Rysiek Sliwinski (ed.), *Philosophical crumbs. Essays dedicated to Ann-Mari Henschen-Dahlquist on the occasion of her seventy-fifth birthday*, Uppsala: Department of Philosophy, Uppsala University, 1999 (*Uppsala philosophical studies 49*), pp. 11-21.

AN IMPOSSIBILITY THEOREM IN POPULATION AXIOLOGY WITH WEAK ORDERING ASSUMPTIONS

Gustaf Arrhenius

Introduction

It has been known for quite a while now that the on-going project of constructing an acceptable population axiology has gloomy prospects. Already in Derek Parfit's seminal contribution to the topic, an informal paradox was presented and later contributions have proved similar results.[1] All of these contributions invoke, however, some version of a principle – the Mere Addition Principle – which is controversial.[2] In Arrhenius (1998), I presented a theorem which didn't invoke this controversial principle but replaced it with logically and intuitively weaker conditions. Still, however, one of the conditions in my theorem shares with these earlier results the presupposition that welfare can be measured on at least an interval scale.[3] One can deny this and, as a matter of

---

[1] See Parfit (1984), pp. 419ff. For an informal proof of a similar result with stronger assumptions, see Ng (1989), p. 240. A formal proof with slightly stronger assumptions than Ng's can be found in Blackorby and Donaldson (1991).

[2] The Mere Addition Principle and its cognates state something to the effect that an addition of people with positive welfare does not make a population worse. See Hudson (1986), Ng (1989), Sider (1991), Blackorby and Donaldson (1991). 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. Ng (1989), p. 244, suggests that those who don't accept the Repugnant Conclusion (see below for a statement of this conclusion) should drop the Mere Addition Principle. Blackorby, Bossert and Donaldson (1995), p. 1305, and (1997), pp. 210-1, 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". See also Feldman (1995) and Kavka (1982).

[3] An *interval* scale is unique up to a positive linear transformation. Given interpersonal comparability, people's gains and losses can be compared. See Roberts (1979), p. 64. The condition in question is the following mildly egalitarian principle: A population with perfect equality is at least as good as a population with the same number of people, inequality, and lower average positive welfare. Since this condition is formulated in terms of average welfare, it presupposes that welfare can be measured on at least an interval scale. Similar egalitarian

fact, one theorist has suggested this as a solution to Parfit's paradox.[4] The theorem that we shall present below invokes much weaker measurability assumptions. None of the conditions we shall use presuppose measurement of welfare on an interval scale. They only presuppose that some but not all lives with positive welfare can be ordered by the relation "has at least as high welfare as".

Another drawback of my earlier theorem is that it assumes complete comparability among populations which only differ in regards to welfare and population size. This assumption rules out that populations which only differ in these respects are incomparable in value.[5] The theorem presented below leaves the door open for such incomparabilities.

Assumptions and Definitions

For the purpose of proving the theorem, it will be useful to introduce a formal notation. It will enable us to state the adequacy conditions and presuppositions in an exact manner. We shall use capital letters to denote populations. Let's define populations A, B, C,… as finite sets of possible 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.[6] Notice that although we have assumed that every possible population is of finite size, we haven't assumed that there is a largest possible population. More exactly, we shall assume that for any possible population with a certain welfare, there is a larger possible population with the same welfare.[7] The number of lives in a population – the population size – is denoted by a subscript: h, i, j, k, m, and n for variables and p, q, r, s and t for constants. For example, A_p denotes a population with p

---

conditions, which thus also presupposes measurement on at least an interval scale, play a crucial role in the paradoxes/theorems of Parfit (1984), Ng (1989), and Blackorby and Donaldson (1991).

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

[5] See Blackorby, Bossert and Donaldson (1997), pp. 218-19, 226 for an example of a principle with this implication.

[6] Examples of possible populations is Rysiek with very high welfare; Rysiek and Erik with very high welfare; and the empty set. An example of an invalid population is Rysiek with very high welfare and Rysiek with very low welfare, that is, a population cannot contain two or more of the same person's possible lives.

[7] 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 welfare. This assumption plays an important role in the earlier paradoxes/theorems in the field, but has never, to the best of my knowledge, been formally noticed.

members. We shall use the corresponding lower-case letters to denote the members of a population: a_1, a_2,…a_i,…a_p, denotes members of population A_p, b_1, b_2,…b_i,...b_p, members of population B_p, and so forth.

