Economics and Philosophy, **16** (2000), 247–266 Copyright © Cambridge University Press

AN IMPOSSIBILITY THEOREM FOR WELFARIST AXIOLOGIES

Gustaf Arrhenius
University of Uppsala

1. INTRODUCTION

A search is under way for a theory that can accommodate our intuitions in population axiology. The object of this search has proved elusive. This is not surprising since, as we shall see, any welfarist axiology that satisfies three reasonable conditions implies at least one of three counter-intuitive conclusions. I shall start by pointing out the failures in three recent attempts to construct an acceptable population axiology. I shall then present an impossibility theorem and conclude with a short discussion of how it might be extended to pluralist axiologies, that is, axiologies that take more values than welfare into account.

I would like to thank Charles Blackorby, Walter Bossert, John Broome, Krister Bykvist, Erik Carlson, Sven Danielsson, David Donaldson, Derek Parfit, Adeze Igboemeka, Wlodzimierz Rabinowicz, Howard Sobel, Wayne Sumner and an anonymous referee for their very detailed and helpful comments on earlier drafts of this paper. The comments from Christoph Fehige, John Gibson, Danny Goldstick, Marc Fleurbaey, Peter Hammond, Brad Hooker, Tom Hurka, Karsten Klint Jensen, Andrew Latus, Jan Odelstad, Arthur Ripstein and Jan Österberg were also very helpful. Earlier versions of this paper were presented at the following conferences: 'Utilitarianism Reconsidered', ISUS, New Orleans, April 1997; 'Utilitarianisme: Analyse et Histoire', Association Charles Gide pour l'Etude de la Pensée Economique and the University of Lille, January 1996; and at the Learned Societies Congress, Canadian Philosophical Association, UCAM, Montreal, June 1995. I would like to thank the participants at these occasions for their stimulating criticism. Financial support through a grant from the Swedish Institute during 1995–96 is gratefully acknowledged.

2. NG'S THEORY X'

Yew-Kwang Ng and Theodore Sider have proposed what Tom Hurka calls 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.

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 the Average Utilitarian Principle ranks populations according to the average welfare Q, and the Total Utilitarian Principle according to the total welfare nQ, theory X' ranks them according to $f(n)Q$. Ng's concave function looks like this:

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

The weighing coefficient k represents how quickly the values of additional people approach zero. The smaller k is, the quicker the values of additional people decline. When n approaches infinity, $f(n)$ 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; that is, theory X' behaves like Average Utilitarianism with large populations and thereby avoids Derek Parfit's Repugnant Conclusion:

> *The Repugnant Conclusion*: For any perfectly equal population with very high positive welfare, there is a population with very low positive welfare which is better.[3]

[1] Hurka (1983) introduced this idea and Ng (1989) and Sider (1991) gave it a precise formulation. Parfit (1984, p. 402), mentions a Variable Value Principle but ignores it since 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.

[3] See Parfit (1984, p. 388). My formulation is more general than Parfit's and he does not

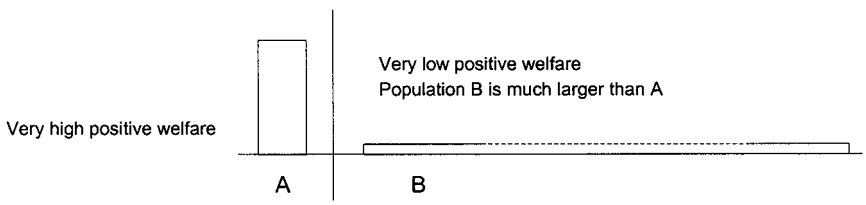

DIAGRAM 1

In Diagram 1, the width of each block represents the number of people, the height represents their lifetime welfare. All the lives in the above diagram have positive welfare, or, as we could also put it, have lives worth living. 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. A hedonist, for example, would typically say that pain is bad and pleasure is good for a person, and that a life without any pain and pleasure has neutral welfare.[4]

People's welfare is much lower in B than in A, since the A-people have very high welfare whereas the B-people have very low positive welfare. The reason for the very low positive welfare in the B-lives could 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 poor quality, for example, working at an assembly line, eating potatoes

demand that the people with very high welfare are equally well off. See Section 5 for a formal definition of this conclusion.

[4] This definition can be combined with other welfarist axiologies, such as desire and objective list theories. Different substantive theories of welfare will probably yield somewhat different answers on exactly where the cut-off point between a life with positive and a life with negative welfare should be drawn. For example, whereas a hedonist would find a life consisting of a few happy days a life worth living, an objective list theorist might find such a life below the threshold of a life with positive welfare. Admittedly, the intuitive force of the Repugnant Conclusion is linked to our understanding of a life with positive (negative) welfare and if we were to radically revise these notions, for example, by claiming that there are no lives that are worth living (see Fehige, 1998), then we would not need to worry about the Repugnant Conclusion. However, as long as a theory of welfare, as a reasonable theory should, roughly respects our commonsense intuitions about the value of life, I do not think that the solution to the problems discussed in this paper essentially turn on exactly where we draw the line between positive and negative welfare and exactly how we spell out our theory of welfare. For a discussion of this issue in connection with the Repugnant Conclusion, see Fehige (1998), Tännsjö (1998) and Arrhenius (1999). Also cf. Parfit (1984, p. 358). A number of alternative definitions of a life with positive (negative, neutral) welfare figures in the literature. For an instructive survey and critical discussion of these, see Broome (1993).

and listening to Muzak.[5] However, since there are many more people in B, the total sum of welfare in B is greater than in A. Hence, Total Utilitarianism ranks B as better than A – an example of the Repugnant Conclusion.[6] Ng's theory X′, on the other hand, ranks A as better than B since with large populations, theory X′ gives average welfare much greater weight than total welfare.

