Prolegomena 22 (2) 2023: 231-236
doi: https://doi.org/10.26362/20230204

Rejection of the Impossibility Theorem for Theory X

B.V.E. HYDE
Durham University, 50 Old Elvet, Durham, United Kingdom DH1 3HN
University of Religions and Denominations, پردیسان، بلوار امامصادق (علیه السلام),
Qom, Iran, 37491 13357
b.v.e.hyde@outlook.com

ORIGINAL RESEARCH PAPER – RECEIVED: 30/3/2023 ACCEPTED: 19/9/2023

ABSTRACT: The principal aims of population axiology are to increase the wellbeing of everyone, to prevent the suffering of future generations, and to make everyone more equal in these respects. A crisis in the pursuit of these goals came when Derek Parfit (1984) suggested that they inevitably result in a repugnant conclusion, that for any happy world, a miserable world of people whose lives were just barely worth living would be better, were it sufficiently populous. Since then, Gustaf Arrhenius (2000) has shown that these same principles also lead to a sadistic conclusion, that it can be better to add people with negative welfare rather than positive welfare when adding people without affecting the original people's welfare. What is more, he showed that there is no welfarist axiology that satisfies these three principles and yet avoids the repugnant conclusion. He called this the impossibility theorem for Theory X. This essay maintains that the ninth premiss of the impossibility theorem contains an invalid inference, and therefore presents a disproof of the theorem.

KEY WORDS: Egalitarianism, population axiology, utilitarianism, welfare, wellbeing.

In classical utilitarianism, the ends justify the means, the end is maximum welfare, and this end is totally impartial. If we are to take this to its logical extremes, we find either a repugnant or a sadistic conclusion:

> The Repugnant Conclusion (RC): For any happy world, a miserable world of people whose lives were just barely worth living would be better, were it sufficiently populous (Parfit 1984: 388).
>
> The Sadistic Conclusion (SC): 'When adding people without affecting the original people's welfare, it can be better to add people with negative welfare rather than positive welfare' (Arrhenius 2000: 251).

Utilitarianism, however, is generally quite useful, and because we like it we want to keep it. In fact, it is so intuitively appealing, and the repugnant conclusion is so difficult to avoid, that some are even willing to accept that their moral theories are repugnant just to retain the utilitarian aspect of them (Zuber et al. 2021). Those unwilling to accept that utilitarian population axiologies are repugnant must, therefore, search for an axiological theory – call it Theory X – that entails the veracity of utilitarianism, but not the repugnant conclusion. More precisely, we want the following principles to obtain:

> The Dominance Principle (DP): 'If worlds *x* and *y* are so related that *x* would be the result of increasing the well-being of everyone in *y* by some amount and adding some new people with worthwhile lives, then *x* is better than *y* with respect to utility' (Huemer 2008: 902).
>
> The Addition Principle (AP): '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' (Arrhenius 2000: 257).
>
> Non-Anti-Egalitarianism (NAE): 'If alternative B has the same set of individuals as in alternative A, with all individuals in B enjoying the same level of utility as each other, and with a higher total utility than A, then, other things being equal, alternative B must be regarded as better than alternative A' (Ng 1989: 238).

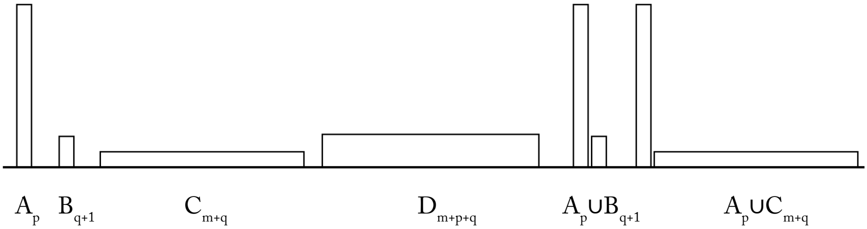
And we want to reject the following:

> The Impossibility Theorem: There is no welfarist axiology that satisfies the dominance principle, the addition principle and non-anti-egalitarianism, and yet avoids the repugnant conclusion (Arrhenius 2000: 261).

If the impossibility theorem can be denied, then we can hope to find a possible Theory X (e.g. Sider 1991: 270); if it obtains, Theory X is impossible (e.g. Arrhenius 2000).

The Impossibility Theorem

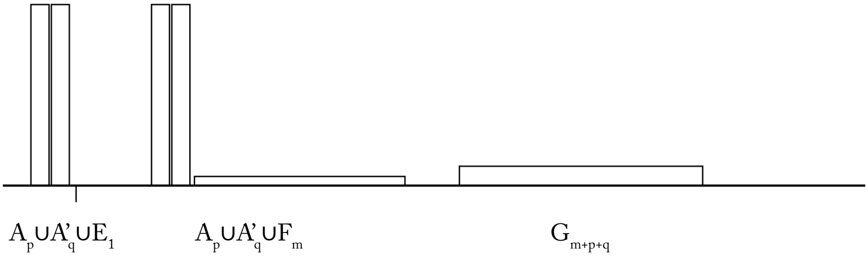
A proof of the impossibility theorem was offered by Gustaf Arrhenius (2000: 261–263). Let us reconstruct it here.