We shall assume that there are possible lives with positive or negative welfare. Furthermore, we assume that there are possible lives with very high positive welfare, very low positive welfare or slightly negative welfare. These assumptions are common in the literature on population axiology and are so common-sense that it is hard to find any further arguments in favour of them.[8] Notice 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. Let's define W_{pw} as the set of all lives with positive welfare, W_{nw} as the set of all lives with negative welfare, W_{vhp} as the set of all lives with very high positive welfare, W_{vlp} as the set of all lives with very low positive welfare, W_{sn} as the set of all lives with slightly negative welfare.

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 than (as) *b*".[9] We shall assume that there are subsets of W_{vhp} and W_{vlp} which are ordered by the "has at least as high welfare as" relation. Let's define W_{vlp1}, W_{vlp2}, and W_{vlp3} as three subsets of W_{vlp} such that in each of these subsets, all lives have the same welfare and lives belonging to W_{vlp3} have higher welfare than lives belonging to W_{vlp2} who in turn have higher welfare than lives belonging to W_{vlp1}.

---

[8] It might be that these categorical statements can be reduced to comparative statements (cf. fn. below). One could hold, for example, that a life has positive (negative) welfare iff it is has higher (lower) welfare than an unconscious life. A number of more or less convincing proposals figure in the literature. For an instructive survey and critical discussion of these, see Broome (1993). At any rate, the truth of this matter wouldn't affect any of the arguments in this paper. If it turns out that categorical statements are reducible to comparative statements, then all that shows is that there is another set of presuppositions which is sufficient for the theorem. This would be interesting if one could show that such a set of presuppositions is logically and/or intuitively weaker than the one we have presented above. That is a complicated matter, however, and a full discussion of this topic would take us too far from the main task of this paper.

[9] From the categorical statements above some comparative statements follow conceptually. If Rysiek has positive and Erik has negative welfare, then Rysiek has higher welfare than Erik; if Rysiek has very high and Erik has very low positive welfare, then Rysiek has higher welfare than Erik; and so forth. Notice that we have not assumed that lives belonging to the same welfare partition share 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.

We shall understand a population axiology as a quasi-ordering of logically possible populations. In other words, we assume that the relation "is at least as good as" should be reflexive, transitive but not necessarily complete over the set of all logically possible populations.[10] This is a minimal and very undemanding definition of a population axiology. Notice that we leave open the possibility that there might be incommensurable populations. 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.[11] Although we shall not defend this claim, this assumption is arguably a minimal adequacy condition for any moral theory.

Adequacy Conditions

We shall now introduce the adequacy conditions which we shall employ in the theorem. We shall give both an informal and a formal statement of the conditions. To illustrate the independence of the different conditions, we shall give examples of theories which violate just one of them.

According to Total Utilitarianism, the value of a population is calculated by summing the welfare of all lives in the population. A well-known implication of Total Utilitarianism is Parfit's Repugnant Conclusion: For any population with very high positive welfare, there is a population with very low positive welfare which is better.[12] Total Utilitarianism implications in this area are of a more general nature, however. Total Utilitarianism violates the following condition:

> *The Quality Addition Condition (informal formulation)*: There is at least one population with very high welfare such that its addition is at least as good as an addition of any population with very low positive welfare, other things being equal.

---

[10] We're using Sen's (1970), p. 9, terminology for orderings. The above definition could be weakened to include only nomologically possible populations, that is, 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 all too unrealistic outcomes. Cf. Parfit (1984), pp. 388-389.

[11] We shall include all the various interpretations of welfare, such as experientialist theories (e.g., hedonism), desire theories (e.g., preferentialism), objective list theories (e.g., perfectionism), and so forth.

[12] See Parfit (1984), p. 388. My formulation is more general than Parfit's.

> *The Quality Addition Condition*: There is an $A_n \subset W_{vhp}$ such that for all $B_m \subset W_{vlp}$, $A_n \cup C_k$ is at least as good as $B_m \cup C_k$, $k \geq 0$, other things being equal.

If the total sum of welfare in the population with very low positive welfare is higher than the total sum of welfare in the population with very high welfare, then, according to Total Utilitarianism, it is better to add the former population rather than the latter. Since such a population can be found for any population with very high welfare, Total Utilitarianism violates the above condition.

