Some Potential Loopholes for Welfarist Axiology

Walter Barta

University of Houston

Fall 2022 (DRAFT)

Introduction

In what follows, we will attempt to point out two technical loopholes in the proof of the "Impossibility Theorem for Welfarist Axiologies" as proposed by Gustaf Arrhenius (Arrhenius, 2000). The two problems arise from different principles, one from an application of the "Addition Principle" and one from an application of the "Dominance Principle" after a counterintuitive combination of the "No Repugnant/Anti-Egalitarian Conclusion" criteria, the former non-fatal and latter fatal. We will show that these problems arise via two methodologies. 1) We appeal to intuitions with counterexample cases. 2) We replace ordinal priority-ordering with cardinal utility functions (von Neumann & Morgenstern, 1944): more specifically, we use a classical utilitarian cardinal utility function, a "greatest good for the greatest number" equivalence between total welfare and the product of the number of persons and the average welfare level for a given population ($W_i = w_i(n_i)$); we redo the proof using this equivalence; then we show that the proof is invalid for our assumed utility function, and how this explains our counterexample cases (List, 2022). If correct, our critique shows that at least one avenue is available to welfarist axiology, which reopens the possibilities of the search for a "Theory X" (Parfit, 1984).

(Note: we include the ordinal and cardinal versions of the reconstructed proof in the appendix for reference, which the reader may find useful to review if they are unfamiliar with Arrhenius's argument.)

The No Repugnant and The No Anti-Egalitarian Criteria Conflict?

The first problem is with the application in step (2) and step (3) of the "No Repugnant Criteria" and "Anti-Egalitarian Criteria," criteria which attempt to exclude two counterintuitive conclusions respectively:

> **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. (p. 248)

And:

> **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. (p. 258)

Although excluding each of these conclusions separately is intuitive enough, when excluded together via the transitivity principle in step (4) the result is counterintuitive. Namely, step (4) ends up contradicting the assumption of the relative welfare levels ($w_4 > w_3 > w_2$) (p. 261). Subsequently in step (13) there is an application of the "Dominance Principle":

> **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. (p. 257)

However, the application of the Dominance Principle here is invalid because it requires the assumption of relative welfare levels ($w_4 > w_3 > w_2$) to be applied (p. 263), an assumption which was contradicted by our analysis of step (4).

Comparing Ordinal Betterness

Using an ordinal betterness comparison version of the argument, deriving step (4) is straightforward and no individual step appears controversial, but step (4) itself is counterintuitive.

In step (1), we simply assume that some positive welfare population A_p exists (p. 262):

$$A_p$$

In step (2), we reject the Repugnant Conclusion by assuming that A_p must be better than some larger population D_{m+p+q} of smaller average positive welfare (p. 262):

$$A_p \geq D_{m+p+q}$$

In step (3), we reject the Anti-Egalitarian Conclusion by assuming that a given large but egalitarian population D_{m+p+q} must be better than a population $A_p \cup C_{m+q}$ that has the same size population and is universal positive welfare (w_3) but happens to be inegalitarian (p. 262):

$$D_{m+p+q} \geq A_p \cup C_{m+q}$$

In step (4), we apply the principle of transitivity to step (2) and step (3) to exclude both the Repugnant and the Anti-Egalitarian Conclusions (p. 262):

$$A_p \geq A_p \cup C_{m+q}$$

However, we can intuit that something may be amiss by looking at the structure of step (4): Population A_p is better than itself A_p plus the additional low positive welfare population C_{m+q}. This result seems to suggest that adding a positive welfare population C_{m+q} to a positive welfare population A_p *is worse* than the original population A_p by itself, which is prima facie counterintuitive.

This conclusion arises from attempting to avoid the "Repugnant Conclusion" and the "Anti-Egalitarian Conclusion" simultaneously and also prepare for the transitivity relation of step (6) in the proof, but the result is that there must be a high level welfare population A_p that is better than lower positive welfare level populations, including ones with both 1) more persons and 2) more inequality.

