+1}.

By transitivity, (6), and (8), it follows that

(9) B is better than A_1.

The Quality Condition, (1), (4) and (7) yield that

(10) A_1 is at least as good as B

which contradicts (9). Hence, the assumption that there is an axiology which satisfies all the adequacy conditions entails a contradiction. Thus, the impossibility theorem must be true. Q.E.D.

10.5 Adequacy Conditions for the Second Theorem

We shall now produce a version of the Mere Addition Paradox with logically weaker and intuitively more plausible conditions than those used elsewhere in the literature. We shall substitute the Mere Addition Principle with the following condition:

The Dominance Addition Condition: An addition of lives with positive welfare and an increase in the welfare in the rest of the population doesn't make a population worse, other things being equal.

The Dominance Addition Condition (exact formulation): For any populations A, B, and C, and any pair of welfare levels $\mathbb{W}_x$ and $\mathbb{W}_y$, $y > 0$, **if** all the lives in A have welfare below $\mathbb{W}_x$, all the lives in B have welfare above or at $\mathbb{W}_x$, $N(A)=N(B)$, and $C \subset \mathbb{W}_y$, **then** $B \cup C$ is not worse than A, other things being equal.

Instead of the Non-Anti Egalitarianism Principle, we shall use

The Inequality Aversion Condition: For any triplet of welfare levels $\mathbb{A}$, $\mathbb{B}$, and $\mathbb{C}$, $\mathbb{A}$ higher than $\mathbb{B}$ and $\mathbb{B}$ higher than $\mathbb{C}$, and for any population A with welfare $\mathbb{A}$, there is a larger population C with welfare $\mathbb{C}$ such that a perfectly equal population B of the same size as $A \cup C$ and with welfare $\mathbb{B}$ is at least as good as $A \cup C$, other things being equal.

The Inequality Aversion Condition (exact formulation): For any triplet $\mathbb{W}_x$, $\mathbb{W}_y$, $\mathbb{W}_z$ of welfare levels, $x > y > z$, and any number of lives $n > 0$, there is a number of lives $m > n$ such that **if** $A \subset \mathbb{W}_x$, $N(A)=n$, $B \subset \mathbb{W}_y$, $N(B)=m+n$, and $C \subset \mathbb{W}_z$, $N(C)=m$, **then** B is at least as good as $A \cup C$, other things being equal.

The two other conditions that we shall employ in this theorem are the Egalitarian Dominance Condition and the Quality Condition.

10.6 The Second Impossibility Theorem

The Second Impossibility Theorem: There is no population axiology which satisfies the Quality Condition, the Inequality Aversion Condition, the Egalitarian Dominance Condition, and the Dominance Addition Condition.

Proof: We show that the contrary assumption leads to a contradiction. Let

(1) $\mathbb{R}(u, v)$ and $\mathbb{R}(1, y)$, $u > y$, be two welfare ranges, and n a number, which satisfy the Quality Condition;

(2) $m > n$ be a number of lives which satisfies the Inequality Aversion Condition for n, $\mathbb{W}_v$, $\mathbb{W}_2$, and $\mathbb{W}_1$;

(3) $E \subset \mathbb{W}_v$, $N(E)=n$;

(4) $D \subset \mathbb{W}_1$, $N(D)=m$;

(5) $C \subset \mathbb{W}_2$, $N(C)=m+n$.

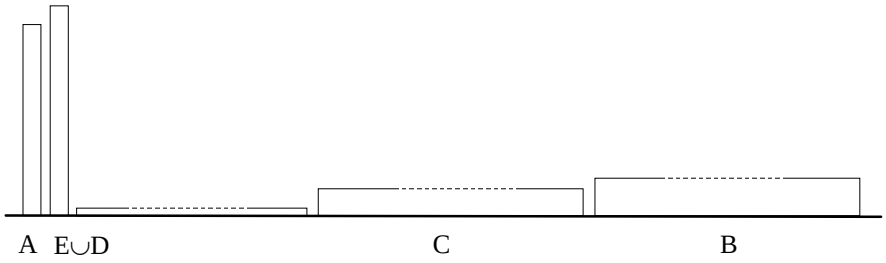

Diagram 10.6.1

The Inequality Aversion Condition and (2)-(5) imply that

(6) C is at least as good as E∪D (see Diagram 10.6.1).

Let

(7) $A \subset \mathbb{W}_u$, $N(A)=n$.

By the definition of a welfare range, we have that $v > u$. Consequently, it follows from (3)-(4), (7) and the Dominance Addition Condition that

(8) E∪D is not worse than A (see Diagram 10.6.1).

Let

(9) $B \subset \mathbb{W}_y$, $N(B)=N(C)$.

It follows from the definition of a welfare range that $y > 2$. Consequently, (5), (9), and the Egalitarian Dominance Condition yield that

(10) B is better than C (see Diagram 10.6.1).

From (1), (7), (9) and the Quality Condition we get that

(11) A is at least as good as B.

By transitivity, it follows from (10) and (11) that

(12) A is better than C

and from (6) and (12) that

(13) A is better than E∪D

which contradicts (8). Hence, the assumption that there is an axiology which satisfies all the adequacy conditions entails a contradiction. Thus, the impossibility theorem must be true. Q.E.D.

10.7 Adequacy Conditions for the Third Theorem

In this theorem, we shall make no use of any condition, such as the Quantity Condition and the Dominance Addition Condition, which, roughly, implies that the contributive value of lives with positive welfare is not negative. Instead, we shall invoke the weaker claim that it is worse to add lives with negative welfare rather than lives with positive welfare, that is, the Non-Sadism Condition:

> *The Non-Sadism Condition*: An addition of any number of lives with positive welfare is at least as good as an addition of any number of lives with negative welfare, other things being equal.

> *The Non-Sadism Condition (exact formulation)*: **If** $A \subset \mathbb{W}_x$, $x > 0$, $B \subset \mathbb{W}_y$, $y < 0$, $N(B) > 0$, **then**, for any population C, A∪C is at least as good as B∪C, other things being equal.

We shall replace the Quality Condition with the Weak Quality Addition Condition:

> *The Weak Quality Addition Condition*: For any population X, there is at least one perfectly equal population with very high welfare such that its addition to X is at least as good as an addition of any population with very low positive welfare to X, other things being equal.

> *The Weak Quality Addition Condition (exact formulation)*: For any population C, there are two positive welfare ranges $\mathbb{R}(x, w)$ and $\mathbb{R}(1, y)$, $x > y$, and a population size n such that **if** $A \subset \mathbb{W}_z$, $z \geq x$, $N(A)=n$, $B \subset \mathbb{R}(1, y)$, **then** $A \cup C$ is at least as good as $B \cup C$, other things being equal.

We shall also make use of

> *The Non-Extreme Priority Condition*: There is a number n of lives such that for any population X, a population consisting of the X-lives, n lives with very high welfare, and a single life with slightly negative welfare, is at least as good as a population consisting of the X-lives and $n+1$ lives with very low positive welfare, other things being equal.

> *The Non-Extreme Priority Condition (exact formulation)*: There are two welfare levels $\mathbb{W}_x$ and $\mathbb{W}_y$, and a welfare range $\mathbb{R}(1, z)$, $x > z, y < 0$, and a number of lives n such that **if** $A \subset \mathbb{W}_u$, $u \geq x$, $N(A)=n$, $B \subset \mathbb{R}(1, z)$, $N(B)=n+1$, and $C \subset \mathbb{W}_y$, $N(C)=1$, **then**, for any population D, $A \cup C \cup D$ is at least as good as $B \cup D$, other things being equal.

We have generalised the Weak Quality Addition Condition and the Non-Extreme Priority Condition in the same way as we generalised the Quality Condition, that is, we have eliminated the concepts "very high positive", "very low positive", and "slightly negative welfare".

The other two conditions which we shall employ in the third theorem are the Egalitarian Dominance and the Inequality Aversion Condition.

10.8 The Third Impossibility Theorem

> ***The Third Impossibility Theorem***: There is no population axiology which satisfies the Egalitarian Dominance, the Inequality Aversion, the Non-Extreme Priority, the Non-Sadism, and the Weak Quality Addition Condition.

Proof. We show that the contrary assumption leads to a contradiction. Let

(1) $\mathbb{W}_x$ and $\mathbb{W}_y$ be two welfare levels, $\mathbb{R}(1, z)$ a welfare range, $x > z, y < 0$, and q a number of lives, which satisfy the Non-Extreme Priority Condition;
(2) $B \subset \mathbb{W}_3$, $N(B)=q+1$;
(3) $\mathbb{R}(u, t)$ and $\mathbb{R}(1, v)$, $u > v$, be two welfare ranges, and p a population size, which satisfy the Weak Quality Addition Condition for B;
(4) $\mathbb{W}_w$ be a welfare level such that $w \geq x$ and $w \geq u$;
(5) $A \subset \mathbb{W}_w$, $N(A)=p$;
(6) $H \subset \mathbb{W}_w$, $N(H)=q$;
(7) $E \subset \mathbb{W}_y$, $N(E)=1$.

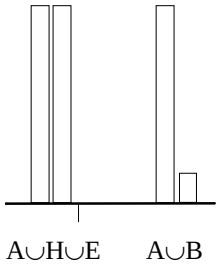

Diagram 10.8.1

It follows from the definition of a welfare range that $\mathbb{W}_3 \subset \mathbb{R}(1, z)$. Accordingly, from (2) we know that $B \subset \mathbb{R}(1, z)$. Consequently, from (1), (4), (6)-(7), and the Non-Extreme Priority Condition we get that

(8) A∪H∪E is at least as good as A∪B (see Diagram 10.8.1).

Let

(9) $r > p+q$ be a number of lives which satisfies the Inequality Aversion Condition for the three welfare levels $\mathbb{W}_w$, $\mathbb{W}_2$, and $\mathbb{W}_1$ and $p+q$ lives at $\mathbb{W}_w$;

(10) $G \subset \mathbb{W}_2$, $N(G)=p+q+r$;

(11) $F \subset \mathbb{W}_1$, $N(F)=r$.

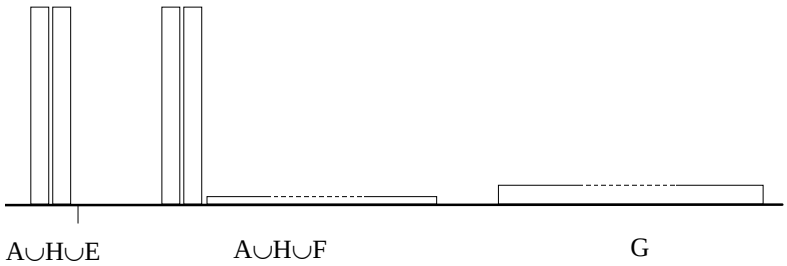

Diagram 10.8.2

Since $A \cup H \subset \mathbb{W}_w$ and $N(A \cup H)=p+q$ (by (5)-(6)), it follows from (9)-(11) and the Inequality Aversion Condition that

(12) G is at least as good as A∪H∪F (see Diagram 10.8.2).

Since the E-life has negative welfare (by (1) and (7)), and the F-lives have positive welfare (by (11)), it follows from the Non-Sadism Condition that

(13) A∪H∪F is at least as good as A∪H∪E (see Diagram 10.8.2).

By transitivity, it follows from (12) and (13) that

(14) G is at least as good as A∪H∪E.

Let

(15) $C \subset \mathbb{W}_3$, $N(C)=p+r-1$.

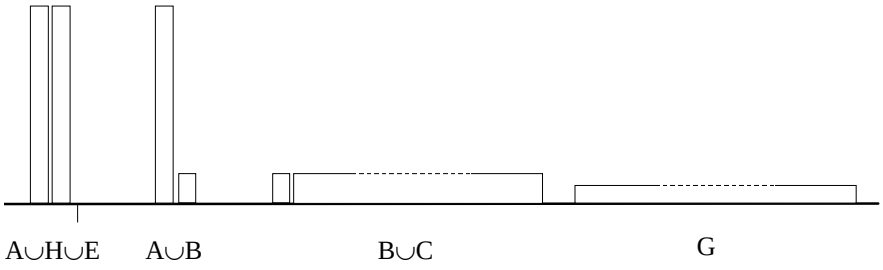

Diagram 10.8.3

Since $\mathbb{W}_3 \subset \mathbb{R}(1, v)$, we can conclude that $C \subset \mathbb{R}(1, v)$, and since $w \geq u$ (by (4)), and $A \subset \mathbb{W}_w$, and $N(A)=p$ (by (5)), it follows from (3) and the Weak Quality Addition Condition that

(16) A∪B is at least as good as B∪C (see Diagram 10.8.3).

Since $B \cup C \subset \mathbb{W}_3$ (by (2) and (15)) and $G \subset \mathbb{W}_2$, (by (10)) and $N(B \cup C) = q+1+p+r-1 = q+p+r = N(G)$, the Egalitarian Dominance Condition implies that

(17) G is worse than B∪C (see Diagram 10.8.3).

By transitivity, it follows from (16) and (17) that

(18) G is worse than A∪B

and from (8) and (18) that

(19) G is worse than A∪H∪E

which contradicts (14). Q.E.D.

10.9 Adequacy Conditions for the Fourth Theorem

The fourth theorem is a version of the third in which we shall replace the Non-Sadism Condition with the even more compelling Weak Non-Sadism Condition:

The Weak Non-Sadism Condition: There is a negative welfare level and a number of lives at this level such that an addition of any number of people with positive welfare is at least as good as an addition of the lives with negative welfare, other things being equal.

The Weak Non-Sadism Condition (exact formulation): There is a welfare level $\mathbb{W}_x$, $x < 0$, and a number of lives n, such that **if** $A \subset \mathbb{W}_x$, $N(A)=n$, $B \subset \mathbb{W}_y$, $y > 0$, **then**, for any population C, $B \cup C$ is at least as good as $A \cup C$, other things being equal.