The Maximax principle ranks populations according to the welfare of the best off: The higher the welfare of the best off, the better the population, and if the best off are equally well off in two populations, then these populations are equally good. Consequently, this principle satisfies the Quality Addition Condition but it violates the following weak egalitarian condition:

> *The Minimal Inequality Aversion Condition (informal formulation)*: For any welfare levels of the best off and the worst off people, and for any n, there is an $m > n$ such that for population of n best off and m worst off, a loss for the n best off can be balanced (compensated for) by a gain for the m worst off such that everybody would be equally well off, other things being equal.

> *The Minimal Inequality Aversion Condition*: For any A_n such that a_i has the same welfare as a_j for all $i, j \leq n$, there is an $m > n$ such that if b_i has the same welfare as b_j for all $i, j \leq m$, c_i has the same welfare as c_j for all $i, j \leq m + n$, and a_i and c_j has higher welfare than b_k for all $i \leq n, j \leq m + n, k \leq m$, then C_{m+n} is at least as good as $A_n \cup B_m$, other things being equal.

Since Maximax only cares about the welfare of the best off, it doesn't matter how many worst off people we can benefit, because this benefit can never compensate for a loss in welfare of the best off.

A principle that satisfies both of the conditions above is Blackorby, Bossert and Donaldson's Critical-Level Utilitarianism. In its simplest form, this is a modified version of Total Utilitarianism.[13] 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. Since the contributive value of lives with positive welfare below the critical

---

[13] 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 above.

level is negative, this theory implies what I call Sadistic Conclusions: An addition of lives with negative welfare can be better than an addition of lives with positive welfare.[14] Let's say that the critical level is 10u. Then the value of a population of 10 people with positive welfare 5u is 10(5-10) = -50, whereas the value of a population consisting of one person with negative welfare -10u is -20. Critical-Level Utilitarianism violates the following condition:

> *The Non-Sadism Condition (informal formulation)*: 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 Non-Sadism Condition*: If $A_n \subset W_{pw}$, $B_m \subset W_{nw}$, $n, m > 0$, then $A_n \cup C_k$ is at least as good as $B_m \cup C_k$, $k \geq 0$, other things being equal.

Maximin is an example of a principle that satisfies the Non-Sadism Condition and the other conditions above. 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 enjoys the same welfare in two populations, then these populations are equally good. This principle violates, however, the following condition:

> *The Minimal Non-Extreme Priority Condition (informal formulation)*: There is an *n* such that a decrease from very high welfare to very low positive welfare for *n* people cannot be balanced (compensated for) by an increase from slightly negative welfare to very low positive welfare for one person, other things being equal.
>
> *The Minimal Non-Extreme Priority Condition*: 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 $C_{n+1} \cup D_k$, $k \geq 0$, other things being equal.

According to the Maximin Principle, 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.

The following and last of our adequacy conditions is as uncontroversial as it gets in population axiology:

---

[14] I introduced this conclusion in Arrhenius and Bykvist (1995). See Arrhenius (1998) for a discussion of this conclusion in connection with Critical-Level Utilitarianism.

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

The Egalitarian Dominance Condition: If a_i has the same welfare as a_j, and a_i has higher welfare than b_j for all *i, j*, then A_n is better than B_n other things being equal.

A principle, let's call it the Complete Indifference Principle, which ranks all possible populations as equally good, violates this condition but trivially satisfies all of the other conditions discussed above.

To properly establish the independence of all the adequacy conditions, we have to show that Total Utilitarianism satisfies all conditions except the Quality Addition Condition, that Maximax satisfies all conditions except the Minimal Inequality Aversion Condition, and so forth. This is a pretty easy but boring task, so we shall leave it as an exercise for the ambitious reader.

The Impossibility Theorem

The Impossibility Theorem: There is no population axiology which satisfies the Dominance, the Minimal Inequality Aversion, the Minimal Non-Extreme Priority, the Non-Sadism, and the Quality Addition Condition.

Proof: We show that the contrary assumption leads to a contradiction. Consider the following populations:

$A_p \subset W_{vhp}$: A population with $p \geq 1$ members with the same very high welfare.

$A'_q \subset W_{vhp}$: A population with $q \geq 1$ members with the same welfare as the A-lives.

$B_{q+1} \subset W_{vlp3}$: A population with $q + 1$ members with very low positive welfare.

$C_{p+r-1} \subset W_{vlp3}$: A population with $p + r - 1$ members with very low positive welfare.

$E_1 \subset W_{sn}$: One person with slightly negative welfare.