To better understand just how counterintuitive this result is, consider the following illustrative cases:

The case:

> **The Sad Poor Island**: An island has a small rich upper class of very well-off people plus a large lower class of poor unhappy people. Then a flood comes and drowns the entire lower class, but this does not affect the upper class at all.

Is the Sad Poor Island better off before or after the flood? Saying this Sad Poor Island is *better off after* the flood seems reasonably intuitive because there is less sadness on the island after the flood than there was before the flood. This need not necessarily commit use to a controversial pro-euthanasia claim—improving the lives of the lower class obviously would have been *even better* than eliminating them—but it is merely committing us to saying that a lack of suffering is preferable to a lot of suffering and an acknowledgement of the less suffering after the flood.

The second case:

> **The Happy Poor Island**: An island has a small rich upper class of very well-off people plus a large lower class of poor but slightly happy people. Then a flood comes and drowns the entire lower class, but this does not affect the upper class at all.

Is the Happy Poor Island better off before the flood or after the flood? Saying that the island is better off *before the flood* seems reasonable because it takes seriously that the happiness (albeit slight) of the lower class makes the island better. Before the flood there are both more people and more happy people. Saying the island is *better off after* the flood seems unreasonable because it suggests that a bunch of slightly happy people make the island worse. After the flood there are both less people and less happy people.

So, returning to Arrhenius's proof, comparing these cases to step (4) of the proof, the counterintuitive Happy Poor Island case corresponds to a situation where $w_3 > 0$ whereas the more intuitive Sad Poor Island case corresponds to a situation where $0 > w_3$. Thus, because C_{m+q} is stipulated to have a *positive* welfare level (p. 262), the proof is applying the Happy Poor Island case. The problem is that, if you believe in step (4) in the proof, then you are committed to the smaller happy A_p population being better than the larger happy population $A_p \cup C_{m+q}$, which is tantamount to the controversial position that the Happy Poor Island is better *after the flood*.

We can attempt to fix step (4) by modifying the Anti-Egalitarian Conclusion to apply to both positive and negative inequalities (like the intuitive Sad Poor Island case):

> **The [Negative-Inclusive] Anti-Egalitarian Conclusion**: A population with perfect equality can be worse than a population with the same number of people, inequality [that can include populations of negative average welfare], and lower average (and thus lower total) positive welfare. (p. 258)

This modification better complies with our intuitions, but it also undermines the proof because the modification contradicts the proof's stipulation of positive welfare level $w_3 > 0$.

Comparing Cardinal Welfare Levels

Looking at the cardinal welfare level comparison version of step (4) can better reveal the nature of the problem and how it undermines the proof (p. 262):

$$w_5(p) \geq\ w_5(p) + w_3(m+q)$$

(Note: the welfare level w_5 has been assigned to specify the arbitrarily high positive welfare level of A_p)

Because $w_5(p)$ is on both sides it can cancel:

$$0 \geq\ 0 + w_3(m+q)$$

And because $m+q$ are positive integers they can disappear:

$$0 \geq w_3$$

However, by stipulation, w_3 is also a "very low positive welfare" (p. 262):

$$w_3 > 0$$

Therefore, our overspecification has resulted in contradiction:

$$0 \geq w_3 > 0$$

Thus, the definition of C_{m+q} as having a positive welfare level w_3 population appears to be impossible because the transitivity relation of step (4) introduces the same population A_p on both sides of the inequality, which means that the addition of population C_{m+q} must be worse than zero, which means that the welfare level $w_3(m+q)$ must be negative, but $m+q$ must be positive, which means that w_3 must be negative: a contradiction. By letting A_p be on both sides of the equation and assuming an added positive population C_{m+q}, the equation becomes overspecified, having more specifications than variables to specify.

For another illustration of just how counterintuitive this is, we can consider the following analogous case in geometry:

> **Self-Sizing Shapes**: consider a shape A that has an area greater than or equal to its own area plus the area of some other shape C.

Thinking about this problem, the initial response may be bafflement: how can an area be greater than or equal to itself plus something? Upon reflection though, we see that the only way for the comparisons of shape A to work is if shape C has a non-positive (zero or negative) area. We can imagine the solution by imagining taking a paper cutout of shape A and cutting the silhouette of shape C out of shape A.