A problem with Ng's principle, however, is that it violates the following plausible condition:

> *The Mere Addition Principle*: For any population, if one adds any number of individuals with positive welfare to create a new population, without affecting the original people's welfare, then this new population is not worse than the original one.[7]

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 person 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 strictly 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 $v < u$ 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 close 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'.[8] Yet, this principle would still not comply with the Mere Addition

[5] See Parfit (1984, p. 388 and 1986, p. 148). Cf. Blackorby, Bossert, and Donaldson (1995, p. 1304), and Ryberg (1996, pp. 154–62). In Arrhenius (1999), I discuss different interpretations of the Repugnant Conclusion in some detail.

[6] For those welfarist axiologies that include animals in the welfare calculus, B could be a population of sheep or some other animal. See, e.g., Singer (1993), Blackorby and Donaldson (1992), Blackorby, Bossert, and Donaldson (1997).

[7] Cf. Hudson (1987), Ng (1989), and Sider (1991). Ng ascribes to Parfit the view that a population principle should satisfy the Mere Addition Principle (Ng, 1989, p. 238) and one might get that impression from Parfit (1984, pp. 420f). 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. Cf. fn 24.

[8] Ng (1989, p. 249). This will also have as a consequence that theory X′ behaves more like Total Utilitarianism even with large populations and yields conclusions similar to the Repugnant Conclusion.

Principle when the population is sufficiently large or when the added people's positive welfare is sufficiently low and implies what I call the Sadistic Conclusion:

> *The Sadistic Conclusion*: When adding people without affecting the original people's welfare, it can be better to add people with negative welfare rather than positive welfare.[9]

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 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.

Ng's principle also has 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*: For any population, if one adds a number of individuals with negative welfare to create a new population, without affecting the original people's welfare, then this new population is worse than the original one.

The 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 person 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 population A if $f(2n)(-u-v)/2 > -uf(n)$. This will be true if $-v > [-f(n)u/f(2n)]+u$. 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

[9] See Section 5 for a formal definition of this conclusion.

n people with welfare $-u$ makes the resulting population *better* than the original one. Consequently, theory X′ violates the very compelling Negative Mere Addition Principle.[10]

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]

Group the individual welfare profiles of a population into two ordered sets:

($u_1 \ldots u_i \ldots u_n$) – 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.
($v_1 \ldots v_j \ldots v_m$) – the welfare profiles of the people with negative welfare, in order of *ascending welfare*.

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

Sider's principle first groups 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 value of this person's welfare will take place and, consequently,

[10] Ng claims that, disregarding the Mere Addition Principle, theory X′ meets all Parfit's requirements on a population axiology and may be exactly the theory he is after (Ng, 1989, p. 245). I doubt that. Parfit rejects Average Utilitarianism exactly on the ground that it does not give enough weight to negative welfare, referring to an example similar to the one 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).

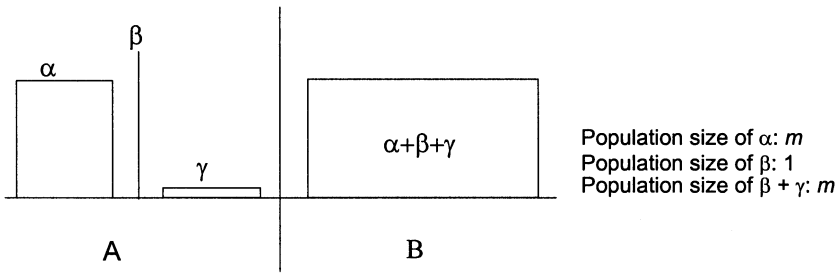

DIAGRAM 2

the more she will detract from the value of the population. The value of the person with the most negative welfare will not be dampened at all.

As Sider has shown, this principle does not violate the Mere Addition Principle.[12] Principle GV 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, principle GV mimics Total Utilitarianism.

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. 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 outcomes in spite of the fact that the γ-people have much higher welfare in outcome 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:

> *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. [13]

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

[13] 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.

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:

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

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 as $zk < x < z$, 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 generalized 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.[14]

If we look on the negative side of welfare our reasons for not advocating principle GV become even stronger. Assume that the world is crowded by lots of people, all living in the same hell full of illness and pain. Let us ponder whether to add two more people. One of these added people will have a life just barely worth living. The other one will have the kind of hellish life that is commonplace in this world. Since the number of unhappy lives is great the negative value of the extra unhappy life will be small – the weight assigned to her life will be small. The extra happy life will be the only happy life in this world and therefore must be assigned the weight one. Consequently, the negative value of the extra unhappy life will be outweighed by the positive value of the life barely worth living. According to Sider's principle, it is better to add the life barely worth living and the hellish life rather than to refrain from creating them.[15]

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

[15] I owe this argument to Krister Bykvist.