We shall replace the Inequality Aversion Condition with

The Non-Elitism Condition: For any triplet of welfare levels $\mathbb{A}$, $\mathbb{B}$, and $\mathbb{C}$, $\mathbb{A}$ slightly higher than $\mathbb{B}$, and $\mathbb{B}$ higher than $\mathbb{C}$, and for any one-life population A with welfare $\mathbb{A}$, there is a population C with welfare $\mathbb{C}$, and a population B of the same size as $A \cup C$ and with welfare $\mathbb{B}$, such that for any population X consisting of lives with welfare ranging from $\mathbb{C}$ to $\mathbb{A}$, $B \cup X$ is at least as good as $A \cup C \cup X$, other things being equal.

The Non-Elitism Condition (exact formulation): For any welfare levels $\mathbb{W}_x$, $\mathbb{W}_y$, $x\text{-}1 > y$, there is a number of lives $n > 0$ such that **if** $A \subset \mathbb{W}_x$, $N(A)=1$, $B \subset \mathbb{W}_y$, $N(B)=n$, and $C \subset \mathbb{W}_{x-1}$, $N(C)=n+1$, **then**, for any $D \subset \mathbb{R}(y, x)$, $C \cup D$ is at least as good as $A \cup B \cup D$, other things being equal.

and the Non-Extreme Priority Condition with

The General Non-Extreme Priority Condition: There is a number n of lives such that for any population X, and any welfare level $\mathbb{A}$, a population consisting of the X-lives, n lives with very high welfare, and one life with welfare $\mathbb{A}$, is at least as good as a population consisting of the X-lives, n lives with very low positive welfare, and one life with welfare slightly above $\mathbb{A}$, other things being equal.

The General Non-Extreme Priority Condition (exact formulation): For any $\mathbb{W}_z$, there is a positive welfare level $\mathbb{W}_u$, and a positive welfare range $\mathbb{R}(1, y)$, u

> y, and a number of lives $n > 0$ such that **if** A⊂$\mathbb{W}_x$, $x \geq u$, B⊂$\mathbb{R}$(1, y), N(A)=N(B)=n, C⊂$\mathbb{W}_z$, D⊂$\mathbb{W}_{z+1}$, N(C)= N(D)=1, **then,** for any E, A∪C∪E is at least as good as B∪D∪E, other things being equal.

10.10 The Fourth Impossibility Theorem

The Fourth Theorem: There is no population axiology which satisfies the Egalitarian Dominance, the Non-Elitism, the General Non-Extreme Priority, the Weak Non-Sadism, and the Weak Quality Addition Condition.

Proof. As before, we shall show that the contrary assumption leads to a contradiction. We shall first prove two lemmas to the effect that the Non-Elitism and the General Non-Extreme Priority Condition each imply another condition. Then we shall show that there is no population axiology which satisfies these two new conditions in conjunction with the Egalitarian Dominance, the Weak Non-Sadism, and the Weak Quality Addition Condition.

10.10.1 Lemma 5.1

Lemma 5.1: The Non-Elitism Condition implies Condition β:

Condition β: For any triplet $\mathbb{W}_x$, $\mathbb{W}_y$, $\mathbb{W}_z$ of welfare levels, $x > y > z$, and any number of lives $n > 0$, there is a number of lives $m > n$ such that **if** A⊂$\mathbb{W}_x$, N(A)=n, B⊂$\mathbb{W}_z$, N(B)=m, and C⊂$\mathbb{W}_y$, N(C)=m+n, **then,** for any D⊂$\mathbb{R}$(z, y+1), C∪D is at least as good as A∪B∪D, other things being equal.

Notice that for D=∅, Condition β is equivalent with the Inequality Aversion Condition. Thus, by proving lemma 5.1, we are also proving the assertion in section 6.3 that Non-Elitism implies the Inequality Aversion Condition.

We shall prove lemma 5.1 by first proving

Lemma 5.1.1: The Non-Elitism Condition entails Condition α.

Condition α: For any welfare levels $\mathbb{W}_x$, $\mathbb{W}_y$, x-1 > y, and for any number of lives $n > 0$, there is a number of lives $m \geq n$ such that **if** A⊂$\mathbb{W}_x$, N(A)=n, B⊂$\mathbb{W}_y$, N(B)=m, C⊂$\mathbb{W}_{x-1}$, N(C)=m+n, **then**, for any D⊂$\mathbb{R}$(y, x), C∪D is at least as good as A∪B∪D, other things being equal.

Proof: Let

(1) $\mathbb{W}_x$ and $\mathbb{W}_y$ be any welfare levels such that x-1 > y;
(2) n be any number of lives such that $n > 0$;
(3) $p > 0$ be a number which satisfies the Non-Elitism Condition for $\mathbb{W}_x$ and $\mathbb{W}_y$.

Let A_1,…, A_{n+1}, B_1,…, B_{n+1}, and C_0,…, C_n, be any three sequences of populations satisfying

(4) $A_i \subset \mathbb{W}_x$; $N(A_i)=1$ for all i, $1 \leq i \leq n$; $A_{n+1}=\varnothing$;
(5) $B_i \subset \mathbb{W}_y$; $N(B_i)=p$, for all i, $1 \leq i \leq n$; $B_{n+1}=\varnothing$;
(6) $C_i \subset \mathbb{W}_{x-1}$; $N(C_i)=p+1$, for all i, $1 \leq i \leq n$; $C_0=\varnothing$.

Finally, let

(7) D be any population such that D⊂$\mathbb{R}$(y, x).

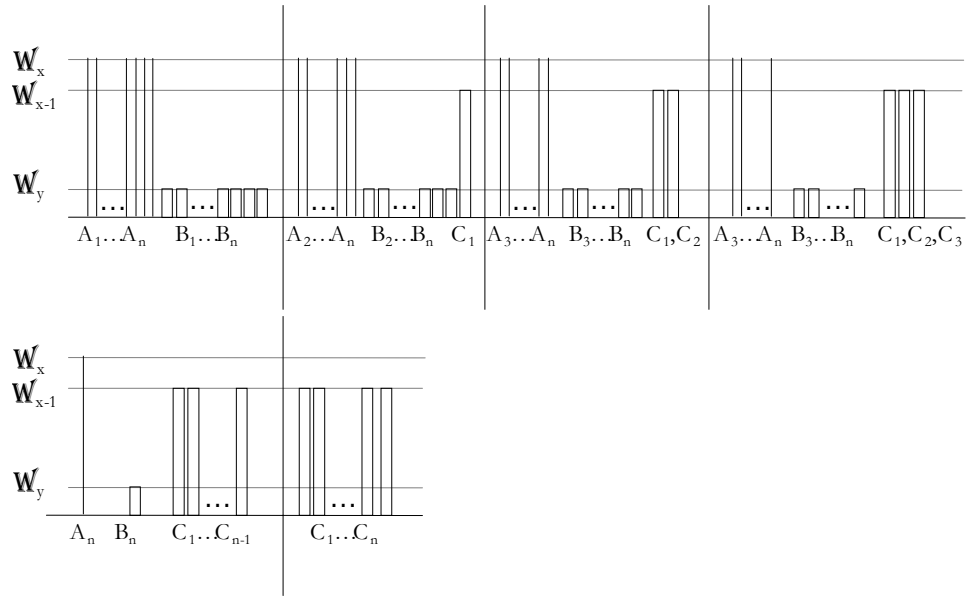

Diagram 10.10.1

The above diagram shows a selection of the involved populations in a case where $n \geq 5$. Dots in between two blocks indicate that there is a number of same sized blocks which have been omitted from the diagram. Population D is omitted throughout.

Since $\mathbb{W}_x$ and $\mathbb{W}_y$ can be any pair of welfare levels separated by at least one welfare level, and D can be any population consisting of lives with welfare ranging from $\mathbb{W}_y$ to $\mathbb{W}_x$, and $N(A_1 \cup \ldots \cup A_n)=n$ (by (4)) can be any number of lives greater than zero, and $N(B_1 \cup \ldots \cup B_n)=np \geq n$ (by (5)), we can show that lemma 5.1.1 is true by showing that $C_1 \cup \ldots \cup C_n \cup D$ is at least as good as $A_1 \cup \ldots \cup A_n \cup B_1 \cup \ldots \cup B_n \cup D$. This suffices since $A_1,\ldots, A_{n+1}$, $B_1,\ldots, B_{n+1}$, $C_1,\ldots, C_n$, and D are arbitrary populations satisfying (4)-(7).

It follows from (3)-(6) and the Non-Elitism Condition that

(8) $C_i \cup E$ is at least as good as $A_i \cup B_i \cup E$ for all i, $1 \leq i \leq n$ and any $E \subset \mathbb{R}(y, x)$

and from (4)-(7) that

(9) $A_{i+1} \cup \ldots \cup A_{n+1} \cup B_{i+1} \cup \ldots \cup B_{n+1} \cup C_0 \cup \ldots \cup C_{i-1} \cup D \subset \mathbb{R}(y, x)$ for all i, $1 \leq i \leq n$.

Letting $E=A_{i+1}\cup\ldots\cup A_{n+1}\cup B_{i+1}\cup\ldots\cup B_{n+1}\cup C_0\cup\ldots\cup C_{i-1}\cup D$, (8) and (9) imply that

(10) $C_i\cup[A_{i+1}\cup\ldots\cup A_{n+1}\cup B_{i+1}\cup\ldots\cup B_{n+1}\cup C_0\cup\ldots\cup C_{i-1}\cup D]$ is at least as good as $A_i\cup B_i\cup[A_{i+1}\cup\ldots\cup A_{n+1}\cup B_{i+1}\cup\ldots\cup B_{n+1}\cup C_0\cup\ldots\cup C_{i-1}\cup D]$ for all *i*, $1 \le i \le n$ (see Diagram 10.10.1).

Transitivity and (10) yield that

(11) $C_n\cup A_{n+1}\cup B_{n+1}\cup C_0\cup\ldots\cup C_{n-1}\cup D$ is at least as good as $A_1\cup B_1\cup A_2\cup\ldots\cup A_{n+1}\cup B_2\cup\ldots\cup B_{n+1}\cup C_0\cup D$

and since $A_{n+1}=B_{n+1}=C_0=\varnothing$ (4-6), line (11) is equivalent to (see Diagram 10.10.1)

(12) $C_1\cup\ldots\cup C_n\cup D$ is at least as good as $A_1\cup\ldots\cup A_n\cup B_1\cup\ldots\cup B_n\cup D$. Q.E.D.

To show that Lemma 5.1 is true, we now need to prove

Lemma 5.1.2: Condition α entails Condition β.

Proof. Let

(1) $\mathbb{W}_x$, $\mathbb{W}_y$, $\mathbb{W}_z$ be any three welfare levels such that $x > y > z$;
(2) $r = x\text{-}y$.

Let $A_1,\ldots, A_{r+1}$ and $B_1,\ldots, B_{r+1}$ be any two sequences of populations, $m_0,\ldots, m_r$ any sequence of integers, and *f* a function satisfying

(3) $m_0 > 0$;
(4) $f(m_i)=m_0+m_1+\ldots+m_i$, for all *i*, $0 \le i \le r$,
(5) $m_i \ge f(m_{i-1})$ satisfies Condition α for $\mathbb{W}_{x-(i-1)}$, $\mathbb{W}_z$, and $f(m_{i-1})$ for all *i*, $1 \le i \le r$;
(6) $A_i\subset\mathbb{W}_{x-(i-1)}$, $N(A_i)=f(m_{i-1})$ for all *i*, $1 \le i \le r+1$;
(7) $B_i\subset\mathbb{W}_z$, $N(B_i)=m_i$, for all *i*, $1 \le i \le r$; $B_{r+1}=\varnothing$.

Finally, let

(8) D be any population such that D⊂ℝ(z, y+1);

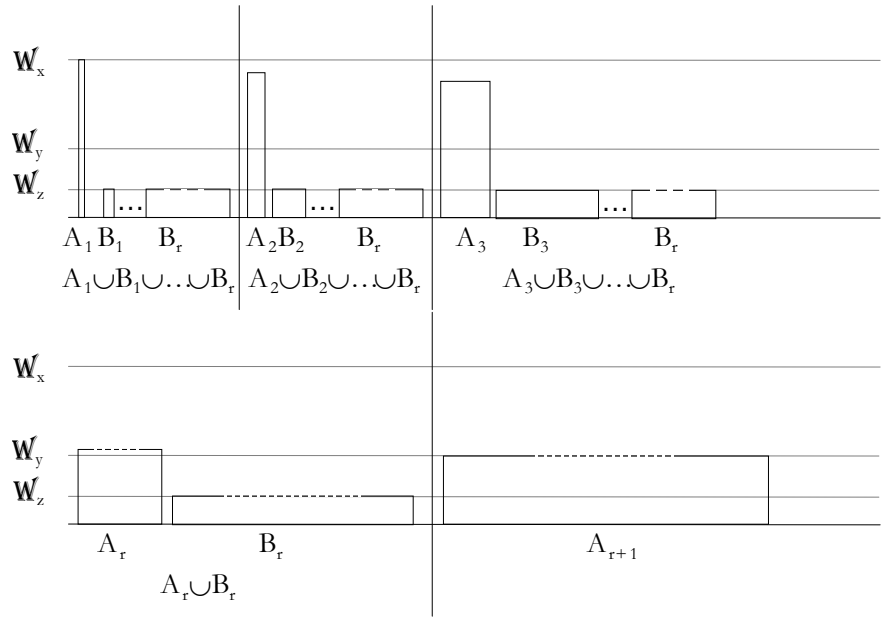

Diagram 10.10.2

The above diagram shows a selection of the involved populations in a case where $r \geq 4$. Population D is omitted throughout.

