Krister Bykvist[1]

Escaping the Impossibility Theorems in Population Ethics[2]

Decision-makers are in a hurry to find morally justified responses to climate change. Population ethicists have thrown a spanner in the works by formulating various impossibility theorems that show that no theory about the value of population change can satisfy all the conditions we think such a theory must satisfy. What shall we do, if we do not know which condition(s) to give up? One relatively unexplored option is to view the satisfaction of a condition as a matter of degree, as Geoff Brennan recently has suggested (in the context of Arrow's impossibility theorem). This opens up the possibility that some theories might overall come closer to full satisfaction of the conditions than others. In my paper, I shall explore various versions of this idea and see how far they will take us. In particular, I will make use of the famous Kemeny-measure of distance and show that this will rule out all population theories that are indifferent between some of the alternative populations in the Mere Addition Paradox. I will also discuss factors beyond distance that are relevant for theory choice.

[1] Dept. of Philosophy, Stockholm University, Institute for Futures Studies, Stockholm

[2] For very helpful feedback, I would like to thank the audience at the internal work-in-progress workshop of the Climate Ethics program held in Spring 2023, especially Gustaf Arrhenius, Tim Campbell, Hilary Greaves, Daniel Ramöller, and Joe Roussos.

1. Introduction

Decision-makers are in a hurry to find morally justified responses to climate change and other urgent issues that involve decisions that will have effects on future populations. But the population ethicists have not been especially helpful. We have thrown a spanner in the works by formulating various impossibility theorems that show that there is no acceptable reaction to climate change if we take into account the value of population change. More precisely, these theorems show that there is no theory about the value of population change that satisfies a set of very plausible conditions we are inclined to think a theory should satisfy.[3] Given that no theory can satisfy all of these conditions, what shall we do?

The main options are to

(1) 'put your hands up in the air': utter despair and moral paralysis, for population ethics is doomed to be inconsistent.[4]

(2) 'not care': argue that it is a mistake to think the impossibility theorems in population ethics are relevant for moral justification;

(3) 'drop a condition': sit down, do some serious philosophical reflection, and try (again) to work out which condition(s) to drop;

(4) 'hedge': keep all the different theories on the table, assign credences to them, compare the values the theories assign to populations, and, in analogy with what we should do under empirical uncertainty, apply some suitable decision-theoretic principle.[5]

(5) 'think that a miss is *not* as good as a mile': instead of just judging whether a theory satisfies or fails to satisfy a certain condition, we can see whether it gets *closer* or further *away* from satisfying the condition, as Geoff Brennan recently has suggested in the context of Arrow's impossibility theorem (Brennan 2015, see also Brennan and Braurmann 2006). Moreover, even if no theory can satisfy all of the conditions, some might come overall *closer* to satisfying them than others. Hopefully, we could seek guidance from the theories that rank higher.

In this paper, I shall explore various versions of the closeness approach and see how far they will take us. To say that one will *explore* something is a philosopher's jargon

[3] See, for instance, one of the leading spanner throwers Arrhenius (forthcoming). For an impossibility theorem in a probabilistic setting, see Arrhenius ibid.

[4] That population ethics is inconsistent is seriously considered in Arrhenius (forthcoming).

[5] For an example of this approach, see Bykvist (2022) and Ord & Greaves (2017).

for saying that one has not yet made up one's mind about the issues, or failed to reach a conclusion. As you will see, I am not sure that the closeness idea can take us far enough. I have excuses for this undecidedness. The issue I am going to discuss involves a lot of uncharted terrain, and the issues are very complex. But I hope this exercise in 'axiological escapology', as we may call it, still can teach us something important, and that it is not just a failed escape act from the chains of the impossibility theorems.

Before I explore the closeness approach, I will introduce the impossibility theorems in population ethics (one simple version, there are many others!), and say a few words about the other alternative reactions to impossibility theorems and why it is worth exploring the closeness approach.

2. Impossibility theorems in population ethics

In general, to show an impossibility theorem is to collect a set of intuitively plausible conditions on a certain kind of theory and prove that they are logically inconsistent. This is what Arrow did for theories of social choice (with interpersonally incomparable ordinal preferences), and this is also what is done in population ethics for theories about how we should value populations. Now, there are many different impossibility theorems in population ethics. Here I will only present a very simple version, since it is easier to work with; it can be discussed informally without getting into technical details. It should be noted that the conditions of this version are not as plausible as the ones of the more complex formal ones.[6]

The conditions for what is often called the Mere Addition Paradox can be stated informally as follows.