4. BLACKORBY, BOSSERT AND DONALDSON'S CRITICAL-LEVEL PRINCIPLE

Blackorby, Bossert and Donaldson's Critical-Level Utilitarianism (CLU) in its simplest form is a modified version of Total Utilitarianism.[16] The 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:

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

In the above formula, n is the number of people, u_i the welfare of individual i, and k is the critical level. Blackorby, Bossert and Donaldson assume a positive critical level, that is, the value of lives with positive welfare below the critical level is going to be negative. Consequently, the Repugnant Conclusion is deflected since the value of a huge population with low but positive welfare will be negative. The Non-Anti-Egalitarianism Principle is satisfied since an increase in the welfare of people below the critical level counts as much as an increase in the welfare of people above the critical level. It is easy to see, however, that CLU implies the Sadistic Conclusion.

In the diagram below, outcome A consists of one person with

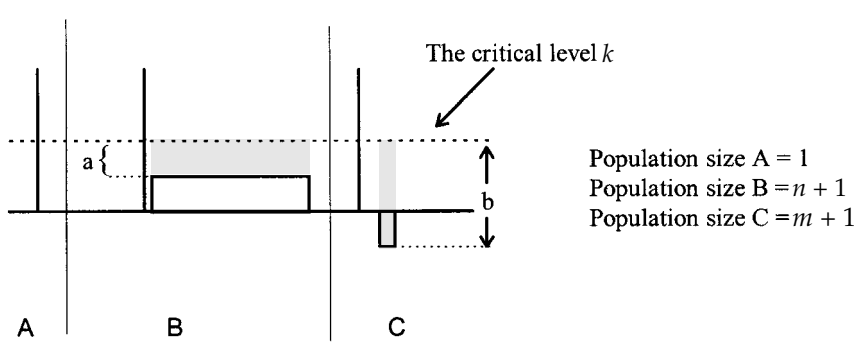

DIAGRAM 3

[16] See Blackorby, Bossert and Donaldson (1997, 1995) and Blackorby and Donaldson (1984). These authors also propose a more refined version of CLU where the 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 in Section 7.

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 has 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 instance of the Sadistic Conclusion.

CLU also implies a stronger version of the Sadistic Conclusion:

> *The Strong Sadistic Conclusion*: For any population consisting of people who all have negative welfare, there is another population whose existence would be *worse* although all its members have positive welfare.

There is always a population with enough 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.

The problem of finding an acceptable population axiology has often been conceived of as a problem of finding the right weighing of average against total welfare or, as it is sometimes expressed, between quality and quantity of welfare.[17] Thus, situations where both the average and the total positive welfare are increased has been seen as unproblematic. Consider the following principle:

> *The Different Number Dominance Principle*: If there is perfect equality in A, and there is higher average and total positive welfare in A as compared to B, then A is better than B.

Assume that everybody in outcome B has positive welfare below the critical level. In A, a number of people with *higher* welfare but below the critical level are added and the welfare of all the original people 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 every person in B.[18] According to CLU, however, A could be worse than B since the negative value of the added people could outweigh the value of the increase in the B-people's welfare.

[17] See, for example, Parfit (1984, pp. 401–3).

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

5. PRECONDITIONS OF THE IMPOSSIBILITY THEOREM

We shall understand a welfarist axiology as an 'at least as good as' ordering of all logically possible populations, solely based on people's welfare.[19] We shall include all the various interpretations of welfare, such as hedonist, preferentialist, objective list and so forth. This leaves the field open for a number of very odd axiologies, for example, one that implies that the higher people's welfare, the worse the population. The following weak principle can, however, be said to be part of any reasonable welfarist axiology:[20]

> *The Dominance Principle*: If population A contains the same number of people as population B, and every person in A has higher welfare than any person in B, then A is better than B.[21]

Another principle which is also, to the best of my knowledge, implicit in all welfarist axiologies in the literature is

> *The Addition Principle*: If it is bad to add a number of people, all with welfare lower than the original people, then it is at least as bad to add a greater number of people, all with even lower welfare than the original people.

Let us say that we can add a person with welfare y to a population where everybody enjoys a welfare above y. If this addition would be bad, then, according to the Addition Principle, it would be at least as bad to add several persons enjoying a welfare below y.

As a matter of fact, the theories presented in the literature entail the stronger principle that if it is bad to add a certain number of people, then it is worse to add an even greater number of people with lower welfare. According to Total Utilitarianism and Sider's principle GV, an addition of any number of people with negative welfare is bad and it is worse to

[19] We are using Sen's (1970, p. 9), terminology for orderings. In other words, we assume that the relation '_ is at least as good as _' should be reflexive, transitive and complete over the set of all logically possible populations. Cf. Arrow's (1963, Ch. VIII), Unrestricted Domain Condition. This definition could be weakened to include only nomologically possible populations, i.e., populations that are compatible with both the laws of logic and natural science. This is not the place to argue the pro and cons of such a restriction but the main idea is to exclude very improbable outcomes. Cf. Parfit (1984, pp. 388–9). We shall return to the completeness property in the last section of the paper.