We can conclude from (3)-(7) that $N(B_1 \cup \ldots \cup B_r) > m_0 = N(A_1)$. Consequently, since $\mathbb{W}_x$, $\mathbb{W}_y$, and $\mathbb{W}_z$ can be any welfare levels such that $x > y > z$, and D can be any population consisting of lives with welfare ranging from $\mathbb{W}_z$ to $\mathbb{W}_{y+1}$, we can show that Condition α implies Condition β by showing that $A_{r+1} \cup D$ is at least as good as $A_1 \cup B_1 \cup \ldots \cup B_r \cup D$. This suffices since $A_1, \ldots, A_{r+1}$, $B_1, \ldots, B_r$, and D are arbitrary populations satisfying (6)-(8).

From (3)-(7) and Condition α, it follows that

(9) $A_{i+1} \cup E$ is at least as good as $A_i \cup B_i \cup E$ for all i, $1 \leq i \leq r$ and any $E \subset \mathbb{R}(z, y+1)$.

and from (7) and (8) that

(10) $B_{i+1} \cup \ldots \cup B_{r+1} \cup D \subset \mathbb{R}(z, y+1)$ for all i, $1 \leq i \leq r$.

Consequently, letting $E=B_{i+1}\cup\ldots\cup B_{r+1}\cup D$, (9) and (10) imply that

(11) $A_{i+1}\cup[B_{i+1}\cup\ldots\cup B_{r+1}\cup D]$ is at least as good as $A_i\cup B_i\cup[B_{i+1}\cup\ldots\cup B_{r+1}\cup D]$ for all *i*, $1 \leq i \leq r$ (see Diagram 10.10.2).

Transitivity and (11) yield that

(12) $A_{r+1}\cup B_{r+1}\cup D$ is at least as good as $A_1\cup B_1\cup\ldots\cup B_{r+1}\cup D$

and since $B_{r+1}=\varnothing$ (7), line (12) is equivalent to (see Diagram 10.10.2)

(13) $A_{r+1}\cup D$ is at least as good as $A_1\cup B_1\cup\ldots\cup B_r\cup D$. Q.E.D.

It follows trivially from lemma 5.1.1 and 5.1.2 that lemma 5.1 is true. Q.E.D.

10.10.2 Lemma 5.2

Lemma 5.2: The General Non-Extreme Priority Condition implies Condition δ.

Condition δ: For any $\mathbb{W}_z$, $z < 0$, and any number of lives $m > 0$, there is a positive welfare level $\mathbb{W}_u$, and a positive welfare range $\mathbb{R}(1, y)$, $u > y$, and a number of lives $n > 0$ such that **if** A⊂$\mathbb{W}_x$, $x \geq u$, B⊂$\mathbb{R}(1, y)$, N(A)=N(B)=n, C⊂$\mathbb{W}_z$, D⊂$\mathbb{W}_3$, N(C)=N(D)=m, **then,** for any E, A∪C∪E is at least as good as B∪D∪E, other things being equal.

We shall prove lemma 5.2 by first proving

Lemma 5.2.1: The General Non-Extreme Priority Condition implies *Condition* χ.

Condition χ: For any $\mathbb{W}_z$, $z < 0$, there is a positive welfare level $\mathbb{W}_u$, and a positive welfare range $\mathbb{R}(1, y)$, $u > y$, and a number of lives $n > 0$ such that **if** A⊂$\mathbb{W}_x$, $x \geq u$, B⊂$\mathbb{R}(1, y)$, N(A)=N(B) =n, C⊂$\mathbb{W}_z$, D⊂$\mathbb{W}_3$, N(C)=N(D)=1, **then,** for any E, A∪C∪E is at least as good as B∪D∪E, other things being equal.

Proof: Let

(1) $\mathbb{W}_z$ be any welfare level such that $z < 0$;
(2) r=3-z;
(3) $\mathbb{W}_{u_i}$ be a positive welfare level, $\mathbb{R}(1, v_i)$ be a positive welfare range, and n_i a number of lives which satisfy the General Non-Extreme Priority Condition for $\mathbb{W}_{z+(i-1)}$ for all i, $1 \leq i \leq r$;
(4) $\mathbb{W}_u$ be a welfare level such that u equals the maximal element in $\{u_i: 1 \leq i \leq r\}$;
(5) $\mathbb{W}_x$ be a welfare level such that $x \geq u$;
(6) y be a number such that y equals the minimal element in $\{v_i: 1 \leq i \leq r\}$.

Let $A_1,\ldots, A_{r+1}$, $B_0,\ldots, B_r$, and $C_1,\ldots, C_{r+1}$, be any three sequences of populations satisfying

(7) $A_i \subset \mathbb{W}_x$, $N(A_i)=n_i$, for all i, $1 \leq i \leq r$; $A_{r+1}=\varnothing$;
(8) $B_i \subset \mathbb{R}(1, y)$, $N(B_i)=n_i$, for all i, $1 \leq i \leq r$; $B_0=\varnothing$;
(9) $C_i \subset \mathbb{W}_{z+(i-1)}$, $N(C_i)=1$, for all i, $1 \leq i \leq r+1$.

Finally, let

(10) E be any population.

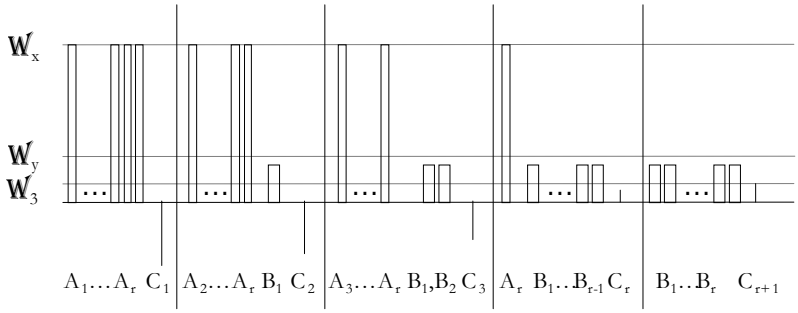

Diagram 10.10.3

The above diagram shows a selection of the involved populations in a case where $r \geq 4$. Population E is omitted throughout.
Since $\mathbb{W}_z$ can be any negative welfare level (by (1)), and $\mathbb{W}_x$ can be any welfare level at least as high as $\mathbb{W}_u$ (by (5)), and since it follows from (3) and (6) that $\mathbb{R}(1, y)$ is a welfare range such that $u > y$, we can show that lemma 5.2.1 is true by showing that $A_1 \cup \ldots \cup A_r \cup C_1 \cup E$ is at least as good as $B_1 \cup \ldots \cup B_r \cup C_{r+1} \cup E$. This suffices since $A_1,\ldots, A_r$, $B_1,\ldots, B_r$, $C_1,\ldots, C_{r+1}$, and E are arbitrary populations satisfying (7)-(10).
The General Non-Extreme Priority Condition and (3)-(9) imply that

(11) $A_i \cup C_i \cup F$ is at least as good as $B_i \cup C_{i+1} \cup F$ for all i, $1 \leq i \leq r$ and any population F.

Letting $F=A_{i+1} \cup \ldots \cup A_{r+1} \cup B_o \cup \ldots \cup B_{i-1} \cup E$, it follows from (11) that

(12) $A_i \cup C_i \cup [A_{i+1} \cup \ldots \cup A_{r+1} \cup B_o \cup \ldots \cup B_{i-1} \cup E]$ is at least as good as $B_i \cup C_{i+1} \cup [A_{i+1} \cup \ldots \cup A_{r+1} \cup B_o \cup \ldots \cup B_{i-1} \cup E]$ for all i, $1 \leq i \leq r$ (see Diagram 10.10.3).

Transitivity and (12) yield that

(13) $A_1 \cup C_1 \cup A_2 \cup \ldots \cup A_{r+1} \cup B_o \cup E$ is at least as good as $B_r \cup C_{r+1} \cup A_{r+1} \cup B_o \cup \ldots \cup B_{r-1} \cup E$.

and since $A_{r+1}=B_0=\varnothing$ (7-8), line (13) is equivalent to (see Diagram 10.10.3)

(14) $A_1 \cup \ldots \cup A_r \cup C_1 \cup E$ is at least as good as $B_1 \cup \ldots \cup B_r \cup C_{r+1} \cup E$. Q.E.D.

To show that Lemma 5.2 is true, we now need to prove

Lemma 5.2.2: Condition χ implies Condition δ.

Proof: Let

(1) $\mathbb{W}_z$ be any welfare level such that $z < 0$;
(2) m be any number such that $m > 0$;
(3) $\mathbb{W}_u$ be a positive welfare level, $\mathbb{R}(1, y)$ be a positive welfare range, and n a number of lives which satisfy Condition χ for $\mathbb{W}_z$;
(4) $\mathbb{W}_x$ be a welfare level such that $x \geq u$.

Let $A_1,\ldots, A_{m+1}$, $B_0,\ldots, B_m$, $C_1,\ldots, C_{m+1}$, and $D_0,\ldots, C_m$, be any four sequences of populations satisfying

(5) $A_i \subset \mathbb{W}_x$, $N(A_i)=n$, for all i, $1 \leq i \leq m$; $A_{m+1}=\varnothing$;
(6) $B_i \subset \mathbb{R}(1, y)$, $N(B_i)=n$, for all i, $1 \leq i \leq m$; $B_0=\varnothing$;
(7) $C_i \subset \mathbb{W}_z$, $N(C_i)=1$, for all i, $1 \leq i \leq m$; $C_{m+1}=\varnothing$;
(8) $D_i \subset \mathbb{W}_3$, $N(D_i)=1$, for all i, $1 \leq i \leq m$; $D_0=\varnothing$.

Finally, let

(9) E be any population.

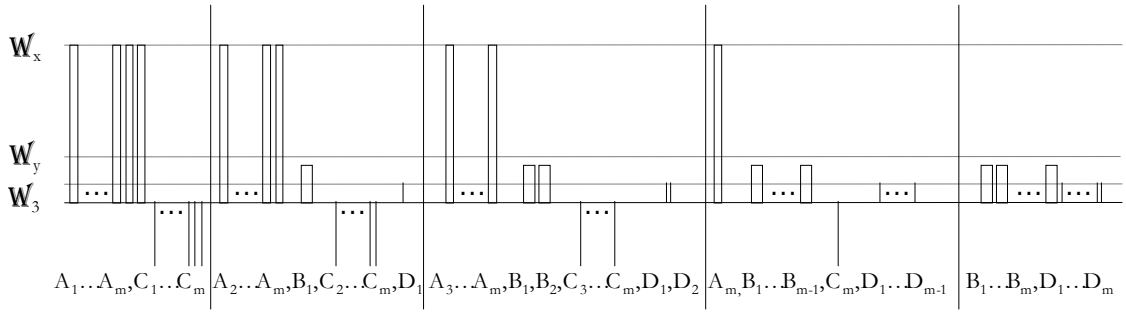

Diagram 10.10.4

The above diagram shows a selection of the involved populations in a case where $m \geq 4$. As before, population E is omitted throughout.

Since $\mathbb{W}_z$ can be any negative welfare level (by (1)), and $\mathbb{W}_x$ can be any welfare level at least as high as $\mathbb{W}_u$ (by (5)), and m can be any number of lives greater than zero, and $\mathbb{R}(1, y)$ is a welfare range such that $u > y$, and n is a number greater than zero (by (3)), we can show that lemma 5.2.2 is true by showing that $A_1 \cup \ldots \cup A_m \cup C_1 \cup \ldots \cup C_m \cup E$ is at least as good as $B_1 \cup \ldots \cup B_m \cup D_1 \cup \ldots \cup D_m \cup E$. This suffices since $A_1, \ldots, A_m$, $B_1, \ldots, B_m$, $C_1, \ldots, C_m$, $D_1, \ldots, D_m$, and E are arbitrary populations satisfying (5)-(9).

It follows from (3)-(8) and Condition χ that

(10) $A_i \cup C_i \cup F$ is at least as good as $B_i \cup D_i \cup F$ for all i, $1 \leq i \leq m$, and any population F

which, for $F = A_{i+1} \cup \ldots \cup A_{m+1} \cup C_{i+1} \ldots \cup C_{m+1} \cup B_0 \cup \ldots \cup B_{i-1} \cup D_0 \ldots \cup D_{i-1} \cup E$, in turn implies

(11) $A_i \cup C_i \cup [A_{i+1} \cup \ldots \cup A_{m+1} \cup C_{i+1} \ldots \cup C_{m+1} \cup B_0 \cup \ldots \cup B_{i-1} \cup D_0 \ldots \cup D_{i-1} \cup E]$ is at least as good as $B_i \cup D_i \cup [A_{i+1} \cup \ldots \cup A_{m+1} \cup C_{i+1} \ldots \cup C_{m+1} \cup B_0 \cup \ldots \cup B_{i-1} \cup D_0 \ldots \cup D_{i-1} \cup E]$ for all i, $1 \leq i \leq m$ (see Diagram 10.10.4).

Transitivity and (11) yield that

(12) $A_1 \cup C_1 \cup A_2 \cup \ldots \cup A_{m+1} \cup C_2 \ldots \cup C_{m+1} \cup B_0 \cup D_0 \cup E$ is at least as good as $B_m \cup D_m \cup A_{m+1} \cup C_{m+1} \cup B_0 \cup \ldots \cup B_{m-1} \cup D_0 \ldots \cup D_{m-1} \cup E$

and since $A_{m+1}=B_0=C_{m+1}=D_0=\varnothing$ (by (5)-(8)), line (12) is equivalent to (see Diagram 10.10.4)

(13) $A_1 \cup \ldots \cup A_m \cup C_1 \ldots \cup C_m \cup E$ is at least as good as $B_1 \cup \ldots \cup B_m \cup D_1 \ldots \cup D_m \cup E$. Q.E.D.