> *Mere Addition*, a population that differs from another only in that it contains some extra lives all worth living is *at least as good* as the smaller population.
>
> *Non-Anti Egalitarianism (NAE)*: a same-sized population with both greater total and average wellbeing, distributed perfectly equally, is better.
>
> *Avoidance of the Repugnant Conclusion (Avoidance of RC)*: a vast population with lives barely worth living is worse than a much smaller population with lives of very high wellbeing.[7]

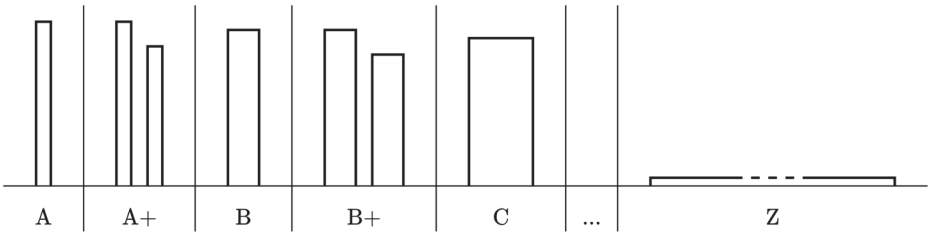
[6] For more plausible and formally developed theorems, see Arrhenius (forthcoming) and Blackorby et al. (2005).

[7] Strictly speaking, a theory avoids RC when it states that a vast population with lives barely worth living is not better than a much smaller population with lives of very high wellbeing. But the stronger

The fudge words 'vast', 'much smaller', 'barely worth living', and 'very high' can be avoided in the more formally precise statements of the conditions.

Here is an illustration of the impossibility of satisfying all the conditions above.

Fig. 1



A is a population with lives of very high wellbeing. By Mere Addition: A+ is at least as good as A. By NAE, B is better than A+. This implies that B is as least as good as A, by transitivity of at least as good as. Now repeat this for B, B+, and C and so on until you reach Z, a vast population with lives barely worth living, and we can conclude that Z is at least as good as A. But this contradicts the claim, stated by Avoidance of RC, that A is better than Z.

Strictly speaking, the conditions that generate the impossibility should include these background conditions:

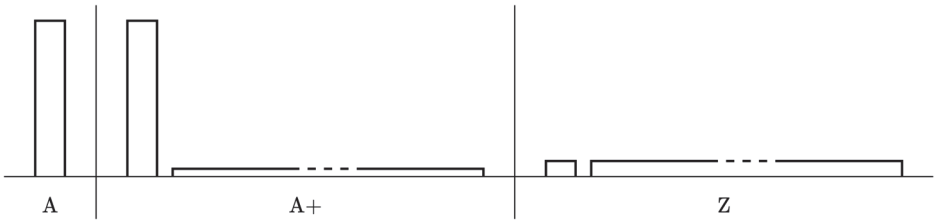
Transitivity: of at least as good as;
Measurability: assumptions about the structure and measurability of wellbeing that make it possible to construct a sequence as the one above and to talk about total, average, and equal wellbeing;
and
Domain Richness: there are possible populations like A, A+, B, B+, ..., and Z.[8]

A much shorter version of the paradox is this.

condition in the text is usually what motivates people to demand that a theory avoids RC.

[8] The strong measurability assumptions can be avoided. See Arrhenius (forthcoming).

Fig. 2



However, one could object here that the move from A to A+ is very problematic because it creates an enormous amount of inequality, or that the move from A+ to Z implausibly requires us to pull down the better off people to the level of barely worth living. However, each move in the long version can be justified by invoking weaker principles than the ones listed above (and relaxing the measurability assumption). But I will work with the short version to avoid unnecessary complexity.

Another simplifying assumption is that I shall work with a stronger version of Mere Addition, according to which population A+ is always *better* than population A. Furthermore, I shall put aside the conditions *Transitivity*, *Measurability*, and *Domain Richness*. I shall also assume that the theories satisfy full comparability (i.e., there are no gaps in the ordering). Finally, I shall assume that the set of alternatives to be compared (the set of populations) are finite, as well as the set of people in each population. I will come back to some of these simplifications later.

3. Reactions to the impossibility theorem

Of the listed options, I think we should not 'throw up our hands in the air' unless we have shown that none of the others work.