The situation with the self-comparing total populations in step (4) is mathematically analogous to the situation with the self-sizing shapes. We can make the general stipulation that any shape/population comparison that is defined self-reflexively, which is to say that it includes itself in its comparison, may have to include the possibility of both additions and subtractions to be self-consistent.

Indeed, far from being a forced analogy, both Parfit (1984) and Arrhenius (2000) themselves (and others) use the convention of quadrilaterals of varying widths and heights to represent total welfares of varying population and average welfare level, so the analogy is apt. In fact, Arrhenius uses such quadrilaterals as visual aids in his own proof, and the glaringly counterintuitive nature of step (4) can be visualized by comparing the quadrilateral representations of A_p and $A_p \cup C_{m+q}$ and wondering how the former could be better than the latter (Arrhenius, 2000, p. 262).

Consequences:

Step (4) could succeed if A_p were compared to some other high welfare level population, not A_p but some population Z_p, let's say, plus an added population C_{m+q}, thus avoiding the strange self-comparison of step (4). But population A_p not Z_p is used in order to fit the transitivity relation in step (6), so that the rest of the proof can follow; using population Z_p would not satisfy the transitivity relation of step (6), which would stop the proof short.

Step (4) is a necessary condition of step (6), which is a step in the subproof steps (5-7) that ends up proving a subconclusion in step (8). However, because it happens that w_3 must be negative, the conclusion of the subproof in step (8) follows from $B_{q+1} \geq C_{m+q}$ because $w_4(q+1) > 0 > w_3(m+q)$ must be true. Thus, the problems arising from the application of the "No Repugnant/Anti-Egalitarian Conclusion" criteria in step (4) does not directly undermine the proof here. This also demonstrates how we arrived at the problem in step (7) in the first place, because we counterintuitively arrived at a bad addition of positive welfare in step (4) and then applied it to step (7), a move that would not have seemed possible but for arriving at step (4) through step (2) and step (3).

However, because w_3 must be negative, the series of welfare levels ($w_4 > w_3 > w_2$) is broken into a disjunction ($w_4 > 0 > w_3 > w_2 \cup w_4 > w_2 > 0 > w_3$) and cannot be used non-disjunctively in any further steps in the proof. We can no longer assume that w_3 is greater than w_2; we must instead consider the disjunction that w_3 may be more than, or less than, or equal to w_2. This consideration affects step (13) indirectly because that step requires the Dominance Principle, which requires that the welfare level of one population be greater than the welfare level of another equally sized population. Because we cannot assume that w_3 is greater than w_2, we cannot unambiguously use the Dominance Principle in step (13). We instead can consider the two disjuncts, 1) where w_3 is greater than w_2 and 2) where w_2 is greater than or equal to w_3. With disjunct 1, the Dominance Principle in step (13) succeeds, step (14) follows and with step (11) forms the contradiction of the Impossibility Theorem. However, with disjunct 2, the Dominance Principle in step (13) fails, step (14) does not follow and instead what follows is an alternative to the Impossibility Theorem. By disjunctive syllogism, if disjunct 1 is shown to lead to impossibility, then we can conclude that disjunct 2 follows. So, ceteris paribus, the application of the "No Repugnant/Anti-Egalitarian Conclusion" Criteria in step (4) and its consequences for the "Dominance Principle" in step (13) undermines the impossibility theorem.

Addition Principle too Broad?

The second problem is that step (7) of the proof makes a broad application of the "Addition Principle" to positive welfare level cases, a principle which says that:

> **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. (p. 257)

It seems like principle like it should be rejected in this case or amended to apply only to certain limited negative welfare level cases.

Comparing Ordinal Betterness

Using an ordinal betterness comparison version of the argument, deriving step (7) follows from our above derivation of step (4), plus a few more straightforward steps, but step (7) is counterintuitive.