It follows trivially from lemma 5.2.1 and 5.2.2 that lemma 5.2 is true. Q.E.D.

10.10.3 Lemma 5.3

Finally, we shall show that the fourth theorem is true by proving

> ***Lemma 5.3***: There is no population axiology which satisfies the Egalitarian Dominance Condition, the Weak Non-Sadism Condition, the Weak Quality Addition Condition, Condition β, and Condition δ.

Proof. We show that the contrary assumption leads to a contradiction. Let

(1) $\mathbb{W}_z$ be a welfare level and m a population size which satisfy the Weak Non-Sadism Condition;
(2) $\mathbb{W}_u$ be a welfare level, $\mathbb{R}(1, y)$ a welfare range, and n a number of lives, which satisfy Condition δ for $\mathbb{W}_z$ and m;
(3) $B_1 \subset \mathbb{W}_3$, $B_2 \subset \mathbb{W}_3$, $N(B_1)=n$, $N(B_2)=m$;
(4) $\mathbb{R}(w, t)$ and $\mathbb{R}(1, v)$, $w > v$, be two welfare ranges, and p a population size, which satisfy the Weak Quality Addition Condition for B;
(5) Let $\mathbb{W}_x$ be a welfare level such that $x \geq w$ and $x \geq u$;
(6) $A \subset \mathbb{W}_x$, $N(A)=p$;
(7) $H \subset \mathbb{W}_x$, $N(H)=n$;
(8) $E \subset \mathbb{W}_z$, $N(E)=m$.

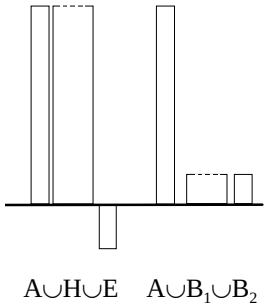

Diagram 10.10.5

It follows from the definition of a welfare range that $\mathbb{W}_3 \subset \mathbb{R}(1, y)$. Accordingly, from (3) we know that $B_1 \subset \mathbb{R}(1, y)$. Consequently, from (2), (3), (7), (8), and Condition δ we get that

(9) A∪H∪E is at least as good as $A \cup B_1 \cup B_2$ (see Diagram 10.10.5).

Let

(10) $r > n+p$ be a number of lives which satisfies Condition β for the three welfare levels $\mathbb{W}_x$, $\mathbb{W}_2$, and $\mathbb{W}_1$ and for $n+p$ lives at $\mathbb{W}_x$;
(11) q be any number of lives such that $q \geq m$ and $q \geq r$;
(12) $G \subset \mathbb{W}_2$, $N(G)=n+p+r$;
(13) $I \subset \mathbb{W}_1$, $N(I)=q-r$;
(14) $F \subset \mathbb{W}_1$, $N(F)=r$.

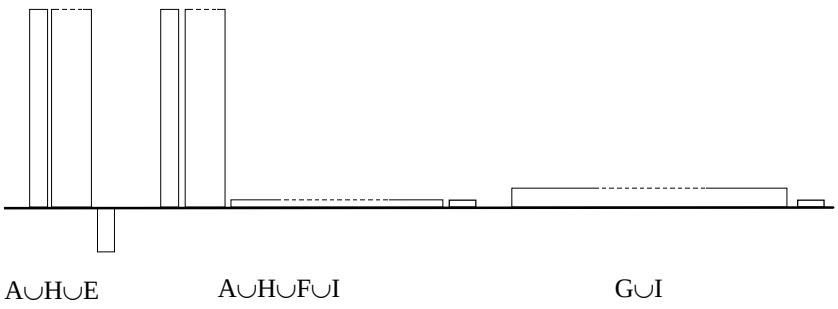

Diagram 10.10.6

Since $A \cup H \subset \mathbb{W}_x$, and $N(A \cup H)=n+p$ (by (6) and (7)), and $I \subset \mathbb{R}(1, 3)$, it follows from (10)-(14) and Condition β that

(15) G∪I is at least as good as A∪H∪F∪I (see Diagram 10.10.6).

Since the F- and the I-lives have positive welfare (by (13) and (14)), it follows from (1), (8) and the Weak Non-Sadism Condition that

(16) A∪H∪F∪I is at least as good as A∪H∪E (see Diagram 10.10.6).

By transitivity, it follows from (15) and (16) that

(17) G∪I is at least as good as A∪H∪E.

Let

(18) C⊂$\mathbb{W}_3$, N(C)=p+q-m.

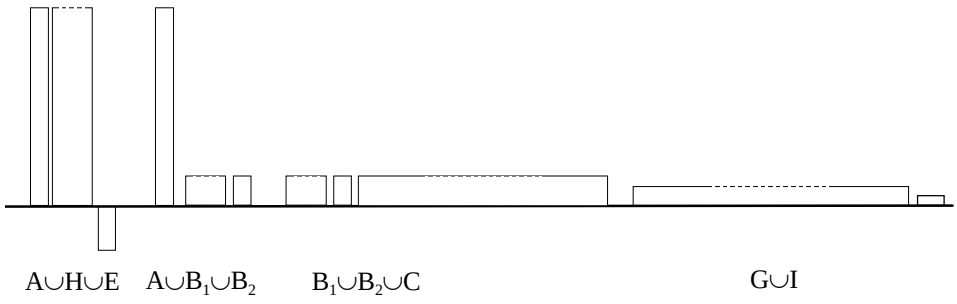

Diagram 10.10.7

Since $\mathbb{W}_3$⊂$\mathbb{R}$(1, v), we can conclude that C⊂$\mathbb{R}$(1, v), and since $x \geq u$ (by (5)), and A⊂$\mathbb{W}_x$ (by (6)), it follows from (4) and the Weak Quality Addition Condition that

(19) A∪B_1∪B_2 is at least as good as B_1∪B_2∪C (see Diagram 10.10.7).

Since B_1∪B_2∪C⊂$\mathbb{W}_3$ (by (3) and (18)) and G∪I⊂$\mathbb{W}_1$∪$\mathbb{W}_2$, (by (12) and (13)) and N(B_1∪B_2∪C)=N(G∪I), the Egalitarian Dominance Condition implies that

(20) G∪I is worse than B_1∪B_2∪C (see Diagram 10.10.7).

By transitivity, it follows from (19) and (20) that

(21) G∪I is worse than A∪B_1∪B_2

and from (9) and (21) that

(22) G∪I is worse than A∪H∪E

which contradicts (17). Q.E.D.

It follows trivially from lemma 5.1, 5.2 and 5.3 that the fourth impossibility theorem is true. Q.E.D.

11

Normative Population Theory

11.1 From Axiology to Morality

So far, our discussion has concentrated on how to evaluate populations in regard to their goodness, that is, how to order populations by the relation "is at least as good as". It is natural to assume that such an ordering partly determines which acts are morally right and wrong. As a matter of fact, most of the theories that we have discussed were originally formulated as normative theories, but we have only discussed the axiological part of these theories. These theories also include a bridging principle from the axiological level to the normative level, namely some form of consequentialism. The most common form of consequentialism is act-consequentialism according to which, roughly, an action is right if and only if it maximises the good. More exactly, we shall define act-consequentialism as follows:

> *Pure Act-Consequentialism*: An action is right (obligatory) if and only if its outcome is at least as good as (better than) that of every alternative. An action is wrong if and only if it is not right.[1]

As we mentioned in section 8.3, some populations can be seen as outcomes of actions, namely populations that consist of all the lives that are part of the outcomes. Which lives are included in the outcome of an action depends, of course, on what we consider the morally relevant outcome of an action. The three most common answers are the possible world that would be the case if the action were

[1] Cf. Carlson (1995), p. 13. My formulation differs slightly from Carlson's (the brief discussion of consequentialism below draws heavily on Carlson's work). The definition of consequentialism that we have suggested has counter-intuitive implications in cases involving outcomes that are incommensurable in value. I don't think, however, that consequentialism necessarily presupposes a complete ordering of the outcomes in a choice situation. See Carlson (1995), p. 25, fn. 48, for some suggested revisions of the definition of consequentialism that can handle incommensurable outcomes.

performed, the total future state of the world that would be the case if the action were performed, and the causal consequences of the action.[2] These three views correspond to three types of populations, namely populations that consist of all the past, present and future lives, or all the present and future lives, or all the lives that are causally affected or consequences of the action. Now, there is nothing in the theorems in chapter 10 that rules out that the involved populations are of these types. And, of course, inconsistent evaluations of outcomes is a devastating problem from a consequentialist perspective.

Some theorists might be willing to abandon transitivity of the relations "is at least as good as" and "is better than" to save consistency. How would that affect the possibility of a consequentialist population morality? The evaluations involved in the impossibility theorems exhibit the following structure: Outcome A_1 is at least as good as A_2, which is at least as good as A_3, …, which is at least as good as A_n, which in turn is better than A_1. In such cases, is there an outcome which is at least as good as all alternative outcomes? We cannot really tell, since without transitivity, we don't know, for example, how A_n relates to A_2. If one could show that A_n is equally as good as A_2 to A_{n-1}, then A_n would be at least as good as all alternative outcomes, and the action which has A_n as outcome would be declared right by consequentialism. So perhaps there is still a possibility for a consequentialist to eschew the impossibility theorems.

A problem, however, is that one could replace the relation "is at least as good as" with "is better than" without much loss of intuitive plausibility in the affected adequacy conditions. This move would yield the following structure of the theorems: Outcome A_1 is better than A_2, which is better than A_3, …, which is better than A_n, which in turn is better than A_1. In cases involving such evaluations it is neither true of any outcome that it is at least as good as all the other outcomes, nor is it true of any outcome that it isn't worse than any other outcome. Consequently, in respect to such cases, consequentialism implies, implausibly, that

[2] See Carlson (1995), pp. 10-12, and ch. 4, for an extensive discussion of the morally relevant outcome of an action.

all the available actions are wrong. Consequentialism requires some form of acyclicity of the ranking of outcomes in a choice situation.[3]

But perhaps we have here a plausible interpretation of the results in chapter 10. In a choice situation involving alternatives like those in the impossibility theorems, we are facing a *moral dilemma*: whatever act we perform we are going to act wrongly.[4] We could claim that the existence of moral dilemmas is part of our moral phenomenology and that it is not surprising that we should face a moral dilemma in situations involving such awesome alternatives as are involved in the theorems.

Although this is a possible interpretation of the impossibility theorems, I don't find much comfort in it. As Jan Österberg suggests, any plausible morality is separately satisfiable:

> *The Condition of Separate Satisfiability*: For any agent and any situation, it is logically possible for her not to act morally wrong.[5]

It is reasonable to claim that it should at least be logically possible for a person not to do the wrong thing. Normative theories which imply that there are moral dilemmas in which all the available actions are wrong, imply that there are situations where it is not even a logical possibility for an agent to do what the theory requires of her.[6] Consequently, since an adequate morality should be separately satisfiable, the impossibility theorems challenge the existence of an acceptable consequentialist morality.

There are other versions of consequentialism apart from act-consequentialism, such as rule-consequentialism according to which an act is right if and only if it can be subsumed under a rule whose general acceptance (or general implementation) would give the best result. We have focused the discussion above on act-

---

[3] The acyclicity I have in mind is that if A_1 is better than A_2, A_2 is better than A_3, …, A_{n-1} is better than A_n, then A_1 is not worse than A_n. Cf. Sen (1970), p. 47. For a discussion of cyclical evaluations, see Danielsson (1996), Carlson (1996), and Rabinowicz (2000).

[4] Following Vallentyne (1988), we could call a dilemma of the above mentioned type a "prohibition dilemma". There are also "obligation dilemmas", that is, situations where more than one action is obligatory.

[5] See Österberg (1988), pp. 127, 145-6. My formulation is weaker than Österberg's which is formulated in terms of the possibility of an agent to act morally right.

[6] Österberg (1988), p. 146, suggests an interesting argument to the effect that the Condition of Separate Satisfiability is entailed by the common idea that "ought" implies "can".

consequentialism, since it is the most popular version and since we think it is the most defensible version (although we shall not defend the latter claim here). It should be clear, however, that the discussion above applies, *mutatis mutandis*, equally well to rule-consequentialism, and, presumably, to other forms of consequentialism too.

Could we solve the problems raised by the impossibility theorems by rejecting consequentialism? I don't think it is that easy. Consequentialists assume that all morally relevant factors can be taken into account in the value of outcomes. One might think that certain moral relevant factors cannot be taken into account in such a manner but should be incorporated on the deontic level in terms of actions that are right or wrong by virtue of being of a certain type. Examples are rights, promises, and actions that involve great personal sacrifice for the agent. One may judge actions that involve violations of people's rights or the breaking of promises as wrong, and actions that involve great personal sacrifice as supererogatory, irrespective of how good the consequences of those actions would be. It is not clear, however, that such theories cannot be formulated as extensionally equivalent consequentialist theories since it is possible to incorporate a wide range of non-welfarist values in a consequentialist theory.[7] Feldman's theory is a case in point. At any rate, some of those critics of consequentialism that take this line do take the consequences of actions into account but they think that there are deontic "constraints" that exclude actions of certain types, or deontic "options" that make certain types of actions permissible. The remaining alternatives are, however, evaluated in a consequentialist manner. They accept what we could call *Ceteris Paribus* Act-Consequentialism:

> *Ceteris Paribus Act-Consequentialism*: Other things being equal, an action is right (obligatory) if and only if its outcome is at least as good as (better than) that of every alternative. An action is wrong if and only if it is not right.