$F_r \subset W_{vlp1}$: A population with $r \geq 1$ members with very low positive welfare.

$G_{p+q+r} \subset W_{vlp2}$: A population of the same size as $A_p \cup A'_q \cup F_r$ with very low positive welfare.

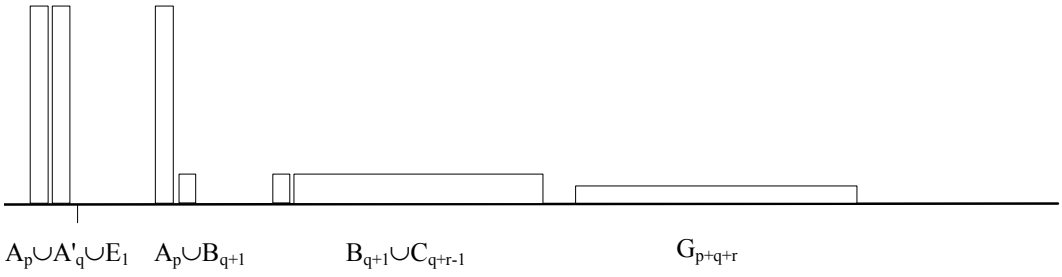

Diagram 1[15]

The Minimal Non-Extreme Priority Condition yields that there is at least some number *n* such that a decrease from very high welfare to very low positive welfare for *n* people cannot be balanced by an increase from slightly negative welfare to very low positive welfare for one person. Define *q* as (one of) this (these) number(s). Accordingly, $A_p \cup A'_q \cup E_1$ is at least as good as $A_p \cup B_{q+1}$ (see Diagram 1). The Quality Addition Condition yields that there is at least one possible population with very high welfare such that its addition is at least as good as an addition of any population with very low positive welfare. Define A_p as (one of) this (these) population(s). Thus, $A_p \cup B_{q+1}$ is at least as good as $B_{q+1} \cup C_{p+r-1}$. The Egalitarian Dominance Condition yields that G_{p+q+r} is worse than $B_{q+1} \cup C_{p+r-1}$. By transitivity, it follows that G_{p+q+r} is worse than $A_p \cup A'_q \cup E_1$.

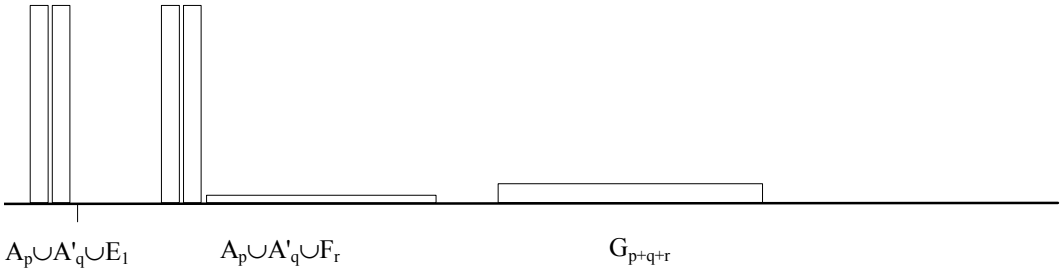

Diagram 2

The Minimal Inequality Aversion Condition yields that there is at least some number $m > p + q$ such that for a population of $p + q$ best off and *m* worst off, a

[15] As is common in the literature on population axiology, we shall use diagrams as an heuristic device to make it easier to follow the steps in the arguments. It is important to remember, however, that these diagrams are nothing more than just heuristic devices. 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 cannot properly be said to represent the average welfare of a population since we haven't assumed that welfare can be measured on at least an interval scale.

loss for the $p + q$ best off can be balanced by a gain for the m worst off such that everybody would be equally well off. Define r as (one of) this (these) number(s). Accordingly, G_{p+q+r} is at least as good as $A_p \cup A'_q \cup F_r$ (see Diagram 2). The Non-Sadism Condition yields that $A_p \cup A'_q \cup F_r$ is at least as good as $A_p \cup A'_q \cup E_1$. It follows by transitivity that G_{p+q+r} is at least as good as $A_p \cup A'_q \cup E_1$. Hence, we have derived a contradiction: G_{p+q+r} is worse than $A_p \cup A'_q \cup E_1$ and G_{p+q+r} is at least as good as $A_p \cup A'_q \cup E_1$. Thus, the impossibility theorem must be true. Q.E.D.