In step (5), we simply assume a betterness relation so that we may contradict it in step (7) (p. 262):

$$A_p \cup C_{m+q} > A_p \cup B_{q+1}$$

In step (6), we apply the principle of transitivity to step (4) and step (5) (p. 262):

$$A_p > A_p \cup B_{q+1}$$

In step (7), we apply the Principle of Addition to step (6) (p. 262):

$$A_p \cup B_{q+1} \geq A_p \cup C_{m+q}$$

However, we can intuit that something may be amiss when we arrive at step (7). In step (7) both B_{q+1} and C_{m+q} are positive populations of $q+1$ and $m+q$ persons respectively, where $q+1 < m+q$, which seems to meet the condition of the Addition Principle that the worse added population be of "greater number"; and, both B_{q+1} and C_{m+q} have positive average welfare levels of w_4 and w_3 respectively, and $w_4 > w_3$, which seems to meet the condition of the Addition Principle that the worse added population have an "even lower welfare".

However, upon further consideration, applying the Addition principle in this case seems counterintuitive. The Addition Principle seems ambiguous regarding the negativity or positivity of the welfare level of the added population and seems to be alternately to apply or not apply depending upon the sign of that welfare level. We can realize this by considering the following two pairs of cases:

First pair of cases:

> **Solely Sad Babies**: Consider one scenario of birthing 2 babies each living 2 solely sad days before dying; and another scenario of birthing just 1 baby that lives just 1 slightly less sad day before dying.

Which scenario is better? It seems somewhat intuitive to say that the second case is better than the first because there is less suffering for less people.

Since the welfare level of the added babies is negative, the Addition Principle seems to apply because adding solely sad babies seems intuitively bad. So, if it is bad to add babies of a negative welfare, then it is *surely even worse* to add even more babies of an even lower negative welfare, ceteris paribus, per the Addition Principle. In other words, for added populations of negative welfare levels the Addition Principle seems intuitive to apply.

Second Case:

> **Barely Happy Babies**: Consider one scenario of birthing 2 babies that live for 2 barely happy days before dying; and another scenario of birthing just 1 baby that lives for just 1 slightly happier day before dying.

Which scenario is better? It seems counterintuitive to say that one scenario is better/worse than the other; rather, it is intuitive to say that neither case is obviously better or worse than the other without further information about the relative happiness levels of the babies in question.

Since the welfare level of the added babies is positive (not negative as in the previous case), the Addition Principle seems to be prima facie not apply because adding barely happy babies is not intuitively bad. Indeed, it is not intuitively bad to add any populations of positive welfare level ceteris paribus. This does not make the Addition Principle generally false, but it means that there may never be bad additions of positive welfare populations that satisfy it, making it only apply to additions of negative welfare populations.

Returning to Arrhenius's proof, comparing these cases to step (7) of the proof, the intuitive Solely Sad Babies case corresponds to a situation where $0 > w_3$ whereas the counterintuitive Barely Happy Babies case corresponds to a situation where $w_3 > 0$. Thus, because B_{q+1} and C_{m+q} are stipulated to have *positive* welfare levels (p. 262), the proof is applying the Barely Happy Babies case. The problem is that, if you believe in step (7) in the proof, then you are committed to applying the Addition Principle to cases of positive welfare additions, which is tantamount to the controversial position that *more* Barely Happy Babies *is worse.*

We can modify the Addition Principle to be more specific to apply just to negative additions (just to Solely Sad Baby cases) that are consistent with our intuitions:

> **The [Negative-Exclusive] Addition Principle**: If it is bad to add a number of people, all with [negative] 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. (p. 257)

This modified principle makes clear our intuition that positive welfare cases should not necessarily be bad to add ceteris paribus, and thus it better complies with our stated intuitions, but it also undermines the proof because the modification contradicts the proof's stipulation of positive welfare level $w_3 > 0$.

Comparing Cardinal Welfare Levels

Looking at the cardinal welfare level comparison version of the application of the Addition Principle in step (7) can better reveal the possibility of this balancing point and how it undermines the proof:

$$w_5(p) + w_4(q+1) \geq w_5(p) + w_3(m+q)$$

It stands to reason that, if both added welfare levels, w_4 and w_3, are positive, it is possible that there is some m of such magnitude that the inequality is broken. If:

$$m > \frac{w_4}{w_3}(q+1) - q$$

Then, the Addition Principle is violated. Because such a value of m seems entirely possible given positive welfare levels, the Addition Principle does not seem like it should hold for positive additional populations.