---

[7] For a discussion of whether any morality can be formulated as a consequentialist morality, see Carlson (1995), Danielsson (1988), Vallentyne (1988), and Bykvist (1996).

In other words, if a choice situation doesn't involve actions that are right or wrong by virtue of a certain deontic constraint or option, then the normative status of the actions are determined by the value of their respective outcomes. Assuming that the involved deontic constraints and options don't concern the number and the welfare of lives in populations that are outcomes of actions (which is a questionable assumption, however), this view clearly runs into the same problem as Pure Act-Consequentialism in respect to the inconsistent or cyclical evaluations of outcomes discussed above.

What the above discussion shows, I think, is that as long as welfare is only taken into account in a consequentialist manner, the impossibility theorems challenge the existence of an acceptable population morality. It is natural, then, to take the next step and ask whether this also holds true for theories that partly or completely take welfare into account in a non-consequentialist manner, that is, theories that partly or completely take welfare into account directly on the normative level instead of taking the route over an ordering of outcomes in regard to their "welfarist" goodness. For example, one could claim that it is always wrong to increase a population with lives not worth living when it is avoidable, or that in the choice between giving a small benefit to one person or a great benefit to many people, one ought to do the latter. Narveson's later theory, quoted in section 8.3, is an example of a more developed effort in this direction. And there are other prominent normative theories for which it is unclear whether the impossibility theorems pose a problem, such as David Gauthier's Mutual Advantage Contractarianism and Richard Arneson's theory of equality of opportunity for welfare.[8] As a matter of fact, David Boonin-Vail, who has proposed a non-consequentialist population morality, suggests that whereas there is no satisfying solution to the axiological Mere Addition Paradox (which he calls "the Goodness Paradox"), the normative version of this paradox (which he calls "the Oughtness Paradox") can be solved and this result deprives the former paradox of its moral significance.[9]

---

[8] See Gauthier (1986), p. 299 and Heath (1997), for the former kind of theory, and Arneson (1989) for the latter. I discuss Gauthier's and Heath's suggestion at length in Arrhenius (2000b).
[9] Boonin-Vail (1996), pp. 279-80, 307.

Although I agree with Boonin-Vail's last point regarding the significance of a solution to the Oughtness Paradox, I don't think it is easier to solve than the Goodness Paradox. As one would suspect, the normative theories mentioned above, including Boonin-Vail's own suggestion, have counter-intuitive implications analogous to the ones of the axiological theories that we have discussed. We shall not discuss these different theories case by case, however, but take a more direct route: We shall investigate whether one can construct normative versions of the axiological impossibility theorems.

11.2 A Normative Structure

We shall now suggest a framework for constructing normative theorems analogous to the axiological theorems in chapter 10. Let's first take a look at a problem facing any such project. Consider the Mere Addition Paradox (see Diagram 6.1.2) again and assume that our normative evaluations are, as I think many would agree, as follows (assuming now that the populations in question are outcomes of actions): In the choice between population A and A∪B, it is permissible to choose either one; in the choice between A∪B and C, we ought to choose C; in the choice between C and A, we ought to choose A; and in the choice among A, A∪B, and C, we ought to choose A, and it would be wrong to choose A∪B or C. Have we contradicted ourselves? As a matter of fact, we haven't. As long as we don't add any more restrictions on our normative evaluations, there is no contradiction involved in the above evaluation. This suggests that evaluations that are contradictory on the axiological level may not be so on the normative level, the reason being that there is no analogue to transitivity on the normative level.[10]

One might think otherwise, however. Gregory Kavka, for example, has suggested the following transitivity principle for moral permissibility: "If it would be permissible to do A if A and B are the alternatives, and would be permissible to do B if B and C were the alternatives, then it is permissible to do A if A, B, and C are the alternatives."[11] Given this requirement on normative judgements, the above evaluations are inconsistent. Since it is permissible to choose C in the choice between C and A∪B (if an action is obligatory, it is of course permissible), and

---

[10] Cf. Boonin-Vail (1996), p. 285.

[11] Kavka (1982), p. 100, fn. 16. Boonin-Vail (1996), p. 283, discusses a similar principle.

permissible to choose A∪B in the choice between A∪B and A, it follows from Kavka's principle that it is permissible to choose C in the choice among A, A∪B, and C. But we said above that in the latter situation, we ought to choose A and it would be wrong to choose C. So we are back in trouble again.

I'm sceptical about Kavka's transitivity principle for moral permissibility, however. Consider the following counter-example suggested by Derek Parfit. Suppose that a woman at some point faces the following options:

> P: Having a handicapped child.
> Q: Having no child.

As Parfit points out, "[i]f this child's handicap would not be severe, and we make certain other assumptions, we can plausibly believe that it would be permissible for the woman to choose either P or Q …".[12] Moreover, this evaluation is, arguably, still plausible if P is replaced by the following alternative:

> R: Having the same child, but in a way that would ensure that he wouldn't be handicapped.

Assume now that all of these three alternatives are available to the woman. According to Kavka's transitivity principle, since P is permissible in the choice between P and Q, and since Q is permissible in the choice between Q and R, it follows that P is permissible in the choice among P, Q, and R. But, as Parfit writes, "[w]e can plausibly believe that, if R were also possible, it would be wrong for this woman to choose P rather than R".[13]

The problem that Kavka's principle runs into suggests an important difference between axiological and normative evaluations. It is usually thought that the intrinsic goodness of an outcome doesn't depend on its relation to other outcomes. If an outcome A is good, or better than another outcome B, then we usually think that this holds irrespective of whether A and B are alternative outcomes in some choice situation, or whether there are other alternative outcomes available. As it is

---

[12] Parfit (1996), p. 311.
[13] Parfit (1996), p. 311.

often put, the intrinsic value of a state of affairs is independent of its relation to other distinct states of affairs. The normative status of actions, however, depends on what other actions are available in a choice situation. For example, it is permissible to inflict harm on somebody if the only other alternative is to inflict even more harm, but if harming is avoidable, then it is wrong.

Notice that I'm neither suggesting that this structural difference is the only difference between axiological and normative concepts, nor that it is necessarily the most prominent one. It is, however, the difference that is important in the present context.[14] Moreover, there are ways of understanding value-concepts that fall into the normative category in my sense, such as using "A is better than B" as synonymous with "A ought to be chosen in a situation where A and B are the alternatives", "A is choice-worthy in a situation where A and B are the alternatives", and the like. As a matter of fact, I think this possible understanding of value-concepts might explain why some theorists have been willing to abandon the transitivity of "better than". I tend to agree with those who take the transitivity of (intrinsically) "better than" as a matter of logic, as part of the meaning of "better than". However, if one takes "A is better than B" as synonymous with, for example, "A ought to be chosen in a situation where A and B are the alternatives", then one can give reasons for the failure of the transitivity of "better than", namely those we gave above in regard to the explicitly normative concepts.

Should we then conclude that Boonin-Vail is right in his conjecture that the Oughtness paradox might be solvable although the Goodness paradox is not? No, I don't think so. What the above discussion shows, I take it, is that we need to take the context dependence of normative status into account when we formulate normative adequacy conditions by partly formulating them in terms of certain features of the choice situation.[15] Consider the following pattern for a normative condition:

---

[14] For a discussion of some other differences between these kinds of concepts, see von Wright (1963), ch. 1, sect. 4, and Danielsson (1999).

[15] We are drawing on a suggestion made by Sen (1995), p. 5, in response to certain criticisms of Arrow's impossibility theorem.

> (i) If action h_1 is of type G and action h_2 is of type B, and both h_1 and h_2 are available in a certain choice situation, then h_2 is forbidden in this choice situation.

The actions P and R in the example above fit this pattern, as the quote from Parfit suggests. We can formulate a normative version of the Egalitarian Dominance Condition along these lines:

> *The Normative Egalitarian Dominance Condition*: If population A is a perfectly equal population of the same size as population B, and every person in A has higher welfare than every person in B, then, in any situation involving a choice between A and B, it is wrong to choose B, other things being equal.

This condition is, I think, as plausible as its axiological counterpart. The *ceteris paribus* condition involved here is a natural extension of the *ceteris paribus* condition used in the discussion of different axiologies (see section 3.1.1): There are neither any constraints (for example, promise-keeping) nor options (for example, great personal sacrifice for the agent which is beyond the call of duty), nor any non-welfarist values in the outcomes (for example, cultural diversity) that gives us a reason to (not) choose one or the other of the involved actions. The only reasons for choosing one or the other of the involved actions arise from the welfare of the lives in the involved populations. Consider a situation where you could, at no cost to yourself (you might even be among the beneficiaries), and without violating any other duties or compromising any other values, choose an outcome in which everybody is equally well off, and better off as compared to another outcome involving the same number of people. Surely it would be wrong to choose the latter outcome in this situation.

We could formulate normative versions of the other axiological adequacy conditions using pattern (i) above, and I think the resulting normative conditions would be compelling. For the theorems we shall prove, however, it suffices to use the following logically weaker construction:

(ii) If action h_1 is of type G and action h_2 is of type B, and both h_1 and h_2 are available in a certain choice situation, and h_1 is wrong in this choice situation, then h_2 is also wrong in this choice situation.

Assume that next Sunday you can help either Erik or Krister with their gardening, and that they both need your help equally as much, and that you haven't promised any one of them your help, and so forth. Now, it is reasonable to claim that in a situation involving these two alternatives, if it would be wrong of you to help Krister, then it would also be wrong of you to help Erik. It could be wrong of you to help Krister if you have promised your elderly aunt to help her next Sunday with the much needed gardening at her house (assuming that the involved acts are mutually exclusive). But if that is the case, then it would also be wrong of you to help Erik.

Apart from the Normative Egalitarian Dominance Condition, we shall formulate all the adequacy conditions used in the theorems below along the lines of pattern (ii). By thus formulating normative versions of the axiological adequacy conditions, we can prove normative versions of all of the axiological theorems in chapter 10. This would be quite repetitious and tiresome, however, so we shall restrict ourselves to two theorems. We shall show that there is no separately satisfiable morality that satisfies the normative versions of the axiological adequacy conditions used in the second and third theorem in the preceding chapter. As this demonstration will make clear, one can similarly make normative versions of the first and fourth theorem too.

For the purpose of proving the theorems, it will be useful to introduce some new terminology. We shall say that a *population morality* at least assigns the normative status wrong to some actions in some possible choice situations. A *choice situation* $\mathscr{C}$ is a set of at least two mutually incompatible actions available to a certain individual or group of individuals at a certain time.[16] Let $\mathscr{A}(A)\subset\mathscr{C}$ be the set of all mutually incompatible actions in a choice situation $\mathscr{C}$ such that if any one of them were performed, then population A would be the case, that is, population A consists of

[16] For the sake of simplicity, we are assuming that a unique choice situation corresponds to any set of agent- and time-identical actions. There are, of course, a number of intricate problems regarding how to individuate actions, what it means for an action to be available to an agent, or a group of agents, and the like. These problems clearly fall outside the scope of this essay, however.

all the lives that would be part of the normatively relevant outcome that would be the case if any action in $\mathscr{A}$(A) were performed. For example, population A could consist of all the lives that are part of the possible world that would be the case if an action in $\mathscr{A}$(A) were performed, or that are part of the total future state of the world that would be the case if an action in $\mathscr{A}$(A) were performed, or that are causally affected by an action in $\mathscr{A}$(A).

11.3 Adequacy Conditions for the Fifth Theorem

The Condition of Separate Satisfiability: For any agent and any situation, it is logically possible for her not to act in a morally wrong way.

The Condition of Separate Satisfiability (exact formulation): In any choice situation $\mathscr{C}$, at least one action is not wrong.

The Normative Egalitarian Dominance Condition: If population A is a perfectly equal population of the same size as population B, and every person in A has higher welfare than every person in B, then, in any situation involving a choice between A and B, it is wrong to choose B, other things being equal.

The Normative Egalitarian Dominance Condition (exact formulation): For any choice situation $\mathscr{C}$, and any welfare level $\mathbb{W}_x$, **if** A$\subset\mathbb{W}_x$, all members of B have welfare below $\mathbb{W}_x$, N(A)=N(B), $\mathscr{A}$(A)$\subset\mathscr{C}$, and $\mathscr{A}$(B)$\subset\mathscr{C}$, **then** all actions in $\mathscr{A}$(B) are wrong, other things being equal.

The Normative Quality Condition: There is at least one perfectly equal population with very high positive welfare such that if it is wrong in a certain situation to choose that population, then it is wrong in the same situation to choose any population with very low positive welfare, other things being equal.

The Normative Quality Condition (exact formulation): There are two positive welfare ranges $\mathbb{R}$(u, v) and $\mathbb{R}$(1, y), $u > y$, and a population size $n > 0$, such that, for any choice situation $\mathscr{C}$, **if** $\mathbb{W}_z\subset\mathbb{R}$(u, v), A$\subset\mathbb{W}_z$, N(A)=n,

B⊂$\mathbb{R}$(1, y), $\mathcal{A}$(A)⊂$\mathcal{C}$, $\mathcal{A}$(B)⊂$\mathcal{C}$, and all actions in $\mathcal{A}$(A) are wrong, **then** all actions in $\mathcal{A}$(B) are wrong, other things being equal.

The Normative Dominance Addition Condition: If it is wrong in a certain situation to add lives with positive welfare and increase the welfare of the rest of the population, then it is also wrong in the same situation to add no lives, other things being equal.

The Normative Dominance Addition Condition (exact formulation): For any choice situation $\mathcal{C}$, and any pair of welfare levels $\mathbb{W}_x$ and $\mathbb{W}_y$, $y > 0$, **if** all the lives in A have welfare below $\mathbb{W}_x$, all the lives in B have welfare above or at $\mathbb{W}_x$, N(A)=N(B), C⊂$\mathbb{W}_y$, $\mathcal{A}$(A)⊂$\mathcal{C}$, $\mathcal{A}$(B∪C)⊂$\mathcal{C}$, and all actions in $\mathcal{A}$(B∪C) are wrong, **then** all actions in $\mathcal{A}$(A) are wrong, other things being equal.