The option of not caring is worth considering, since some might argue that there is something deeply mistaken about the framework. Why should we think that the value of populations matters, and why think that it is a function of the wellbeing of its members? This smacks of old-fashioned utilitarianism. Why should non-utilitarians care about this? However, this concern isn't exclusive to utilitarians; everyone should care about how our current actions affect the wellbeing of future people. One should not be completely indifferent to the possibility of making future lives miserable, or barely worth living. Of course, this is not to deny that many other things matter as well, such as our rights and duties to contemporaries, but one factor to take into account is the wellbeing of future people. If all else is equal, this is the factor that determines what we should do.

The idea that only individual wellbeing matters for the value of a population can

also be relaxed. For example, the value of a life can be in part determined by moral, artistic, and athletic achievements. Finally, we can even state the impossibility theorems directly in terms of reasons or obligations to bring about various population changes, without invoking the value of populations or the value of outcomes.[9] But, again, for simplicity, I will stick to the wellbeing framework in this paper.

The ideal option is of course to drop a condition, if we know which one to drop. The problem is that we do not know, or many of us do not. Despite extensive philosophical reflection on these issues, there is still wide-spread disagreement and undecidedness among population ethicists. Since climate change requires immediate action, we need some guidance on what to do now, even though we are undecided about which condition to drop.

Hedging could be an option here (which I have explored in Bykvist (2022)). The idea is to view the choice situation like this, where numbers represent the value of populations A, A+, and Z according to some theories, T1, T2, and T3:

Fig. 3

Alternatives	**T1** p1	**T2** p2	**T3** p3
A	1	1	1
A+	2	-1	2
Z	3	0	0

Each theory considered satisfies two out of the three conditions. T1 satisfies Mere Addition and NAE, but not Avoidance of RC. T2 satisfies NAE and Avoidance of RC, but not Mere Addition. T3 satisfies Mere Addition and Avoidance of RC, but not NAE.

To decide which population to bring about, we need to somehow weigh the probabilities (credences for the different theories) p1, p2, and p3, against the values assigned to the populations by the theories T1, T2, and T3. One major challenge for this approach is to show that it makes sense to compare values across *different* theories. While I think it does make sense in some cases, this is controversial.[10] Thus, it is worthwhile to explore the last option.

[9] For a deontic impossibility theorem that only invokes 'ought' and 'permissible', see Arrhenius (2021).

[10] For a critical discussion of some existing proposals and a defence of a new one, see MacAskill et al (2020) and Riedener (2021).

4. Satisfying a condition is not an *all or nothing* affair

The guiding idea of this approach is that instead of just judging that a theory satisfies or fails to satisfy a certain condition, we can say that a theory gets closer or further away from satisfying the condition (for short, 'closer or further away from the condition'). Moreover, even if no theory can satisfy all of the conditions, some might come overall closer to satisfying them than others. This degree of closeness can be understood in different ways, but a plausible closeness account must validate:

> *Closeness Dominance*
> If, for every condition C, T1 is at least as close to C as T2 is, and for some condition C', T1 is closer to C' than T2 is, then T1 is closer overall to satisfying the conditions than T2 is.
>
> *Equal Closeness*
> If, for every condition C, T1 is exactly as close to C as T2 is, then T1 is exactly as close to all conditions as T2 is.
>
> *Inclusion*
> If T1's C-violations are a proper subset of T2's C-violations, then T1 is closer to C than T2 is.
>
> *Identity*
> If T1's C-violations are exactly the same as T2's C-violations, then T1 is exactly as close to C as T2 is.

While these principles have some applicability, but the first two require closeness comparisons between different theories regarding a certain condition. None of them requires closeness comparisons across conditions, i.e., that one theory is closer to a certain condition than another theory is to another condition. But this also shows its limitations. Ideally, we would like to make overall closeness comparisons when theories differ in how close they are to a whole set of conditions.

I shall consider three accounts of closeness: a value-based approach (defended by Brennan (2015)), a proportion-based approach, and a ranking-distance approach —which is the one I will end up favouring if combined with a proportion-based approach. To simplify the discussion, I will assume that all conditions have the same weight. This is unrealistic, since we might have more confidence in some conditions than in others. In section 7, I will briefly discuss the significance of dropping this idealization.

5. Value-based approach

To explain the motivation behind his value-based approach, Brennan usefully invokes an analogy with bananas and apples. Suppose you wish to eat 10 bananas and 7 apples a week, but you can't afford this fruit consumption. You should not declare yourself an *apple person* or a *banana person* and forget about the other fruit. You should trade off the fruits so that you get an ideal combination of apples and bananas, which normally means that you will give up some of both. Brennan suggests something similar for impossibility theorems. When you realize that not all conditions can be jointly satisfied by a theory, you should not just go for some conditions and forget about the others. You should trade off some conditions against others until you find a theory that is best in terms of an ideal trade-off between the different conditions.