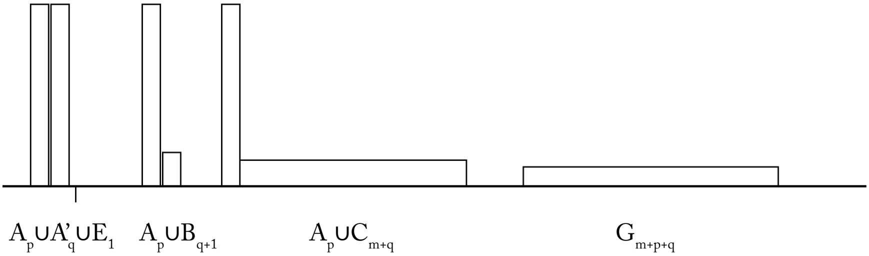


A_p: A population with *p* members with very high welfare.
B_{q+1}: A population with *q*+1 members with very low positive welfare w_4.

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 $A_p \cup C_{m+q} < w_4$.
D_{m+p+q}: A population of the same size as $A_p \cup C_{m+q}$ with very low positive welfare w_4.

The first part is the proof that $A_p \cup B_{q+1} \geq A_p \cup C_{m+q}$. **(1)** ¬RC entails that there is a possible population (A_p) with very high welfare that is at least as good as any population with very low positive welfare. **(2)** D_{m+p+q} has low welfare, so ¬RC → $A_p \geq D_{m+p+q}$. **(3)** NAE → $D_{m+p+q} \geq A_p \cup C_{m+q}$. By transitivity, (2) & (3) → **(4)** $A_p \geq A_p \cup C_{m+q}$. Assume that **(5)** $A_p \cup B_{q+1} < A_p \cup C_{m+q}$. By transitivity, it follows from (4) and (5) that **(6)** $A_p \cup B_{q+1} < A_p$. Since $m \geq 2$, (6) & AP → **(7)** $A_p \cup B_{q+1} \geq A_p \cup C_{m+q}$, which contradicts (5). $A_p \cup B_{q+1} < A_p \cup C_{m+q} \rightarrow \bot$, so, by *modus tollens*, **(8)** $A_p \cup B_{q+1} \geq A_p \cup C_{m+q}$ (Q.E.D.).





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 $A_p \cup A'_q \cup F_m < w_2$.
G_{m+p+q}: A population the size of $A_p \cup A'_q \cup F_m$ with very low positive welfare w_2.

The second part starts with the premiss that **(9)** ¬SC → $A_p \cup A'_q \cup F_m \geq A_p \cup A'_q \cup E_1$. By virtue of NAE, **(10)** $G_{m+p+q} \geq A_p \cup A'_q \cup F_m$, and by transitivity from (9) and (10), **(11)** $G_{m+p+q} \geq A_p \cup A'_q \cup E_1$. **(12)** DP → $G_{m+p+q} < A_p \cup C_{m+q}$. By transitivity, it follows from (8) and (12) that $G_{m+p+q} < A_p \cup A'_q \cup E_1$, which contradicts (11). Hence, the assumption that the impossibility theorem fails leads to a contradiction and, by *modus tollens*, the impossibility theorem must be true (Q.E.D.).

The Impossibility Theorem Denied

If we will grant that the impossibility theorem is valid, then we are left with three options (Arrhenius et al. 2017: 1) abandon some of the principles underlying the theorem, 2) become moral skeptics, or 3) explain away the significance of the proofs. The problem we have is that we simply cannot deny any of the three principles (the first option) without also becoming moral skeptics (the second), and if we are to be skeptical about morality, then the enterprise of population axiology loses its significance. If we are to care at all, and we will take it as a term of engagement with our contemporaries that we should, then the only option available to us is to either explain away the significance of the impossibility theorem, or to accept either, or both, the repugnant and sadistic conclusions.

We must grant the first part of the proof because the conclusion ($A_p \cup B_{q+1} \geq A_p \cup C_{m+q}$) is consistent with Theory X. The second part of the proof is where we are forced into accepting increasingly unpleasant conclusions as inferences from pleasant principles. This first occurs in premiss 9. However, we maintain that premiss 9 contains an invalid inference.

It does not follow from the denial of the sadistic conclusion that $A_p \cup A'_q \cup F_m \geq A_p \cup A'_q \cup E_1$. What follows is that $\neg (A_p \cup A'_q \cup E_1 > A_p \cup A'_q \cup F_m)$. Namely, denying the sadistic conclusion is to deny that it is better to add a lesser number of people with negative welfare than a greater number of people with low but positive welfare. It does not follow from this that adding a greater number with low welfare is *better than* (or equal to) adding a lesser number with negative welfare. Only an either/or entails that if it is not the one then it is the other, but value cannot be placed on such a binary scale: it is quite understandable to say that we want both options, or that we want neither, and it does not follow from our wanting one, or choosing one, that we did not want the other; likewise, it does not follow from our distaste of the sadistic conclusion that we would want its opposite.