The Normative Inequality Aversion Condition: For any triplet of welfare levels $\mathbb{A}$, $\mathbb{B}$, and $\mathbb{C}$, $\mathbb{A}$ higher than $\mathbb{B}$ and $\mathbb{B}$ higher than $\mathbb{C}$, and for any population A with welfare $\mathbb{A}$, there is a larger population C with welfare $\mathbb{C}$ such that if it is wrong in a certain situation to choose a perfectly equal population B of the same size as A∪C and with welfare $\mathbb{B}$, then it is also wrong in the same situation to choose A∪C, other things being equal.

The Normative Inequality Aversion Condition (exact formulation): For any triplet $\mathbb{W}_x$, $\mathbb{W}_y$, $\mathbb{W}_z$ of welfare levels, $x > y > z$, and any number of lives $n > 0$, there is a number of lives $m > n$ such that, for any choice situation $\mathcal{C}$, **if** A⊂$\mathbb{W}_x$, N(A)=n, B⊂$\mathbb{W}_y$, N(B)=m+n, C⊂$\mathbb{W}_z$, N(C)=m, $\mathcal{A}$(A∪C)⊂$\mathcal{C}$, $\mathcal{A}$(B)⊂$\mathcal{C}$, and all actions in $\mathcal{A}$(B) are wrong, **then** all actions in $\mathcal{A}$(A∪C) are wrong, other things being equal.

11.4 The Fifth Impossibility Theorem

The Fifth Impossibility Theorem: There is no separately satisfiable morality which satisfies the Normative Quality Condition, the Normative Inequality Aversion Condition, the Normative Egalitarian Dominance Condition, and the Normative Dominance Addition Condition.

Proof: We show that the contrary assumption leads to a contradiction. Let

(1) $\mathbb{R}$(u, v) and $\mathbb{R}$(1, y), $u > y$, be two welfare ranges, and n a number, which satisfy the Normative Quality Condition;
(2) $m > n$ be a number of lives which satisfies the Normative Inequality Aversion Condition for n, $\mathbb{W}_v$, $\mathbb{W}_2$, and $\mathbb{W}_1$;
(3) A$\subset\mathbb{W}_u$, N(A)=n;
(4) B$\subset\mathbb{W}_y$, N(B)=m+n;
(5) C$\subset\mathbb{W}_2$, N(C)=N(B);
(6) D$\subset\mathbb{W}_1$, N(D)=m;
(7) E$\subset\mathbb{W}_v$, N(E)=n.

Finally, let

(8) $\mathcal{A}$(A)$\cup\mathcal{A}$(E$\cup$D)$\cup\mathcal{A}$(C)$\cup\mathcal{A}$(B) be all of the available actions in $\mathcal{C}$.

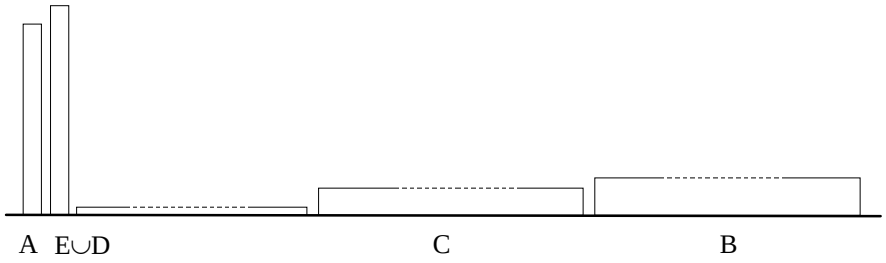

Diagram 11.4.1

It follows from the definition of a welfare range that $y > 2$. Consequently, (4), (5), and the Normative Egalitarian Dominance Condition yield that

(9) all actions in $\mathcal{A}$(C) are wrong (see Diagram 11.4.1).

The Normative Inequality Aversion Condition, (2),(5)-(7), and (9) imply that

(10) all actions in $\mathcal{A}$(E$\cup$D) are wrong (see Diagram 11.4.1).

By the definition of a welfare range, we have that $v > u$. Consequently, it follows from (3), (6)-(7), (10), and the Normative Dominance Addition Condition that

(11) all actions in $\mathscr{A}$(A) are wrong (see Diagram 11.4.1).

From (1), (3)-(4), (11) and the Normative Quality Condition we get that

(12) all actions in $\mathscr{A}$(B) are wrong (see Diagram 11.4.1).

It follows from the Condition of Separate Satisfiability that

(13) not all actions in $\mathscr{C}$ are wrong,

and from (9)-(12) that

(14) all actions in $\mathscr{C}$ are wrong,

which contradicts (13). Hence, the assumption that there is a morality which satisfies all the adequacy conditions entails a contradiction. Thus, the impossibility theorem must be true. Q.E.D.

11.5 Adequacy Conditions for the Sixth Theorem

> *The Normative Non-Sadism Condition*: If it is wrong in a certain situation to add any number of lives with positive welfare, then it is also wrong in the same situation to add any number of lives with negative welfare, other things being equal.
>
> *The Normative Non-Sadism Condition (exact formulation)*: For any choice situation $\mathscr{C}$, and for any population C, **if** A$\subset\mathbb{W}_x$, $x > 0$, B$\subset\mathbb{W}_y$, $y < 0$, N(B) > 0, $\mathscr{A}$(A$\cup$C)$\subset\mathscr{C}$, $\mathscr{A}$(B$\cup$C)$\subset\mathscr{C}$, and all actions in $\mathscr{A}$(A$\cup$C) are wrong, **then** all actions in $\mathscr{A}$(B$\cup$C) are wrong, other things being equal.
>
> *The Normative Weak Quality Addition Condition*: For any population X, there is at least one perfectly equal population with very high welfare

such that if it is wrong in a certain situation to add this population to X, then it is also wrong in the same situation to add any population with very low positive welfare to X, other things being equal.

The Normative Weak Quality Addition Condition (exact formulation): For any population C, there are two positive welfare ranges $\mathbb{R}(x, w)$ and $\mathbb{R}(1, y)$, $x > y$, and a population size n such that, for any choice situation $\mathscr{C}$, **if** $A \subset \mathbb{W}_z$, $z \geq x$, $N(A)=n$, $B \subset \mathbb{R}(1, y)$, $\mathscr{A}(A \cup C) \subset \mathscr{C}$, $\mathscr{A}(B \cup C) \subset \mathscr{C}$, and all actions in $\mathscr{A}(A \cup C)$ are wrong, **then** all actions in $\mathscr{A}(B \cup C)$ are wrong, other things being equal.

The Normative Non-Extreme Priority Condition: There is a number n of lives such that for any population X, if it is wrong in a certain situation to choose a population consisting of the X-lives, n lives with very high welfare, and a single life with slightly negative welfare, then it is also wrong in the same situation to choose a population consisting of the X-lives and $n+1$ lives with very low positive welfare, other things being equal.

The Normative Non-Extreme Priority Condition (exact formulation): There are two welfare levels $\mathbb{W}_x$ and $\mathbb{W}_y$, and a welfare range $\mathbb{R}(1, z)$, $x > z$, $y < 0$, and a number of lives n such that, for any choice situation $\mathscr{C}$, **if** $A \subset \mathbb{W}_u$, $u \geq x$, $N(A)=n$, $B \subset \mathbb{R}(1, z)$, $N(B)=n+1$, $C \subset \mathbb{W}_y$, $N(C)=1$, $D \subset \mathscr{L}$, $\mathscr{A}(A \cup C \cup D) \subset \mathscr{C}$, $\mathscr{A}(B \cup D) \subset \mathscr{C}$, and all actions in $\mathscr{A}(A \cup C \cup D)$ are wrong, **then** all actions in $\mathscr{A}(B \cup D)$ are wrong, other things being equal.

The other two conditions which we shall employ in the sixth theorem are the Normative Egalitarian Dominance and the Normative Inequality Aversion Condition.

11.6 The Sixth Impossibility Theorem

The Sixth Impossibility Theorem: There is no separately satisfiable morality which satisfies the Normative Egalitarian Dominance, the Normative Inequality Aversion, the Normative Non-Extreme Priority,

the Normative Non-Sadism, and the Normative Weak Quality Addition Condition.

Proof. We show that the contrary assumption leads to a contradiction. Let

(1) $\mathbb{W}_x$ and $\mathbb{W}_y$ be two welfare levels, $\mathbb{R}(1, z)$ a welfare range, $x > z$, $y < 0$, and q a number of lives, which satisfy the Normative Non-Extreme Priority Condition;
(2) B$\subset\mathbb{W}_3$, N(B)=q+1;
(3) $\mathbb{R}$(u, t) and $\mathbb{R}$(1, v), $u > v$, be two welfare ranges, and p a population size, which satisfy the Normative Weak Quality Addition Condition for population B;
(4) $\mathbb{W}_w$ be a welfare level such that $w \geq x$ and $w \geq u$;
(5) $r > p+q$ be a number of lives which satisfies the Normative Inequality Aversion Condition for the three welfare levels $\mathbb{W}_w$, $\mathbb{W}_2$, and $\mathbb{W}_1$ and $p+q$ lives at $\mathbb{W}_w$;
(6) C$\subset\mathbb{W}_3$, N(C)=$p+r$-1;
(7) G$\subset\mathbb{W}_2$, N(G)=$p+q+r$;

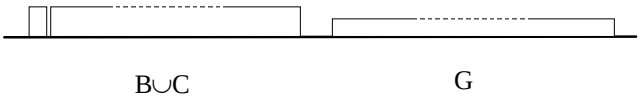

Diagram 11.6.2

Since B∪C$\subset\mathbb{W}_3$ (by (2) and (6)) and G$\subset\mathbb{W}_2$, (by (7)) and N(B∪C)=q+1+p+r-1=q+p+r=N(G), the Normative Egalitarian Dominance Condition implies that

(8) all actions in $\mathbb{A}$(G) are wrong (see Diagram 11.6.2).

Let

(9) A$\subset\mathbb{W}_w$, N(A)=p;
(10) E$\subset\mathbb{W}_y$, N(E)=1;

(11) H⊂$\mathbb{W}_w$, N(H)=q;
(12) F⊂$\mathbb{W}_1$, N(F)=r.

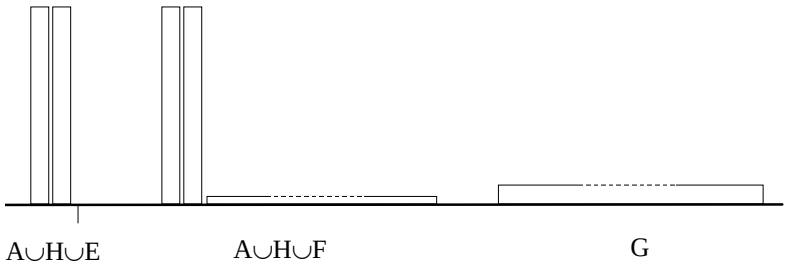

Diagram 11.6.3

Since A∪H⊂W_w and N(A∪H)=p+q (by (9) and (11)), it follows from (5), (7)-(8), and the Normative Inequality Aversion Condition that

(13) all actions in $\mathscr{A}$(A∪H∪F) are wrong (see Diagram 11.6.3).

Since the E-life has negative welfare (by (1) and (10)), and the F-lives have positive welfare (12), it follows from (13) and the Normative Non-Sadism Condition that

(14) all actions in $\mathscr{A}$(A∪H∪E) are wrong. (see Diagram 11.6.3).

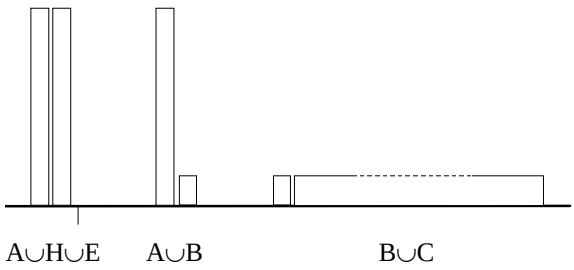

Diagram 11.6.4

It follows from the definition of a welfare range that $\mathbb{W}_3$⊂$\mathbb{R}$(1, z). Accordingly, from (2) we know that B⊂$\mathbb{R}$(1, z). Consequently, from (1), (4), (10)-(11), (14) and the Normative Non-Extreme Priority Condition we get that

(15) all actions in $\mathscr{A}$(A∪B) are wrong (see Diagram 11.6.4).

Since $\mathbb{W}_3 \subset \mathbb{R}(1, v)$, we can conclude that $C \subset \mathbb{R}(1, v)$, and since $w \geq u$ (by (4)), and $A \subset \mathbb{W}_w$, and $N(A)=p$ (by (9)), it follows from (3), (15), and the Normative Weak Quality Addition Condition that

(16) all actions in $\mathscr{A}(B \cup C)$ are wrong (see Diagram 11.6.4).

Let

(17) $\mathscr{A}(A \cup B) \cup \mathscr{A}(A \cup H \cup E) \cup \mathscr{A}(A \cup H \cup F) \cup \mathscr{A}(B \cup C) \cup \mathscr{A}(G)$ be all of the available actions in $\mathscr{C}$.

It follows from the Condition of Separate Satisfiability that

(18) not all actions in $\mathscr{C}$ are wrong,

and from (8), (13)-(17) that

(19) all actions in $\mathscr{C}$ are wrong,

which contradicts (18). Q.E.D.