[20] We shall shortly give a formal definition of the Dominance Principle and the other adequacy conditions.

[21] As one of the referees pointed out to me, the Dominance Principle is similar to but stronger than the Pareto principle formulated in terms of welfare. The former principle but not the latter also applies to cases involving different people in the compared populations.

add a greater number of people, each of whose welfare is even more negative. For Average Utilitarianism, an addition of people with welfare below the average is bad, and it is worse to add more people further below the average rather than less people closer to the average. Ng's theory X′ yields that some additions below the average are bad, and if it is bad to add a person with some welfare y below the average, then it is worse to add several persons with welfare below y. According to Critical Level Utilitarianism, it is bad to add people with welfare below the critical level and it is worse to add more people further below the critical level rather than less people closer to the critical level. For our purposes, however, the weaker principle formulated above is enough.

We saw above that Ng's theory X′ and Blackorby *et al.*'s Critical Level Utilitarianism respected the Non-Anti-Egalitarianism Principle but violated the Mere Addition Principle whereas the reverse holds for Sider's principle. This is not surprising since it is easy to show that no principle can satisfy the Mere Addition Principle, the Non-Anti-Egalitarianism Principle and avoid the Repugnant Conclusion.[22] Consequently, if one is not prepared to accept the Repugnant Conclusion, one has to reject either the Mere Addition or the Non-Anti-Egalitarianism Principle. The Non-Anti-Egalitarianism Principle has strong intuitive support.[23] We shall, however, adopt a weaker condition, namely avoidance of the following conclusion:

> *The Anti-Egalitarian Conclusion*: A population with perfect equality can be worse than a population with the same number of people, inequality, and lower average (and thus lower total) positive welfare.

The Mere Addition Principle, on the other hand, has been questioned by many authors and they have suggested dropping it to avoid the

[22] For an informal presentation of this result, see Ng (1989, p. 240). A formal exposition of a similar result can be found in Blackorby and Donaldson (1991). This paradox is similar to, but not the same as, Parfit's famous 'Mere Addition Paradox'. See Parfit (1984, pp. 419f). Cf. fn 24 below.

[23] As mentioned above, Sider's theory violates this principle. Sider rejects his own theory, however, just because it favours unequal distributions of welfare. See Sider (1991, p. 270, fn 10). Ng states that 'Non-Antiegalitarianism is extremely compelling'. See Ng (1989, p. 239, fn 4). Blackorby, Bossert and Donaldson (1997, p. 210), hold that 'weak inequality aversion is satisfied by all ethically attractive … principles'. Fehige (1998, p. 12), asks rhetorically '… 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?'. In personal communication, Parfit suggests that the Non-Anti-Egalitarianism Principle might not be convincing in cases where the quality of the good things in life are much worse in the perfectly equal population. We might assume, however, that the good things in life are of the same quality in the compared populations, but that in the perfectly equal population these things are equally distributed. Cf. the discussion of appeals to non-welfarist values in the last section.

Repugnant Conclusion.[24] We shall drop it as a criterion for an acceptable axiology, but show that this move does not get us out of a more general paradox. We shall show that any theory that satisfies the above mentioned requirements implies the Sadistic Conclusion or violates an intuitive principle that we shall now introduce.

It is easy to find a principle that satisfies the Dominance and the Addition Principles and avoids the Repugnant, the Sadist and the Anti-Egalitarian Conclusions. The Maximin Principle, for example, avoids these conclusions by giving priority to the worst-off. If the worst-off in A is better off than the worst-off in B, then A is better than B. This principle clearly avoids the Repugnant and the Anti-Egalitarian Conclusion and it does not imply the Sadistic Conclusion since the addition of a person with negative welfare always makes an outcome worse than any addition of people with positive welfare. It should be equally obvious that the Maximin Principle satisfies the Dominance and the Addition Principle. It violates, however, the following principle:

> *The Minimal Non-Extreme Priority Principle*: There is a number n such that an addition of n people with very high welfare and a single person with slightly negative welfare is at least as good as an addition of the same number of people but with very low positive welfare.

According to the Maximin Principle, if one population contains a person with negative welfare, and another does not, then the latter population is always better and the difference in positive welfare does not matter at all. In other words, a slight gain in welfare for one person outweighs a large loss for any number of people. Principles that violate the Minimal Non-Extreme Priority Principle give too much weight to negative welfare since they do not allow for any trade-offs between negative and positive welfare.[25]

The Minimal Non-Extreme Priority Principle is the last of our criteria for an acceptable axiology. It will be useful to list the presuppositions necessary for the theorem and state the adequacy conditions in a formal

[24] Ng (1989, p. 244), suggests that those who do not accept the Repugnant Conclusion should drop the Mere Addition Principle. Blackorby, Bossert and Donaldson (1995, p. 1305), and (1997, pp. 210–11), argue that if we have to choose between the Repugnant Conclusion and the Mere Addition Principle, then the latter must be rejected. Fehige (1998), holds that 'it's intrinsically wrong to bring people into existence who will have at least one unfulfilled preference'. In personal communication, Parfit rejects the Mere Addition Principle in cases where the added people in A+ have very low positive welfare as is the case in the above paradox. In his referee report he writes that 'if the extra people in A+ have lives that are only just worth living, most people find it easy to believe that A+ would be worse than A'. See also Feldman (1995) and Kavka (1982).