For step (7) of the proof, the original and added populations in question are stipulated to be of a low positive welfare level, and thus the application of the Addition Principle seems counterintuitive (like the Barely Happy Babies, not like the Solely Sad Babies). Thus, step (7) is only valid if we give some account of why the Addition Principle applies in positive added welfare level cases.

For another illustration of just how counterintuitive this is, we can consider the following analogous case in geometry:

> **Relative Shapes**: consider a rectangle B that has a width of w_B and a length of l_B and rectangle C that has a width of $w_B - x$ and a length of $l_B + y$.

Can we say, without further information, which shape has the larger area? Thinking about this problem, the obvious answer should be that the answer entirely depends upon the values of x and y. It would be premature to assign relative are until then.

The situation with the relative total populations in step (7) is mathematically analogous to the situation with the relative shapes. We cannot make the general stipulation that any shape/population comparison without fully specifying the variables on which the area depends.

Consequences:

Step (7) is a necessary condition of a subproof steps (5-7) that ends up proving a subconclusion in step (8), and step (8) is one of the necessary conditions for step (14), and step (14) combined with step (11) is one of the two final conditions in the contradiction that proves the Impossibility Theorem. So, ceteris paribus, the broad application of the Addition Principle to step (7) undermines the proof.

Conclusions

So, two loopholes have been discovered and delineated in Arrhenius's proof of the Impossibility Theorem for Welfarist Axiology. Notably, the "Addition Principle" loophole does not appear fatal to the proof's conclusion, though the "Dominance Principle" loophole appears fatal to the proof's conclusion.

The loophole caused by the "Addition Principle" in step (7) is not necessarily fatal to the conclusion because it affects step (7), which is a premise in the subproof (5-7) that proves step (8). However, as it turns out, step (7) is not necessary because we end up being able to prove step (8) by a different path. This different path is the discovery that w_3 must be intuitively assumed negative via step (4), and thus step (8) is true and proven by step (4) without the consequent subproof (5-7). However, while not fatal itself, this problem has the unfortunate side effect of obscuring the importance of the signs of the welfare level, making other problems harder to spot.

The loophole in step (4) caused by the "No Repugnant/Anti-Egalitarian Conclusion" Criteria also only directly affects the subproof (5-7), but it is fatal to the conclusion indirectly anyways because it contradicts the assumed relative ordering of welfare levels ($w_4 > w_3 > w_2$), which is necessary to assume the "Dominance Principle" in step (13). This invalidity of the Dominance Principle permits a disjunction by which the Impossibility Theorem is avoided. This loophole is threefold tricky to uncover: 1) because assuming the relative ordering of welfare levels is not a formally stated step in the proof, nor an assumed principle, its violation in step (4) is easy to overlook; 2) because the consequences of violating the relative ordering of welfare levels does not come until step (13) it is doubly easy to overlook; and 3) because the "Addition Principle" also ambiguously underspecifies the importance of the negativity of the welfare level, the possibility of negative welfare levels is muffled by step (7).

Notably, these critiques are illuminated by thought experiments in terms of an analysis of ordinal values, but they are even more clearly illuminated by a cardinal comparison of additive average level welfares multiplied by their populations into total welfares, enabled by the appropriate assumptions (Neumann, Morgenstern, 1944). This is perhaps a hint that such a granular cardinal utility analysis, as opposed to a preference-ordering-based ordinal analysis, is methodologically helpful for solving welfare axiology. Using such analysis we have shown that the Impossibility Theorem is only valid for certain utility functions, and in fact is invalid for the most theoretically important welfarist principle, Parfit's Impersonal Total Beneficence Principle (Parfit, 1984, p. 69, p. 387), as well as the most historically important cardinal utility function, that of classical utilitarianism (Bentham, 1789).