Discussion

In our discussion above 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 Minimal Inequality Aversion Condition wouldn't make sense. The theorem is pretty flexible, on the other hand, in regards to measurement of welfare. It doesn't matter whether welfare is measurable on an ordinal, interval or ratio scale, for example.[16] The theorem is valid as long as lives with very low or very high positive welfare are ordered by the relation "has at least as high welfare as".

It is interesting to compare the information demands of the present theorem with that of Arrow's famous impossibility theorem.[17] 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.[18] Not surprisingly then, the standard remedy for Arrowian impossibility results is to introduce some kind of interpersonal comparability of welfare.[19] But with interpersonal comparability of welfare, and some minimal demands on the orderings of lives, we run into the impossibility theorem presented in this paper.[20]

---

16 An ordinal scale is unique up to an order-preserving transformation, whereas a ratio scale is unique up to a similarity transformation. See Roberts (1979), p. 64.

17 See Arrow(1963). Notice that Arrow's result appears already in a fixed population size setting.

18 See Roemer (1996), pp. 26-36. I suspect that Roemer's theorem can be extended to cover non-interpersonal comparable measurement on any scale at least as strong as the ordinal scale but I haven't yet found nor figured out a proof for this.

19 Roemer (1996), p. 36, among many others, suggests this.

20 I would like to thank Krister Bykvist, Erik Carlson, Sven Danielsson, Adeze Igboemeka, Jan Odelstad, Derek Parfit, Rysiek Sliwinski, Howard Sobel and Wayne Sumner for their very

References

Gustaf Arrhenius, "An Impossibility Theorem for Welfarist Axiologies", mimeo., Uppsala Universitet, 1998.

Gustaf Arrhenius and Krister Bykvist, *Interpersonal Compensations and Moral Duties to Future Generations: Moral Aspects of Energy Use*, Uppsala Prints and Preprints in Philosophy, #21, Uppsala Universitet, 1995.

Kenneth J. Arrow, *Social Choice and Individual Values*, 2nd ed., New Haven and London: Yale UP, 1963.

Charles Blackorby, Walter Bossert, and David Donaldson, "Critical-Level Utilitarianism and the Population-Ethics Dilemma", *Economics and Philosophy*, 13, 197-230, 1997.

Charles Blackorby, Walter Bossert, and David Donaldson, "Intertemporal Population Ethics: Critical-level Utilitarian Principles", *Econometrica* 65, 1303-1320, 1995.

Charles Blackorby and David Donaldson, "Pigs and Guinea Pigs: A Note on the Ethics of Animal Exploitation", *The Economic Journal*, 102, 1345-69, November, 1992.

Charles Blackorby and David Donaldson, "Normative Population Theory: A Comment", *Social Choice and Welfare*, 8:261-7, 1991.

Charles Blackorby and David Donaldson, "Social Criteria for Evaluating Population Change", *Journal of Public Economics*, 25, 13-33, 1984.

John Broome, "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*, Peter Koslowski and Yuichi Shionoya (eds.), pp. 70-84, Springer-Verlag, 1993.

Fred Feldman, "Justice, Desert and the Repugnant Conclusion", *Utilitas*, Vol. 7, No. 2, November, 1995.

Christoph Fehige, "A Pareto Principle for Possible People", in Christoph Fehige and Ulla Wessels (eds.), *Preferences*, de Gruyter: Berlin/New York, 1998.

James Griffin, *Well-Being,* Clarendon Press: Oxford UP, 1986.

J. L. Hudson, "The Diminishing Marginal Value of Happy People", *Philosophical Studies* 51, 123-37, 1987.

Gregory Kavka, "The Paradox of Future Individuals", *Philosophy & Public Affairs* 11, 93-112, 1982.

Yew-Kwang Ng, "What Should We Do About Future Generations? Impossibility of Parfit's Theory X", *Economics and Philosophy* 5(2), 235-253, 1989.

Derek Parfit, *Reasons and Persons*, Oxford: Oxford UP, 1984.

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

John E. Roemer, *Theories of Distributive Justice*, Cambridge, Mass., Harvard UP, 1996

---

helpful and detailed comments on my earlier ideas in this area which facilitated the writing of this paper.

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

Theodore R. Sider, "Might Theory X Be a Theory of Diminishing Marginal Value?", *Analysis* 51 (4), 265-271, 1991.