[25] See Arrhenius and Bykvist (1995, Ch. 3), for a discussion of the weight of negative welfare.

manner. A population is a finite set of possible lives. 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. A population is a set of lives under the restriction that one and the same person only occurs once in one and the same population. Unions of populations are also populations given that the aforementioned restriction is satisfied. Let A, B, C, $A \cup B$, and so on, denote populations. The number of lives in a population – the population size – is denoted by a subscript. For example, A_p denotes a population with p members. Let a_i denote a member of population A, b_i a member of population B, and so forth.

We shall assume that there are possible lives with positive or negative welfare. Furthermore, we shall assume that there are possible lives with very high positive welfare, very low positive welfare and slightly negative welfare.[26] Let W_{pw} be the set of all lives with positive welfare, W_{nw} the set of all lives with negative welfare, W_{vhp} the set of all lives with very high positive welfare, W_{vlp} the set of all lives with very low positive welfare, and W_{sn} the set of all lives with slightly negative welfare. We shall also assume that there are lives of at least four different levels of welfare in W_{vlp}.

The welfare statements above are all *categorical*, that is, of the general form '*a* has such-and-such welfare'. We also need to make some *comparative* welfare statements such as '*a* has higher (lower, the same) welfare as *b*'.[27] We shall assume that lives with positive welfare can be ordered by the relation 'has at least as high welfare as' such that people's welfare can be numerically represented in a manner that allows us to compare gains and losses of welfare.[28] The numerical representation of a life a_i's welfare is given by the function $v(a_i)$.

[26] That there are such possible lives is a common assumption, explicitly or implicitly, in the literature on population axiology and is, I think, a very plausible and common-sensical assumption. Again, we are not claiming that it is apparent how the above classification of lives looks in every detail – cf. fn 4. Notice also that we are not assuming that the above partitions of possible lives are exhaustive. There might, of course, be lives with neutral welfare but also some peculiar lives that cannot be grouped into any of these sets.

[27] From the categorical statements above some comparative statements follow conceptually. If John has positive and Chandra has negative welfare, then John has higher welfare than Chandra; if John has very high and Chandra has very low positive welfare, then John has higher welfare than Chandra and so forth. Notice that we have not assumed that lives belonging to the same welfare partition have the same level of welfare. It seems reasonable to assume that there are different levels of welfare among lives with, for example, very high positive welfare.

[28] This assumption is necessary for the application of the Anti-Egalitarian Conclusion which involves comparisons of average welfare. Comparisons of gains and losses of welfare presuppose measurement on a scale *at least* as strong as an interval scale, that is, a scale which is at least unique up to a positive linear transformation. See Roberts (1979, p. 64). As John Broome pointed out to me, there are some reasons for classifying a

We can now state our conditions as follows:

The Dominance Condition: If a_i has higher welfare than b_j for all $a_i \in An$, $b_j \in B_n$, then A_n is better than B_n.

The Addition Principle: If a_i has higher welfare than b_j, and b_j has higher welfare than c_h, for all $a_i \in A_k$, $b_j \in B_n$, $c_h \in C_m$, and A_k is better than $A_k \cup B_n$, and $m > n$, then $A_k \cup B_n$ is at least as good as $A_k \cup C_m$.

The Anti-Egalitarian Conclusion: There are populations A_n and B_n such that for all a_i, $a_j \in A_n$, a_i has the same welfare as a_j, but for some b_i, $b_j \in B_n$, b_i has lower welfare than b_j, and $[v(a_1)+\ldots v(a_n)]/n > [v(b_1)+\ldots v(b_n)]/n$, and B_n is better than A_n.

The Minimal Non-Extreme Priority Principle: There is an n such that if $A_n \subset W_{vhp}$, $B_1 \subset W_{sn}$, $C_{n+1} \subset W_{vlp}$, then $A_n \cup B_1 \cup D_k$ is at least as good as $C_{n+1} \cup D_k$, $k \geq 0$.

The Repugnant Conclusion: For any $A_n \subset W_{vhp}$ such that a_i has the same welfare as a_j for all a_i, $a_j \in A_n$, there is a $B_m \subset W_{vlp}$ such that B_m is better than A_n.

The Sadistic Conclusion: There are populations A_n, B_m, C_k such that $A_n \subset W_{pw}$, $B_m \subset W_{nw}$, $n, m > 0$, $k \geq 0$, and $B_m \cup C_k$ is better than $A_n \cup C_k$.

The primary claim of this paper is that any axiology that satisfies the Dominance, the Addition, and the Minimal Non-Extreme Priority Principle implies the Repugnant, the Anti-Egalitarian, or the Sadistic Conclusion. Let us turn to the proof.