As a consequence, the Impossibility Theorem for Welfarist Axiology seems to be undermined, chiefly because it could not specify a population that met the "No Repugnant/Anti-Egalitarian Conclusion" criteria that was consistent with the assumed domain of positive welfare levels, making the keystone application of the "Dominance Principle" invalid. To the extent that this loophole exists in other Impossibility Theorems of the same kind, those theorems should fail as well, though other theorems could perhaps also survive the above method; indeed, Arrhenius presents a different version of his impossibility theorem that we have not touched upon here, but it follows a roughly similar proof, and it is thus expected to fail for similar reasons (Arrhenius, 2009, 2011a, 2011b). Of course, if any impossibility theorem does succeed, then the options available seem largely unsatisfactory (Temkin, 2012, pp. 9-10). This critique does not provide a positive solution of population ethics, but the proof makes more options possible (Ryberg & Tännsjö, 2004; Arrhenius et al., 2022). So, we can accept these principles (at least) without abandoning hope in a self-consistent welfarist ethic.

Acknowledgements

Special thanks to Dr. Bruce Gordon at Houston Baptist University for guidance on formal logic and Dr. David Phillips for guidance on welfare ethics (and both for feedback on early versions of this project).

References

Arrhenius, G., 2000a, "An Impossibility Theorem for Welfarist Axiology", *Economics and Philosophy*, 16: 247–266.

Arrhenius, G. (2009). "One more axiological impossibility theorem." In L. G. Johansson, J. Österberg, & R. Sliwinski (Eds.), *Logic, Ethics and All That Jazz: Essays in Honour of Jordan Howard Sobel*, 23–37. Uppsala Philosophical Studies.

Arrhenius, G., (2011a). "The Impossibility of a Satisfactory Population Ethics" in H. Colonius & E. Dzhafarov (eds.) 2011. https://link.springer.com/article/10.1007/s11098-021-01621-4

Arrhenius, G. (2011b). The impossibility of a satisfactory population ethics. In E. N. Dzhafarov & L. Perry (Eds.), *Descriptive and Normative Approaches to Human Behavior*, 1–26. World Scientific Publishing Company.

Arrhenius, Gustaf, Jesper Ryberg, and Torbjörn Tännsjö, "The Repugnant Conclusion", *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/win2022/entries/repugnant-conclusion/>.

Bentham, Jeremy, 1789. *An Introduction to the Principles of Morals and Legislation.*, Oxford: Clarendon Press, 1907.

List, Christian, "Social Choice Theory", *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/win2022/entries/social-choice/>.

von Neumann, J., and Morgenstern, O., 1944, *Theory of Games and Economic Behavior*, Princeton: Princeton University Press.

Parfit, Derek, *Reasons and Persons*, Oxford: Clarendon Press, (Oxford, 1984). p. 390.

Ryberg, J. and T. Tännsjö, 2004 (eds.), *The Repugnant Conclusion. Essays on Population Ethics*, Dordrecht: Kluwer Academic Publishers.

Temkin, L.S., 2012, *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*, New York: Oxford University Press.

Appendix: A Reconstruction of Arrhenius's Original Argument

We have here provided a reconstruction of Arrhenius's original proof of the Welfarist Impossibility Theorem along with the text of the proof as it is stated in his original article, including all assumptions and definitions (Arrhenius, 2000, pp. 261-263). The reconstructed proof supplied here is meant as a supplement to the original text, not a replacement of it, so the full text is not supplied here.

Given Definitions

Given that:

$$m, p, q > 0$$

$$w_5, w'_5 \gg w_4 > w_3 > w_2 > w_1 > 0 > w_0$$

$$A_p, A'_q \gg B_{q+1}, C_{m+q}, D_{m+p+q}, F_m, G_{m+p+q} > 0 > E_1$$

Where m, p, q are real number populations. And where levels w_4 through w_1 are "four very low positive [average] welfare levels" (Arrhenius, 2000, p. 261). And where the capital letters represent populations with differing persons and welfare levels.