How does Brennan's approach work more exactly? First, we need to identify for each condition 'the underlying value' that this condition is 'supposed to promote' (Brennan 2015). Then, we form a metric that 'reflects the degree to which a procedure fails' to meet the condition. A theory fails to meet a condition when the theory promotes the underlying value below a certain *threshold* level.

This suggests that a theory's closeness to a condition is the difference between the amount of value 'promoted' by the theory and the threshold of value set by the condition. A theory's closeness to the set of conditions is then some strictly decreasing function of all the value differences between the theory and the conditions.

Since Brennan talks about Arrow's Impossibility Theorem and Sen's Liberal Paradox, the conditions he has in mind are: *Universal Domain*, *Independence of Irrelevant Alternatives*, *Pareto*, *Non-Dictatorship*, *Transitivity of 'at least as good as'*, and *Minimal Liberty*. Brennan concedes that developing a metric for each of these conditions is a great challenge. But he suggests that for some of the conditions it is pretty easy. For example, he claims that we can measure how well a theory does in terms of Universal Domain by the proportion of possible individual rankings that have to be ruled out. Furthermore, when he considers the Pareto-principle, he suggests that the value it promotes is preference satisfaction and that the metric should be defined in terms of *distance from a Pareto-optimal frontier* (the set of Pareto-optimal social states).

This measure should, with suitable constraints on the underlying values, be able to satisfy the general principles: Closeness Dominance, Closeness Equality, Inclusion, Identity. But there are some problems with the account, especially if we want to apply it to the Mere Addition Paradox.

First, it seems very questionable that each of these conditions has a *unique* underlying value that is supposed to be promoted. Putting aside *Transitivity* and

Domain Richness, which might be exceptions, what is the *unique* promotion-worthy value underlying the *Mere Addition Principle*, the *Avoidance of RC*, and *NAE*, respectively? Each of these conditions can be accepted for a variety of reasons, and from very different evaluative standpoints. That is especially clear for *NAE*, which can be accepted by pure egalitarians, total utilitarians, average utilitarians, and leximiners. But it is also clear that the Mere Addition Principle can be accepted by total utilitarians and person-affecting views, and the *Avoidance of RC* can be accepted by average utilitarians, critical level utilitarians, leximiners, and various perfectionist theories. Indeed, that a condition can be accepted by very different moral outlooks is one of the main reasons why we assume it in the first place, since a condition that only a few outlooks would accept can more easily be rejected.

Second, even if we assume that there are values underlying each condition, why assume that there is a threshold for each of these values? And if there is threshold, how do we decide where it is?

Finally, and more importantly, Brennan asks us to assess theories according to how well they trade off the values underlying the conditions. But this is odd, since the conditions were supposed to constrain value trade-offs. For example, to accept the Avoidance of RC is to accept that no number of barely worth living people can together be more valuable than a smaller number of very well-off people. So, when Brennan asks us to judge theories according to how well they trade off various values, we seem to be back to where we started. We have a set of values and we want to know how to aggregate them. For instance, we want to know how to weigh quality of wellbeing against quantity of wellbeing. The impossibility theorems were generated by listing all plausible conditions on such trade-offs. Unless Brennan can show us which condition(s) to drop, we have not escaped the impossibility theorems.

6. The proportion-based approach

According to this approach, a theory's closeness to a condition C is identified with the proportion of its C-violations. The greater proportion of C-violations a theory has, the further away the theory is from condition C. (The account could of course be restated in terms of proportions of satisfactions of a condition.)[11]

Overall closeness is then some strictly decreasing function of the closeness measures for each condition. For example, if we can measure the exact proportion of violations for each condition, we can average these measures to get the overall closeness to the set of all conditions.

[11] A similar approach has been defended by Campbell and Kelly (1994), who construct a measure of degrees of Non-Dictatorship satisfaction in terms of the percentage of total alternatives someone has dictatorial power over.