6. THE IMPOSSIBILITY THEOREM

The Impossibility Theorem: There is no welfarist axiology that satisfies the Dominance, the Addition, and the Minimal Non-Extreme Priority Principle and avoids the Repugnant, the Sadistic and the Anti-Egalitarian Conclusion.

Proof: We show that the contrary assumption leads to a contradiction. Let $w_4 > w_3 > w_2 > w_1$ be four very low positive welfare levels. Consider the following populations (see Diagram 4):

A_p: A population with p members with very high welfare.
B_{q+1}: A population with q+1 members with very low positive welfare w_4.

numerical representation involving an interval scale *and* a stipulation that some objects should be represented by positive (negative) numbers as a ratio scale. There are also, to my mind, reasons for not classifying it as such a scale, but a proper treatment of this issue would take us too far afield.

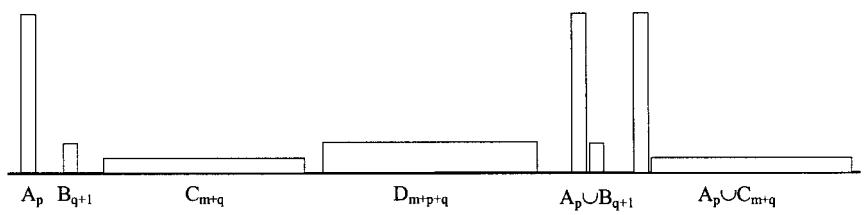

DIAGRAM 4

C_{m+q}: A population with $m+q$ members, $m \geq 2$, with very low positive welfare w_3 such that the average welfare of $A_p \cup C_{m+q}$ is less than w_4, that is, $[v(a_1) + \ldots v(a_p) + v(c_1) \ldots + v(c_{m+q})]/(m+p+q) < w_4$.
D_{m+p+q}: A population of the same size as $A_p \cup C_{m+q}$ with very low positive welfare w_4.

We start by proving that $A_p \cup B_{q+1}$ is at least as good as $A_p \cup C_{m+q}$ by showing that the contrary assumption leads to a contradiction.

Avoidance of the Repugnant Conclusion yields that there is at least one possible population with very high welfare that is at least as good as any population with very low positive welfare. Let **(1)** A_p be such a population. Since D_{m+p+q} is a population with very low welfare, avoidance of the Repugnant Conclusion yields that **(2)** A_p is at least as good as D_{m+p+q}. Avoidance of the Anti-Egalitarian Conclusion yields that **(3)** D_{m+p+q} is at least as good as A_pC_{m+q}. By transitivity, it follows from (2) and (3) that **(4)** A_p is at least as good as $A_p \cup C_{m+q}$. Assume that **(5)** $A_p \cup B_{q+1}$ is worse than $A_p \cup C_{m+q}$. By transitivity, it follows from (4) and (5) that **(6)** $A_p \cup B_{q+1}$ is worse than A_p. Since $m \geq 2$, it follows from (6) and the Addition Principle that **(7)** $A_p \cup B_{q+1}$ is at least as good as $A_p \cup C_{m+q}$ which contradicts (5). Hence, if we assume that $A_p \cup B_{q+1}$ is worse than $A_p \cup C_{m+q}$, then we get a contradiction. Thus, **(8)** $A_p \cup B_{q+1}$ is at least as good as $A_p \cup C_{m+q}$. Q.E.D.

Now consider the following populations (see Diagram 5):

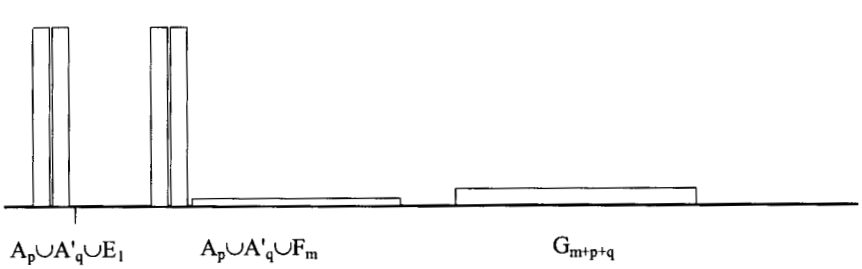

DIAGRAM 5

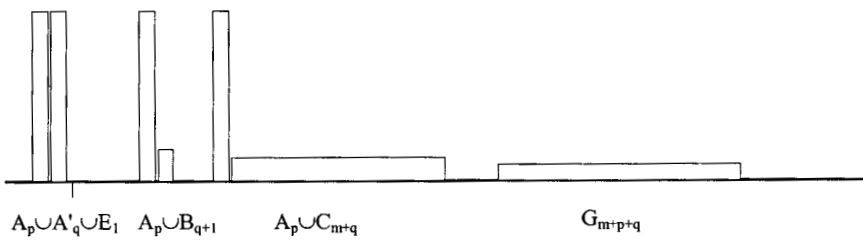

DIAGRAM 6

A_p, B_{q+1}, C_{m+q}: The same as above.
A'_q: A population with q members with very high welfare.
E_1: One person with slightly negative welfare.
F_m: A large population with very low positive welfare w_1 such that the average welfare of $A_p \cup A'_q \cup F_m$ is less than w_2, that is, $[v(a_1)+\ldots v(a_p)+\ldots v(a'_1)+\ldots v(a'_q) + v(f_1)+\ldots v(f_m)]/(m+p+q) < w_2$.
G_{m+p+q}: A population of the same size as $A_p \cup A'_q \cup F_m$ with very low positive welfare w_2.