12

Summary

We began this essay by considering a case in which our evaluations of three possible future scenarios seemed to be inconsistent. As we said, if the evaluations involved in such cases stand up to scrutiny, then our considered moral beliefs are mutually inconsistent. Since consistency is, arguably, a necessary condition for moral justification, we would thus seem to be forced to conclude that there is no moral view which can be justified. In other words, cases involving future generations constitute a serious challenge to the existence of a satisfactory moral theory. Hence the title of this essay.

As the paradox in chapter 1 was presented, however, it was hopelessly vague and didn't force us to draw any such conclusions. In chapter 2, we tried to clarify the concepts involved in the paradox and proceeded in chapters 3-9 with a discussion of the suggested population axiologies in the literature and putative solutions to the paradox. We showed that none of the population axiologies in the literature stood up to scrutiny. In our discussion, we proposed a number of adequacy conditions for an acceptable population axiology. In chapter 10, we proved that there is no population axiology that can satisfy these conditions.

The axiological theorems in chapter 10 presuppose that the relation "is at least as good as" is transitive. Some theorists find this a matter of logic, claiming that it is part of the meaning of "better than" and "equally as good as". Although we are inclined to agree, one might think otherwise, and argue that the impossibility theorems actually demonstrate that these relations are not transitive. As we suggested in chapter 11, there is no uncontroversial analogue to transitivity for normative concepts, and if one takes "A is better than B" as synonymous with, for example, "A ought to be chosen in a situation where A and B are the alternatives", then "better than" won't be transitive. The main question of chapter 11 was whether, as some theorists have suggested, the paradoxes in population theory could be resolved by shifting the discussion to the normative level. We showed that

this isn't the case by proving normative versions of two of the theorems from chapter 10.

In our discussion we have assumed that welfare is at least sometimes interpersonally comparable. Without this assumption, claims such as "Rysiek is better off than Erik" wouldn't be meaningful. In other words, conditions such as the Egalitarian Dominance Condition and the Inequality Aversion Condition, in their normative or axiological guise, wouldn't make sense. The adequacy conditions and the theorems are quite undemanding, however, in regard to measurement of welfare. It doesn't matter whether welfare is measurable on an ordinal, interval or ratio scale, for example. The conditions and theorems only presuppose that lives are quasi-ordered by the relation "has at least as high welfare as".

It is interesting to compare the information demands of the present theorems with that of Arrow's famous impossibility theorem.[1] It has been shown that Arrow's theorem holds true both for measurement on the ordinal and interval scale as long as there is no interpersonal comparability of welfare.[2] Not surprisingly then, the standard remedy for Arrowian impossibility results is to introduce some kind of interpersonal comparability of welfare.[3] But with interpersonal comparability of welfare, and some minimal demands on the orderings of lives, we come up against the impossibility theorems presented in this essay.

Where does this leave us? Roughly, I think we have three options. We can (i) bite the bullet and abandon some of the adequacy conditions, (ii) we can become moral sceptics and accept that our considered moral beliefs are not epistemically justified, or (iii) we can try to find a way to explain away the relevance of the results in this essay for moral justification. Of these three options, I cannot come to terms with (i) and (ii). We haven't discussed the third option much in this essay (see, however, the discussion of relevant test cases in section 2.3). In light of the results presented here, however, I think it deserves further attention. Derek Parfit, for instance, has suggested that we might be able to "quarantine" the impossibility theorems and the resulting scepticism to cases involving different numbers of people only.[4] He compares cases involving different sized populations with cases

---

[1] See Arrow (1963). Notice that Arrow's result appears already in a fixed population size setting.

[2] See Sen (1970), pp. 123-5, 128-30, and Roemer (1996), pp. 26-36. It would be surprising if Sen's and Roemer's theorems cannot be extended to cover non-interpersonally comparable measurement on any scale at least as strong as the ordinal scale.

[3] Roemer (1996), p. 36, among many others, suggests this.

[4] Personal communication.

involving "infinite quantities of welfare": Although it is very difficult to formulate a welfarist theory that can handle such cases in an acceptable way, this problem doesn't undermine our confidence in the theories that can handle cases that only involve finite quantities of welfare. One could argue similarly that the results in the present work shouldn't undermine our beliefs in moralities that can handle same number cases in a satisfactory way.

Parfit's suggestion raises interesting but unresolved epistemological questions. For example, how could one justify "quarantining" certain areas of inquiry? Parfit's analogy to cases that involve infinite quantities of welfare is not convincing since, as we argued in chapter 2, we have, in regard to such cases, pretty good epistemological reasons for "quarantining": Intuitions about cases that involve infinities are notoriously unreliable, and the very concept of "infinity" is complicated and hard to understand. The idea of finite but different sized populations is not hard to understand, and although some of our intuitions about different number cases might be unreliable, I don't think this is true of the ones we have made use off in the theorems. At least, we must find a reason why we should be sceptical about them. At any rate, "quarantining" would be just half of a solution since a vast number of decisions affect the number of people in the future.

There could be other ways of explaining away the relevance of the impossibility theorems for moral justification and although I regret to report that I have no proposal of my own, a closer investigation of option (iii) might bring to light an as yet unforeseen solution to the problems discussed in this essay. Perhaps we should not take the paradoxes of future generations as a challenge to the existence of a satisfactory moral theory, but as a challenge to some of our beliefs about moral justification and about the purpose and scope of moral theory.

Appendix A

List of Conditions, Principles and Conclusions

Here's a list of conditions, principles, and conclusions that we refer to in several chapters. Exact formulations of the conditions can be found in chapter 10.

The Egalitarian Dominance Condition: If population A is a perfectly equal population of the same size as population B, and every person in A has higher welfare than every person in B, then A is better than B, other things being equal.

The General Non-Extreme Priority Condition: There is a number *n* of lives such that for any population X, and any welfare level $\mathbb{A}$, a population consisting of the X-lives, *n* lives with very high welfare, and one life with welfare $\mathbb{A}$, is at least as good as a population consisting of the X-lives, *n* lives with very low positive welfare, and one life with welfare slightly above $\mathbb{A}$, other things being equal.

The Inequality Aversion Condition: For any triplet of welfare levels $\mathbb{A}$, $\mathbb{B}$, and $\mathbb{C}$, $\mathbb{A}$ higher than $\mathbb{B}$, and $\mathbb{B}$ higher than $\mathbb{C}$, and for any population A with welfare $\mathbb{A}$, there is a larger population C with welfare $\mathbb{C}$ such that a perfectly equal population B of the same size as A∪C and with welfare $\mathbb{B}$ is at least as good as A∪C, other things being equal.

The Mere Addition Principle: An addition of people with positive welfare does not make a population worse, other things being equal.

The Negative Mere Addition Principle: An addition of people with negative welfare makes a population worse, other things being equal.

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

The Non-Elitism Condition: For any triplet of welfare levels 𝔸, 𝔹, and ℂ, 𝔸 slightly higher than 𝔹, and 𝔹 higher than ℂ, and for any one-life population A with welfare 𝔸, there is a population C with welfare ℂ, and a population B of the same size as A∪C and with welfare 𝔹, such that for any population X consisting of lives with welfare ranging from ℂ to 𝔸, B∪X is at least as good as A∪C∪X, other things being equal.

The Non-Extreme Priority Condition: There is a number *n* of lives such that for any population X, a population consisting of the X-lives, *n* lives with very high welfare, and a single life with slightly negative welfare is at least as good as a population consisting of the X-lives and *n*+1 lives with very low positive welfare, other things being equal.

The Non-Sadism Condition: An addition of any number of people with positive welfare is at least as good as an addition of any number of people with negative welfare, other things being equal.

The Quality Addition Principle: There is at least one perfectly equal population with very high welfare such that its addition to any population X is at least as good as an addition of any population with very low positive welfare to X, other things being equal.

The Quality Condition: There is at least one perfectly equal population with very high welfare which is at least as good as any population with very low positive welfare, other things being equal.

The Quantity Condition: For any pair of positive welfare levels 𝔸 and 𝔹, such that 𝔹 is slightly lower than 𝔸, and for any number of lives *n*, there is a greater number of lives *m*, such that a population of *m* people at level 𝔹 is at least as good as a population of *n* people at level 𝔸, other things being equal.

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.

The Reversed Repugnant Conclusion: For any population with very high positive welfare, there is a better population consisting of just one person with slightly higher welfare, other things being equal.

The Weak Non-Sadism Condition: There is a negative welfare level and a number of lives at this level such that an addition of any number of people with positive welfare is at least as good as an addition of the lives with negative welfare, other things being equal.

The Weak Quality Addition Condition: For any population X, there is at least one perfectly equal population with very high welfare such that its addition to X is at least as good as an addition of any population with very low positive welfare to X, other things being equal.

Appendix B

The Logical Relation Between the Repugnant Conclusion and the Quality Condition

We shall show that given full comparability, a theory which avoids the Repugnant Conclusion and satisfies the Egalitarian Dominance Condition implies the Quality Condition. The negation of the Repugnant Conclusion can be stated as follows:

> *NON-RC*: There is at least one population of at least ten billion people with very high welfare such that no larger population with very low positive welfare is better, other things being equal.

Given full comparability, this is equivalent to:

> *NON-RC2*: There is at least one population of at least ten billion people with very high welfare which is at least as good as any larger population with very low positive welfare, other things being equal.

Let A be a population with very high welfare which satisfies NON-RC2. Let B be a perfectly equal population of the same size as A such that every person in B has higher welfare than every person in A. According to the Egalitarian Dominance Condition, B is better than A. Consequently, B is better than any larger population with very low welfare. It also follows from the Egalitarian Dominance Condition that B is better than any same sized population with very low welfare.

We now have to consider two cases: (a) B is at least as good as any smaller population with very low welfare, or (b) B is worse than some smaller population with very low welfare.

If (a) is true, then the Quality Condition is satisfied, since B is then at least as good as all populations with very low welfare. If (b) is true, then let $\mathcal{P}_1$ be the set of

all populations with very low welfare which are better than B. Now, for any member of $\mathcal{P}_1$ there is a same sized perfectly equal population with very high welfare which is better according to the Egalitarian Dominance Condition. Let $\mathcal{P}_2$ be a set of perfectly equal populations with very high welfare such that for any member X of $\mathcal{P}_1$, there is one, but only one, member Y of $\mathcal{P}_2$ such that Y is better than X according to the Egalitarian Dominance Condition. It follows that any member of $\mathcal{P}_2$ is better than all same sized populations in $\mathcal{P}_1$. Since B is of finite size, and all members of $\mathcal{P}_1$ are smaller than B, the set consisting of all equivalence classes on $\mathcal{P}_1$ in respect to the relation "is of the same size as" is finite. Since any member of $\mathcal{P}_2$ is better than all same sized populations in $\mathcal{P}_1$, and there is only one member of $\mathcal{P}_2$ which is better than any given member of $\mathcal{P}_1$ according to the Egalitarian Dominance Condition, $\mathcal{P}_2$ is of finite size. Given full comparability and that $\mathcal{P}_2$ is of finite size, there is a population in $\mathcal{P}_2$ which is at least as good as all other members of $\mathcal{P}_2$. Let C be a member of $\mathcal{P}_2$ which is at least as good as all other members of $\mathcal{P}_2$. It follows that C is better than any member of $\mathcal{P}_1$ and, consequently, that C is better than B. Since C is better than B, C is better than all populations with very low welfare which are larger than B or of the same size as B. Moreover, C is better than all populations with very low welfare which are smaller than B, since C is better than all members of $\mathcal{P}_1$ and, of course, better than all smaller populations with very low welfare which B is better than or equally as good as. Consequently, C is a perfectly equal population with very high welfare which is better than all populations with very low welfare. Thus, the Quality Condition is satisfied. In other words, given full comparability among populations, avoidance of the Repugnant Conclusion and satisfaction of the Egalitarian Dominance Condition together entail the Quality Condition. Q.E.D.

Bibliography

(Square brackets indicate year of reprint)

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

Arneson, R. "Equality of Opportunity for Welfare", *Philosophical Studies* 56, pp. 77-93, 1989.

Arrhenius, G. "Population Ethics and Variable Value Theories", mimeo, University of Toronto, 1995.

——— "An Impossibility Theorem for Welfarist Axiologies", *Economics and Philosophy*, forthcoming, October 2000a.

——— "Mutual Advantage Contractarianism and Future Generations", *Theoria*, forthcoming, 2000b.

Arrhenius, G., and Bykvist, K. *Interpersonal Compensations and Moral Duties to Future Generations: Moral Aspects of Energy Use*, Uppsala Prints and Preprints in Philosophy, No. 21, Uppsala: Department of Philosophy, University of Uppsala, 1995.

Arrow, K. J. *Social Choice and Individual Values*, 2nd ed., New Haven and London: Yale University Press, 1963.

Barry, B. "Utilitarianism and Preference Change", *Utilitas*, Vol 1., pp. 278-282, 1989.

Bigelow, J., and Pargetter, R. "Morality, Potential Persons and Abortion", *American Philosophical Quarterly*, Vol. 25, No. 2, pp. 173-81, April, 1988.

Blackorby, C., and Donaldson, D. "Social Criteria for Evaluating Population Change", *Journal of Public Economics*, 25, pp. 13-33, 1984.

——— "Normative Population Theory: A Comment", *Social Choice and Welfare* 8, pp. 261-7, 1991.

——— "Pigs and Guinea Pigs: A Note on the Ethics of Animal Exploitation", *The Economic Journal*, 102, pp. 1345-69, November, 1992.

Blackorby, C., Bossert, W., and Donaldson, D. "Intertemporal Population Ethics: Critical-Level Utilitarian Principles", *Econometrica* 65, pp. 1303-1320, 1995.

——— "Critical-Level Utilitarianism and the Population-Ethics Dilemma", *Economics and Philosophy*, 13, pp. 197-230, 1997.