The levels w_5 and w'_5 are added to represent the average welfare level of the population A_p and A'_p. The level w_0 is added to represent the average welfare level of the population of E_1.

Note: in our version of the article there is an apparent typo "$A_p C_{m+q}$" which we have assumed to be "$A_p \cup C_{m+q}$" in order to make sense of the sequence of the argument (p. 262).

The populations are also more formally defined:

Population Definitions per Arrhenius, 2000 pp. 261-263
Ap: A population with p members with very high welfare.
Bq+1: A population with q+1 members with very low positive welfare w4.
Cm+q: A population with m+q members, m > 2, with very low positive welfare w3 such that the average welfare of Ap ∪ Cm+q is less than w4, that is, [v(a1) +. . .v(ap) + v(c1). . .+ v(cm+q)]/(m+p+q) < w4.
Dm+p+q: A population of the same size as Ap ∪ Cm+q with very low positive welfare w4.
A'q: A population with q members with very high welfare.
E1: One person with slightly negative welfare.
Fm: A large population with very low positive welfare w1 such that the average welfare of Ap ∪ A'q ∪ Fm is less than w2, that is, [v(a1)+. . . v(ap)+. . . v(a'1)+. . . v(a'q) + v(f1)+. . . v(fm)]/(m+p+q) < w2.
Gm+p+q: A population of the same size as Ap ∪ A'q ∪ Fm with very low positive welfare w2.

Given Assumptions

Four principles must be stated that go into our critique of the proof of the Impossibility Theorem:

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. (p. 257)

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 (p. 259)

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. (p. 251)

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. (p. 248)

In addition to the four principles provided above, we also need two additional principles for the proof to work, plus the Impossibility Theorem itself:

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. (p. 258)

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. (p. 257)

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. (p. 261)

The principles are also more formally defined:

Principles Formalized by Arrhenius (pp. 261)
The Dominance Condition: If ai has higher welfare than bj for all ai ∈ An, Bj ∈ Bn, then An is better than Bn.
The Addition Principle: If ai has higher welfare than bj, and bj has higher welfare than ch, for all ai ∈ Ak, bj ∈ Bn, ch ∈ Cm, and Ak is better than Ak ∪ Bn, and m > n, then Ak ∪ Bn is at least as good as Ak ∪ Cm.
The Anti-Egalitarian Conclusion: There are populations An and Bn such that for all ai, aj ∈ An, ai has the same welfare as aj, but for some bi, bj ∈ Bn, bi has lower welfare than bj, and [v(a1)+...v(an)]/n > [v(b1)+...v(bn)]/n, and Bn is better than An.
The Minimal Non-Extreme Priority Principle: There is an n such that if An ⊂ Wvhp, B1 ⊂ Wsn, Cn+1 ⊂ Wvlp, then An ∪ B1 ∪ Dk is at least as good as Cn+1 ∪ Dk, k > 0.
The Repugnant Conclusion: For any An ⊂ Wvhp such that ai has the same welfare as aj for all ai, aj ∈ An, there is a Bm ⊂ Wvlp such that Bm is better than An.
The Sadistic Conclusion: There are populations An, Bm, Ck such that An ⊂ Wpw, Bm ⊂ Wnw, n, m > 0, k ⊂ 0, and Bm ∪ Ck is better than An ∪ Ck.
Where W are defined sets: Wpw is positive, Wnw is negative, Wvhp is very high positive, Wvlp is very low positive, and Wsn is somewhat negative welfare level.