Avoidance of the Sadistic Conclusion yields that **(9)** $A_p \cup A'_q \cup F_m$ is at least as good as $A_p \cup A'_q \cup E_1$. Since $A_p \cup A'_q \cup F_m$ is an anti-egalitarian alternative relative to G_{m+p+q}, **(10)** the latter population is at least as good as the former. By transitivity, (9) and (10), it follows that **(11)** G_{m+p+q} is at least as good as $A_p \cup A'_q \cup E_1$.

The Minimal Non-Extreme Priority Principle yields that there is at least some number such that an addition of such a number of people with very high welfare and a single person with slightly negative welfare is at least as good as an addition of the same number of people but with very low positive welfare. Let q be such a number. Accordingly, **(12)** $A_p \cup A'_q \cup E_1$ is at least as good as $A_p \cup B_{q+1}$ (see Diagram 6). The Dominance Principle yields that **(13)** G_{m+p+q} is worse than $A_p \cup C_{m+q}$. From (8) above, we know that $A_p \cup B_{q+1}$ is at least as good as $A_p \cup C_{m+q}$. By transitivity, it follows from (8), (12), and (13) that **(14)** G_{m+p+q} is worse than $A_p \cup A'_q \cup E_1$ which contradicts (11). Hence, the assumption that there is an axiology that satisfies all of the adequacy conditions leads to a contradiction. Thus, the impossibility theorem must be true. Q.E.D.

7. DISCUSSION

Sider conjectured that 'perhaps' a theory that can accommodate our considered beliefs in population axiology is 'just around the bend, and is a compromise axiological theory'.[29] Rather, what we seem to have found

[29] Sider (1991, p. 270).

around the bend is an impossibility theorem for the existence of an acceptable welfarist theory.

Which are the weakest links in the theorem? That the evaluations involved in the theorem are inconsistent is, of course, a prima-facie reason to give up at least one of them. We should scrutinize these beliefs and look for possible reasons for jettisoning them. The prospects look gloomy, however. The arguments that have been put forward for accepting, for example, the Repugnant Conclusion, are less than satisfactory.[30]

Another weak link could be that we have put too high demands on a welfarist axiology by assuming that it is a *complete* ordering of all possible populations. Perhaps some populations are incomparable in value, that is, some populations might neither be worse, nor equally as good as, nor better than some other populations. Thus, we should not expect more than a quasi-ordering of all possible populations.

It is true that the theorem above will not work without the completeness assumption. For example, in the proof above, we claimed that 'avoidance of the Repugnant Conclusion yields that there is at least one population with very high welfare that is at least as good as any population with very low positive welfare'. If an axiology is only a quasi-order, then this inference is not valid. It only follows that there is at least one population with very high welfare that is at least as good as, or *incomparable* to, any population with very low positive welfare.[31]

I am not completely convinced by this objection, however. It is important to remember that we have discussed welfarist principles. For an appeal to incommensurability to have any credibility as an escape from an impossibility theorem against welfarist axiologies, one must produce a good *welfarist* reason for incommensurability. The apparent source for incommensurability inside a welfarist framework is to reject interpersonal comparability of welfare. This move would certainly yield extensive incommensurability among populations but, I surmise, too extensive to be plausible and, moreover, it would lead to Arrovian impossibility theorems.[32] And given interpersonal comparability of

[30] See, for example, Tännsjö (1988). Carlson (1998, pp. 301–4) expresses some doubts regarding the unacceptability of the Sadistic Conclusion. Griffin's (1986, pp. 85–9, 338–40), interesting suggestion that there may be 'discontinuities' in the measurement of welfare can be interpreted to imply that sometimes we cannot compare people's welfare on a scale that makes sense of talk about average welfare. If so, then the Anti-Egalitarian Conclusion would not be applicable in such cases. In Arrhenius (1999), I show that this move does not solve the problem.

[31] The completeness assumption also plays a role in the *reductio* in the first part of the theorem, and in the application of the Anti-Egalitarian and the Sadistic Conclusion.

[32] See, for example, Roemer (1996, pp. 26–36) for a proof of Arrovian impossibility theorems with different measurement assumptions but no interpersonal comparability of welfare. Notice that these impossibility results already arise in a fixed population setting.

welfare, it is unclear why a welfarist would think that, for example, a population with very high welfare is incommensurable with a population with very low positive welfare.[33] The possibility of incommensurability among populations inside a welfarist framework deserves further attention since it might provide an escape route from the impossibility theorem, but I doubt that we will find a satisfactory solution here.[34]

Incommensurability among populations is pretty plausible 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 would not 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 of welfare but the other is better in respect of liberty, then these two population would be incommensurable if the above pluralism were true.