Boonin-Vail, D. "Don't Stop Thinking About Tomorrow: Two Paradoxes About Duties to Future Generations", *Philosophy and Public Affairs*, Vol. 25, No. 4, pp. 267-307, Fall 1996.

Braybrooke, D. *Meeting Needs*, Princeton: Princeton University Press, 1987.

Broad, C. D. *Five Types of Ethical Theory*, London: Routledge and Kegan Paul, [1979] 1930.

Broome, J. *Weighing Goods: Equality, Uncertainty, and Time*, Oxford: Basil Blackwell, 1991.

——— *Ethics out of Economics*, Cambridge: Cambridge University Press, 1999.

Bykvist, K. "Utilitarian Deontologies?", pp. 1-16 in W. Rabinowicz (ed.), *Preference and Value: Preferentialism in Ethics*, Studies in Philosophy, Dept. of Philosophy, Lund University, Vol. 1, 1996.

——— *Changing Preferences: A Study in Preferentialism*, F.D. Diss., Uppsala University, 1998.

Carlson, E. *Consequentialism Reconsidered*, Dordrecht: Kluwer Academic Publisher, 1995.

——— "Cyclical Preferences and Rational Choice", *Theoria* 62, pp. 144-160, 1996.

——— "Consequentialism, Distribution and Desert", *Utilitas*, Vol. 9, No. 3, pp. 307-18, November 1997.

——— "Mere Addition and Two Trilemmas of Population Ethics", *Economics and Philosophy* 14, pp. 283-306, 1998a.

——— "Aggregating Harms – Should We Kill to Avoid Headaches?" in *Two Short Papers on Harm and Value Aggregation*, Uppsala Prints and Preprints in Philosophy, No. 4, Uppsala: Department of Philosophy, University of Uppsala, 1998b.

Cowen, T. "What Do We Learn from the Repugnant Conclusion?", *Ethics* 106, pp. 754-75, July 1996.

Crisp, R. *Ideal Utilitarianism: Theory and Practice*, D.Phil. Diss., University of Oxford, 1988.

——— "Utilitarianism and the Life of Virtue", *The Philosophical Quarterly*, Vol. 42, No. 167, pp. 139-60, April 1992.

Danielsson, S. "Konsekvensetikens gränser" in *Filosofiska Utredningar*, Stockholm: Thales, 1988.

——— "The Refutation of Cyclic Evaluations", *Theoria* 62, pp. 161-168, 1996.

——— "The Norm-Value Distinction in 1963" pp. 325-330 in G. Meggle (ed.) *Actions, Norms, Values: Discussions with Georg Henrik von Wright*, Berlin: de Gruyter, 1999.

Dasgupta, P. "Lives and Well-Being", *Social Choice and Welfare* 5, pp. 103-126, 1988.

Donagan, A. *The Theory of Morality*, Chicago: University of Chicago Press, 1977.

Fehige, C. "A Pareto Principle for Possible People", mimeo, Univ. of Pittsburgh, Center for Philosophy of Science, 1992.

——— "A Pareto Principle for Possible People", pp. 509-43 in C. Fehige and U. Wessels (eds.), *Preferences*, Berlin: de Gruyter, 1998.

Feldman, F. "Adjusting Utility for Justice: a Consequentialist Reply to the Objection from Justice", *Philosophy and Phenomenological Research* LV, 3, pp. 567-85, 1995a. Reprinted in Feldman (1997).

——— "Justice, Desert, and the Repugnant Conclusion", *Utilitas*, Vol. 7, No. 2, pp. 189-206, November 1995b. Reprinted in Feldman (1997).

——— *Utilitarianism, Hedonism, and Desert: Essays in Moral Philosophy,* Cambridge: Cambridge University Press, 1997.

Finnis, J. *Natural Law and Natural Rights*, Clarendon Law Series, Oxford: Oxford University Press, 1980.

Gauthier, D. *Morals By Agreement*, Oxford: Clarendon Press, 1986.

Glanz, J. "Cosmic Motion Revealed", *Science*, Vol. 282, No. 5397, 18 December, pp. 2156-7, 1998.

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

Goodin, R. E. "Actual Preferences, Actual People", *Utilitas*, Vol. 3, No. 1, pp. 113-9, May, 1991.

Griffin, J. *Well-Being: Its Meaning, Measurement, and Moral Importance*, Oxford: Clarendon Press, [1990] 1986.

Hansson, S. O. *Structures of Value: An Investigation of the Statics and Dynamics of Values and Norms*, Ph.D. Diss., Lund: Lund Philosophy Reports, Vol. 1, 1998.

Hare, R. M. *Moral Thinking: Its Levels, Method, and Point*, Oxford: Clarendon Press, 1981.

——— "Possible People" in *Essays on Bioethics*, Oxford: Clarendon Press, 1993.

Harsanyi, J. C. "Utilities, Preferences and Substantive Goods", Working Papers No. 101, WIDER, The United Nations University, December 1992.

Heath, J. "Intergenerational Cooperation and Distributive Justice", *Canadian Journal of Philosophy*, Vol. 27, No. 3, pp. 361-76, September 1997.

Heyd, D. "Procreation and Value: Can Ethics Deal With Futurity Problems?", *Philosophia* (Israel), No. 18, pp. 151-170, July, 1988.

Hobbes, T. *Leviathan*, M. Oakeshott (ed.), Collier Books, New York, [1973] 1962.

Holtug, N. "In Defence of the Slogan", pp. 64-89 in W. Rabinowicz (ed.), *Preference and Value: Preferentialism in Ethics*, Studies in Philosophy, Dept. of Philosophy, Lund University, Vol. 1, 1996.

Hudson, J. L. "The Diminishing Marginal Value of Happy People", *Philosophical Studies* 51, pp. 123-37, 1987.

Hurka, T. "Value and Population Size", *Ethics* 93, pp. 496-507, 1983.

——— *Perfectionism*, New York: Oxford University Press, 1993.

Kavka, G. S. "The Paradox of Future Individuals", *Philosophy and Public Affairs*, Vol. 11, No. 2, pp. 93-112, Spring 1982.

Klint Jensen, K. *Om afvejning af værdier*, Ph.D. Diss., University of Copenhagen, 1996.

Kymlicka, W. *Contemporary Political Philosophy*, Oxford: Clarendon Press, 1990.

McTaggart, J. M. E. *The Nature of Existence*, Cambridge: Cambridge University Press, 1927.

Mitchell, E. T. *A System of Ethics*, New York: Charles Scribner's Sons, 1950.

Mongin, P. "Spurious Unanimity and the Pareto Principle", mimeo, Université de Cergy-Pontoise, 1997.

Moore, G. E. *Principia Ethica*, Cambridge: Cambridge University Press, [1966] 1903.

Munthe, C. *Livets slut i livets början: en studie i abortetik*, Stockholm: Thales, 1992.

Narveson, J. "Utilitarianism and New Generations", *Mind* 76, pp. 62-72, Jan 1967.

——— "Moral Problems of Population", *The Monist* 57, pp. 62-86, Jan 1973.

——— "Future People and Us", pp. 38-60 in R. I. Sikora and B. Barry (eds.), *Obligations to Future Generations*, Philadelphia: Temple University Press, 1978.

Ng, Y-K. "What Should We Do About Future Generations? Impossibility of Parfit's Theory X", *Economics and Philosophy* 5 (2), pp. 235-253, 1989.

Norcross, A. "Comparing Harms: Headaches and Human Lives", *Philosophy and Public Affairs*, Vol. 26, No. 2, pp. 135-167, 1997.

——— "Great Harms from Small Benefits Grow: How Death Can Be Outweighed by Headaches", *Analysis* 58 (2), pp. 152-158, April 1998.

Odelstad, J. *Mätning och beslut: Sju uppsatser om meningsfullhet, amalgamering och begreppet funktion*, Philosophical Studies published by the Philosophical Society and the Department of Philosophy, No. 43, Uppsala: University of Uppsala, 1990.

Österberg, J. *Self and Others*, Dordrecht: Kluwer Academic Publisher, 1988.

——— "Utilitarism och möjliga varelser", mimeo, Uppsala Universitet, 1992.

——— "Value and Existence: the Problem of Future Generations", pp. 94-107 in S. Lindström, R. Sliwinski, and J. Österberg (eds.), *Odds and Ends*, Uppsala Philosophical Studies, Uppsala: Department of Philosopy, Uppsala University, 1996.

Parfit, D. *Reasons and Persons*, Oxford: Clarendon Press, [1991] 1984.

——— "Overpopulation and the Quality of Life", pp. 145-64 in P. Singer (ed.), *Applied Ethics*, Oxford University Press, 1986.

——— "Does Equality Matter?" in J. R. Richards (ed.), *Philosophical Problems of Equality*, Milton Keynes: The Open University, 1993.

——— "Acts and Outcomes: A Response to Boonin-Vail", *Philosophy and Public Affairs*, Vol. 25, No. 4, pp. 308-317, Fall 1996.

Persson, I. "Ambiguities in Feldman's Desert-adjusted Values", *Utilitas*, Vol. 9, No. 3., pp. 319-28, November 1997.

Quinn, W. "The Puzzle of the Self-Torturer", *Philosophical Studies* 59, pp. 79-90, 1990.

Rabinowicz, W., and Österberg, J. "Value Based on Preferences: On Two Interpretations of Preference Utilitarianism", *Economics and Philosophy* 12, pp. 1-27, 1996.

Rabinowicz, W. "Money Pump with Foresight" pp. 201-234 in W. Rabinowicz (ed.) *Value and Choice: Some Common Themes in Decision Theory and Moral Philosophy*, Lund: Lund Philosophy Reports, Vol. 1, 2000.

Rachels, S. "Counterexamples to the Transitivity of Better Than", *Australasian Journal of Philosophy*, Vol. 76, No. 1, pp. 71-83, March 1998.

Rawls, J. *A Theory of Justice*, Cambridge, Mass.: Harvard University Press, 1971.

——— "The Priority of Right and Ideas of the Good", *Philosophy and Public Affairs*, Vol. 17, No. 4, pp. 251-276, Fall 1988.

Raz, J. *The Morality of Freedom*, Oxford: Clarendon Press, 1986.

Roberts, F.S. *Measurement Theory with Applications to Decisionmaking, Utility, and the Social Sciences*, Encyclopedia of Mathematics and Its Applications, Vol. 7, Cambridge : Cambridge Univ. Press, 1984.

Roemer, J. E. *Theories of Distributive Justice*, Cambridge, Mass.: Harvard University Press, 1996.

Ryberg, J. *Topics on Population Ethics*, Ph.D. Diss., University of Copenhagen, 1996.

Sen, A. *Collective Choice and Social Welfare*, Mathematical Economics Texts 5, 1970.

——— "Equality of What?" in S. McMurrin (ed.), *The Tanner Lecture on Human Values*, Cambridge: Cambridge University Press, 1980.

——— *Inequality Reexamined,* Cambridge, Mass.: Cambridge University Press, 1992.

——— "Capability and Well-Being" in M. Nussbaum and A. Sen (eds.), *The Quality of Life*, Oxford: Clarendon Press, 1993.

——— "Rationality and Social Choice", *American Economic Review*, Vol. 85, No. 1, pp. 1-25, March 1995.

Sider, T. R. "Might Theory X Be a Theory of Diminishing Marginal Value?", *Analysis* 51 (4), pp. 265-271, 1991.

Sidgwick, H. *The Methods of Ethics*, 7th edn., London: Macmillan, [1967] 1907.

Singer, P. *Practical Ethics*, 2nd ed., Cambridge: Cambridge University Press, 1993.

Sumner, L. W. *Welfare, Happiness, and Ethics*, Oxford: Clarendon Press, 1996.

——— "Review of Fred Feldman's Utilitarianism, Hedonism, and Desert: Essays in Moral Philosophy, *Ethics* 109 (1), 176-9, October, 1998.

Temkin, L. S. "Intransitivity and the Mere Addition Paradox", *Philosophy and Public Affairs*, Vol. 16, No. 2, pp. 138-87, Spring 1987.

——— *Inequality*, Oxford: Oxford University Press, 1993a.

——— "Harmful Goods, Harmless Bads", pp. 291- 324 in R. G. Frey and C. W. Morris (eds.), *Value, Welfare and Morality*, Cambridge: Cambridge University Press, 1993b.

——— "Weighing Goods: Some Questions and Comments", *Philosophy and Public Affairs*, Vol. 23, No. 4, pp. 350-80, Fall 1994.

——— "A Continuum Argument for Intransitivity", *Philosophy and Public Affairs*, Vol. 25, No. 3, pp. 175-210, Summer 1996.

Tännsjö, T. *Göra Barn*, Sesam förlag, Borås, 1991.

——— *Hedonistic Utilitarianism*, Edinburgh: Edinburgh University Press, 1998.

Vallentyne, P. "Gimmicky Representations of Moral Theories", Metaphilosphy, Vol. 19, No. 34, 1988.

——— "Taking Justice Too Seriously", *Utilitas*, Vol. 7, No. 2, pp. 207-16, November 1995.

von Wright, G. H. *The Varieties of Goodness*, Bristol: Thoemmes Press [1993], 1963.

Warren, M. "Do Potential People Have Moral Rights?", pp. 14-30 in R. I. Sikora and B. Barry (eds.) *Obligations to Future Generations*, Philadelphia: Temple University Press, 1978.

Wedberg, A. "Om S. S. Stevens' klassifikation av måttskalor: Ett begreppsanalytiskt försök", in A-M. Henschen-Dahlquist et al. (eds.) *Sanning Dikt Tro: Till Ingemar Hedenius*, Stockholm: Albert Bonniers förlag AB, 1968.

Whewell, W. *Lectures on Moral Philosophy*, 1st ed., 1852.