Stated Proof

	Stated Steps per Arrhenius, 2000 pp. 262-263
1	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) Ap be such a population.
2	Since Dm+p+q is a population with very low welfare, avoidance of the Repugnant Conclusion yields that (2) Ap is at least as good as Dm+p+q.
3	Avoidance of the Anti-Egalitarian Conclusion yields that (3) Dm+p+q is at least as good as ApCm+q.
4	By transitivity, it follows from (2) and (3) that (4) Ap is at least as good as Ap ∪ Cm+q.
5	Assume that (5) Ap ∪ Bq+1 is worse than Ap ∪ Cm+q.
6	By transitivity, it follows from (4) and (5) that (6) Ap ∪ Bq+1 is worse than Ap.
7	Since m > 2, it follows from (6) and the Addition Principle that (7) Ap ∪ Bq+1 is at least as good as Ap ∪ Cm+q which contradicts (5).
8	Hence, if we assume that Ap ∪ Bq+1 is worse than Ap ∪ Cm+q, then we get a contradiction. Thus, (8) Ap ∪ Bq+1 is at least as good as Ap ∪ Cm+q. Q.E.D.
9	Avoidance of the Sadistic Conclusion yields that (9) Ap ∪ A'q ∪ Fm is at least as good as Ap ∪ A'q ∪ E1.
10	Since Ap ∪ A'q ∪ Fm is an anti-egalitarian alternative relative to Gm+p+q, (10) the latter population is at least as good as the former.
11	By transitivity, (9) and (10), it follows that (11) Gm+p+q is at least as good as Ap ∪ A'q ∪ E1.
12	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) Ap ∪ A'q ∪ E1 is at least as good as Ap ∪ Bq+1 (see Diagram 6).
13	The Dominance Principle yields that (13) Gm+p+q is worse than Ap ∪ Cm+q. From (8) above, we know that Ap ∪ Bq+1 is at least as good as Ap ∪ Cm+q.
14	By transitivity, it follows from (8), (12), and (13) that (14) Gm+p+q is worse than Ap ∪ A'q ∪ E1 which contradicts (11).
15	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.

Formalized Proof

	Ordinal Population Betterness Comparison	**Cardinal Welfare Level Comparison**	**Step**
1	A_p	$w_5(p)$	Assumption
2	$A_p \geq D_{m+p+q}$	$w_5(p) \geq w_4(m+p+q)$	No Repugnant C.
3	$D_{m+p+q} \geq A_p \cup C_{m+q}$	$w_4(m+p+q) \geq w_5(p) + w_3(m+q)$	No Anti-Egalitarian C.
4	$A_p \geq A_p \cup C_{m+q}$	$w_5(p) \geq w_5(p) + w_3(m+q)$	Transitivity 1+2+3
5	$A_p \cup C_{m+q} > A_p \cup B_{q+1}$	$w_5(p) + w_3(m+q) > w_5(p) + w_4(q+1)$	Assumption
6	$A_p > A_p \cup B_{q+1}$	$w_5(p) > w_5(p) + w_4(q+1)$	Transitivity 4+5
7	$A_p \cup B_{q+1} \geq A_p \cup C_{m+q}$	$w_5(p) + w_4(q+1) \geq w_5(p) + w_3(m+q)$	Addition P. +6
8	$A_p \cup B_{q+1} \geq A_p \cup C_{m+q}$	$w_5(p) + w_4(q+1) \geq w_5(p) + w_3(m+q)$	Contradiction 5+7
9	$A_p \cup A'_p \cup F_m \geq A_p \cup A'_p \cup E_1$	$w_5(p) + w'_5(q) + w_1 m \geq w_5(p) + w'_5(q) + w_0$	No Sadistic C.
10	$G_{m+p+q} \geq A_p \cup A'_p \cup F_m$	$w_2(m+p+q) \geq w_5(p) + w'_5(q) + w_1 m$	No Anti-Egalitarian C.
11	$G_{m+p+q} \geq A_p \cup A'_p \cup E_1$	$w_2(m+p+q) \geq w_5(p) + w'_5(q) + w_0$	Transitivity 9+10
12	$A_p \cup A'_p \cup E_1 \geq A_p \cup B_{q+1}$	$w_5(p) + w'_5(q) + w_0 \geq w_5(p) + w_4(q+1)$	Min-Non-Ex-Priority P.
13	$A_p \cup C_{m+q} > G_{m+p+q}$	$w_5(p) + w_3(m+q) > w_2(m+p+q)$	Dominance P.
14	$A_p \cup A'_p \cup E_1 > G_{m+p+q}$	$w_5(p) + w'_5(q) + w_0 > w_2(m+p+q)$	Transitivity 8+12+13
15			Contradiction 11+14