Pluralism need not imply incommensurability. There might be other values than welfare but the gains or losses in these values can be weighed against gains and losses in welfare. Still, it might be that an appeal to other values could deflect the impossibility theorem by calling into question some of the principles applied in the theorem. For example, one might object to the Dominance Principle on the grounds of desert: one might regard a general increase in welfare as a change for the worse if all the people who become better off are malefactors – they do not deserve it.[35]

The objection from pluralism, with or without incommensurability, seems to be easily met, however, since we can simply assume that the populations that we consider do not involve such values or are equally good in regard to such values. In other words, we can restrict the domain of the adequacy conditions to the set of all logically possible populations that only involve welfarist values or such populations that are equal in respect to all other values apart from welfarist values.

If it is true that there are no other values that we can appeal to in the outcomes that we have considered in this paper, then the impossibility theorem is even more troubling. Since it is reasonable to claim that an acceptable normative theory has to take welfare into account, the impossibility theorem cast doubts on the whole project of finding a normative theory that coheres with our considered moral beliefs.

[33] See Blackorby, Bossert and Donaldson (1997, pp. 218–19, 226) for an example of a principle with this implication. Their principle also avoids the Sadistic Conclusion by rendering the compared populations incommensurable.

[34] See Arrhenius (1999) for a theorem without the completeness assumption.

[35] See, for example, Temkin (1994, p. 353–6). Cf. Feldman (1995).

REFERENCES

Arrhenius, Gustaf. 1999. *Population Axiology*, Ph.D. dissertation. Department of Philosophy, University of Toronto

Arrhenius, Gustaf and Krister Bykvist. 1995. *Interpersonal Compensations and Moral Duties to Future Generations: Moral Aspects of Energy Use*. Uppsala Prints and Preprints in Philosophy, No. 21, Uppsala Universitet

Arrow, Kenneth J. 1963. *Social Choice and Individual Values*. 2nd edn. Yale University Press

Blackorby, Charles, Walter Bossert and David Donaldson. 1995. Intertemporal population ethics: critical-level utilitarian principles. *Econometrica*, 65:1303–20

Blackorby, Charles, Walter Bossert and David Donaldson. 1997. Critical-level utilitarianism and the population-ethics dilemma. *Economics and Philosophy*, 13:197–230

Blackorby, Charles and David Donaldson. 1992. Pigs and guinea pigs: a note on the ethics of animal exploitation. *The Economic Journal*, 102:1345–69

Blackorby, Charles and David Donaldson. 1991. Normative population theory: a comment. *Social Choice and Welfare*, 8:261–7

Blackorby, Charles and David Donaldson. 1984. Social criteria for evaluating population change. *Journal of Public Economics*, 25:13–33

Broome, John. 1993. Goodness is reducible to betterness: the evil of death is the value of life. In *The Good and the Economical: Ethical Choices in Economics and Management*, pp. 70–84. Peter Koslowski and Yuichi Shionoya (eds.).

Feldman, Fred. 1995. Justice, desert and the repugnant conclusion., *Utilitas*, Vol. 7, No. 2

Carlson, Erik. 1998. Mere addition and two trilemmas of population ethics. *Economics and Philosophy*, 14:283–306

Fehige, Christoph. 1998. A Pareto principle for possible people. In *Preferences*. Christoph Fehige and Ulla Wessels (eds.). de Gruyter

Griffin, James. 1986. *Well-Being: Its Meaning, Measurement, and Moral Importance*. Clarendon Press

Hudson, J. L. 1987. The diminishing marginal value of happy people. *Philosophical Studies*, 51:123–37

Hurka, Thomas. 1983. Value and population size. *Ethics*, 93:496–507

Kavka, Gregory. 1982. The paradox of future individuals. *Philosophy & Public Affairs*, 11:93–112

Ng, Yew-Kwang. 1989. What should we do about future generations? Impossibility of Parfit's theory X. *Economics and Philosophy*, 5:235–53

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

Parfit, Derek. 1986. Overpopulation and the quality of life. In *Applied Ethics*, pp. 145–64. P. Singer (ed.). Oxford University Press

Roberts, Fred S. 1979. *Measurement Theory with Applications to Decisionmaking, Utility, and the Social Sciences*, Addison-Wesley Publishing Company

Roemer, John E. 1996. *Theories of Distributive Justice*. Harvard University Press

J. Ryberg. 1996. *Topics on Population Ethics*. Ph.D. dissertation. University of Copenhagen

Sen, Amartya. 1970. *Collective Choice and Social Welfare*. Mathematical Economics Texts 5

Sider, Theodore R. 1991. Might theory X be a theory of diminishing marginal value? *Analysis*, 51:265–71

Singer, Peter. 1993. *Practical Ethics*, 2nd edn. Cambridge University Press

Tännsjö, Torbjörn. 1998. *Hedonistic Utilitarianism*. Edinburgh University Press

Temkin, Larry S. 1994. Weighing goods: some questions and comments. *Philosophy and Public Affairs*, Vol. 23, No. 4