I think this account is on to something, for proportions of violations seem to be a relevant factor for closeness. But it cannot be the whole story, for not all violations are *on a par*. If a condition states that X-alternatives are *better* than Y-alternatives, then a theory that says that X-alternatives are *equal in value* to Y-alternatives seems closer to the condition than a theory that says that X-alternatives are *worse* than Y-alternatives. This suggests the following general principle:

> If T1 swaps C's ranking of the alternatives and T2 ties the alternatives, then T2 is closer to C than T1 is, other things being equal.

In short, swaps take a theory further away than ties, other things being equal.

Consider the Avoidance of RC, and the A- and Z-populations from above. A theory that states that A-populations are equally as good as Z-populations is closer to Avoidance of RC than a theory that says that A-populations are worse than Z-populations. So, even if two theories can have the same proportion of C-violations, one can come closer to C than the other because its violations are ties rather than swaps.

Of course, this does not disqualify the proportion-account, if we understand it as saying that the proportion of violations matter, when *other things are equal*:

> If T1's proportion of C-violations are greater than T2's, then T1 is closer to C than T2 is, *other things being equal*.

7. Ranking-distance approach

The ranking-distance approach defines a theory's closeness to a condition in terms of the *distance* between the theory's ranking and the ranking(s) given by a condition. There are three notions that need to be explained here: a theory, the notion of a distance, and the notion of the ranking(s) given by a condition. For simplicity, I will work with a course-grained notion of a theory, according to which a theory is just an ordinal ranking of populations. A more fine-grained notion of a theory would include an *explanation* of why a given ranking of states of affairs holds.

The notion of distance I am going to work with is the popular Kemeny-metric (Kemeny 1959), which has been used in the contexts of information technology and social choice. It has the virtue of being very simple, capturing some intuitive ideas about distance and closeness, being impartial between conditions, and not requiring anything more than an ordinal ranking of alternatives. (I aim to deal with alternative metrics in the future.)

The intuitive idea is that the distance between two rankings is the number of

minimal changes one has to apply in order to get from one ranking to the other. To define it more precisely, note first that a ranking R can be represented as a set of ordered pairs of alternatives, such that a pair (x, y) belongs to R if and only if R ranks x at least as highly as y. Now, the distance between two rankings, R1 and R2, is simply the number of ordered pairs that belong to either R1 or R2 but not to both of these rankings. Finally, the *total* distance between a ranking R and a *set* of rankings is the *sum* of distances between R and each ranking in the set.

Consider the following example of distances between individual rankings (i.e., all the minimal moves required to swap the top-ranked and the bottom-ranked alternatives).

Fig. 4

x		x		x		x, z		z		z		z		
y	=>	y, z	=>	z	=>	y	=>	x	=>	x, y	=>	y		
z				y				y				x		
	1		1		1		1		1		1			

Here, the distance between each adjacent pair of rankings is 1. The distance between the first and the last rankings is 6.

The notion of the rankings given by a condition is more difficult to spell out. One option is to think about the rankings given by a condition C as *all* the possible complete rankings that satisfy C. The distance between the theory and C is then the total distance between the theory and the set of all the C-complying rankings. This is a non-starter, however. On this account, no theory can be at zero distance to a condition (thus, no theory is maximally close to a condition), since any theory is at a non-zero distance to some of the C-complying rankings.[12] But we know that some theories do satisfy and thus come maximally close to some of the conditions.

Another option is to take all the rankings that satisfy condition C and then identify the ranking(s) that minimizes the distance to all other C-satisfying rankings. Call these the *representative C-ranking(s)*. The closeness of a theory to a condition C is defined by the distance between the theory and the representative C-ranking(s).

On this account, a theory can be maximally close to a condition, but the obvious problem is that this holds only if it is *identical* to the representative C-ranking.[13] Any

[12] This holds for all plausible measures of a ranking's total distance to a set of rankings, for only a ranking that is identical to all rankings in the set is overall maximally close to the set.

[13] This holds for all measures of distance, for only identical rankings are maximally close to each other on any adequate distance measure.

other theory is at a non-zero distance from the condition. So, the account will violate the trivial constraint that if two different theories both satisfy a condition C, then they are both maximally close to satisfying it, i.e., they both have the distance value 0 with regard to C.

A better idea is to look only at the relevant *sub-rankings* in theories. If a condition ranks x versus y (given that x and y stand in the appropriate relation), then we only look at how theories rank x versus y. This solves the problems with the previous accounts, since if the condition ranks x over y, then *any* theory that ranks x over y is maximally close to the condition given that the distance is 0.

This account also validates the principle that a swap takes us further away than a tie comes out as obviously true.[14] x = y is always closer to x > y than y > x is. The distance between x = y and x > y is 1, and the distance between x < y and x > y is 2, as the following diagram shows.