This is, essentially, a denial of the circumstantial transitivity of value: it is denied that value is transitive in the sense that, if *x* is more preferable than *y*, and *y* more preferable than *z*, then *x* is more preferable than *z*. If we are to make such a denial, another criticism follows: Some values seem transitive *ceteris paribus* but may not be transitive *circumstantially*. Observe that in the diagram $A_p \cup A'_q \cup E_1$ has just one person with slightly negative welfare, whereas $A_p \cup A'_q \cup F_m$ has many miserable people whose lives are still barely worth living. In such a circumstance, we would clearly prefer the former world to the latter: the sadistic conclusion does not seem so sadistic *in this circumstance*, and we would not therefore reject it, even though we reject the principle, *ceteris paribus*, that it is better to add a lesser number of people with negative welfare than a greater number of people with low but positive welfare.

Perhaps it would be said that, even in this circumstance, we would still reject the sadistic conclusion, and we would say that $A_p \cup A'_q \cup F_m \geq A_p \cup A'_q \cup E_1$. Ivan Karamazov says that 'it's not worth the tears of that one tortured child' (Dostoyevsky 1912: bk. ii, ch. iv). However, it should be observed that we regularly do sacrifice the one for the many. We give them the title of 'martyr' or 'hero' and are done with the matter. It is a romantic ideal to claim to a life at the expense of nobody; it is the very essence of life – observe the evolutionary principle, or what Schopenhauer (1819) called the 'will to live' – to clamber upon one another and live at one another's expense. Normally we are content living at the considerable expense of a great many. To live gloriously at the minor expense of just one is something that, realistically, every one of us would accept.

A retort that we can imagine is that we are not sacrificing but adding people. The 'sacrifice' of the one does nothing to improve the lives of the many, but only contributes towards a foolish 'average' – a statistic. Indeed, this is quite true, if in fact we are adding. But we are not. Observe a footnote from Michael Huemer (2008: 902, n. 7): 'The notion of "adding" people to a world need not be taken to denote a temporal process; rather, when we have imagined a possible world, we "add" people to it by imagining another world just like the first but with additional people'. And in this possible world with additional people, those that are very happy are so at the expense of the miserable one. He is, in fact, a necessary sacrifice, and one that we are all willing to make.

So what we have found is that, in our search for Theory X, we may hope to find it. Derek Parfit was right to be optimistic about it, as he was until the very end of his life, at which time he believed himself to have found a solution to the theoretical problems that had prompted him to seek Theory X in the first place (Parfit 2017). Or, at least, he was right to not be pessimistic – unlike Gustaf Arrhenius, who was wrong in this respect.

References

Arrhenius, G. 2000. "An impossibility theorem for welfarist axiologies," *Economics and Philosophy*, 16(2): 247–266.

Arrhenius, G., J. Ryberg, and T. Tännsjö. 2017. "The repugnant conclusion." In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* Stanford: Stanford University Metaphysics Research Lab. (https://plato.stanford.edu/entries/repugnant-conclusion/) (2 November 2023)

Dostoyevsky, F. 1912. *The Brothers Karamazov.* C. Garnett, transl. (New York: The Lowell Press).

Huemer, M. 2008. "In defence of repugnance," *Mind*, 117(468): 899–933.

Ng, Yew-Kwang. 1989. "What should we do about future generations? Impossibility of Parfit's Theory X," *Economics and Philosophy*, 5(2): 235–253.

Parfit, D. 1984. *Reasons and Persons* (Oxford: Clarendon Press).

Parfit, D. 2017. "Future people, the non-identity problem, and person-affecting principles," *Philosophy and Public Affairs*, 45(2): 118–157.

Schopenhauer, A. 1819. *The World as Will and Representation* (Leipzig: Johann Friedrich Hartknoch).

Sider, Th. R. 1991. "Might Theory X be a theory of diminishing marginal value?" *Analysis*, 51(4): 265–271.

Zuber, S., N. Venkatesh, T. Tännsjö, Ch. Tarsney, H. O. Stefánsson, K. Steele, D. Spears, J. Sebo, M. Pivato, T. Ord, Yew-Kwang Ng, M. Masny, W. MacAskill, N. Lawson, K. Kuruc, M. Hutchinson, J. E. Gustafsson, H. Greaves, L. Forsberg, M. Fleurbaey, D. Coffey, S. Cato, C. Castro, T. Campbell, M. Budolfson, J. Broome, A. Berger, N. Beckstead, and G. B. Asheim. 2021. "What should we agree on about the repugnant conclusion?" *Utilitas*, 33(4): 379–383.