Fig. 5

x				y
y	=>	x, y	=>	x
	1		1	

Now the conditions we are discussing do not just rank two populations; they rank *any pair* of populations that stand in certain relations to each other, spelled out by the relevant condition: in any (A, A+)-pair, the A+-population is better than the A-population; in any (Z, A+)-pair, the Z-population is better than the A+-population; in any (A, Z)-pair, the A-population is better than the Z-population. So, in order to decide how close a theory is to a certain condition C it is not enough to look at how close a theory comes to C's ranking of a certain pair; we need to look at how close it comes to the C's ranking of *each* pair, or each *C-ranking,* as we may call them. More precisely, to see how close a theory T is to C, the idea is to first look at how close the theory comes to each C-ranking. The distance between T and C is then the sum of the distances between T and each C-ranking. In order to see how close a theory is to *all* conditions, we should sum the distances between the theory and each condition.

[14] Other distance measures validate this too. For example, the Duddy-Piggins measure (Duddy & Piggins 2012) and the Cook-Seiford measure (Cook & Seiford 1978). Note, however, that not all distance measures will validate this. For example, the Hamming distance measure (Hamming 1950) will not validate it, since it defines the difference between two rankings as the number of (unordered) pairs of objects for which the rankings disagree. This means that the Hamming distance between the first and the second ranking is 1 and so is also the distance between the first and the third ranking.

The picture is this. Assume that we have three theories, T1, T2, and T3, which provide the following rankings.

Fig. 6

T1	T2	T3
A1+ > A1 A2+ > A2 ⋮ Ak+ > Ak	A1+ = A1 A2+ > A2 ⋮ Ak+ > Ak	A1+ < A1 A2+ < A2 ⋮ Ak+ < Ak
Z1 < A1+ Z2 < A2+ ⋮ Zl < Al+	Z1 < A1+ Z2 > A2+ ⋮ Zl > Al+	Z1 = A1+ Z2 = A2+ ⋮ Zl = Al+
A1 > Z1 A2 > Z2 ⋮ Am > Zm	A1 > Z1 A2 < Z2 ⋮ Am < Zm	A1 > Z1 A2 > Z2 ⋮ Am > Zm

Note that T2 illustrates the possibility that how well a theory fares with respect to a condition can vary from one case to another (e.g., A1+ = A1, but Ai+ > Ai, for all other i). These theories will show the following closeness distances to MA, NAE, and Avoidance of RC.

Fig. 7

	T1	T2	T3
Mere Addition (MA)			
A1+ > A1	0	1	2
A2+ > A2	0	0	2
⋮	⋮	⋮	⋮
Ak+ > Ak	0	0	2
Distance to MA	a1 = 0 + 0 + …+ 0	b1 = 1 + 0 +…+ 0	c1 = 2 + 2 +…+ 2
NAE			
Z1 > A1+	2	2	1
Z2 > A2+	2	0	1
⋮	⋮	⋮	⋮
Zl > Al+	2	0	1
Distance to NAE	a2 = 2 + 2 +…+2	b2 = 2 + 0 +…+ 0	c2 = 1 + 1 +…+1
Avoidance of RC			
A1 > Z1	0	0	0
A2 > Z2	0	2	0
⋮	⋮	⋮	⋮
Am > Zm	0	2	0
Distance to ARC	a3 = 0 + 0 +…+ 0	b3= 0 + 2 +…+ 2	c3 = 0 + 0 + …+ 0
Total distance to (MA, NAE, ARC)	a1 + a2 + a3	b1 + b2 + b3	c1 + c2 + c3

This account clearly satisfies Inclusion, Closeness Dominance, and Equal Closeness. It also provides a measure of overall closeness to all conditions.

It is also sensitive to the number of violations: if the violations are uniform across cases, all a tie or all a swap, then more violations take us further away from a condition. This is easier to see if we introduce the notion of a *violation vector* that represents how close a theory comes to a condition in different cases. The first value in the vector shows the distance in the first case, the second, the distance in the second case, and so on. If the violation vector for theory T with respect to condition C is (0,

0, x) and for T' it is (0, x, x), where x > 0, then T is closer to C than T' is. But if the violations are not uniform, then one theory can be closer to a condition than another even if the first has more violations. For example, if the violation vector for one theory is (0, 1, 1, 1) and for the other it is (0, 0, 2, 2), then the first theory is closer. If you think this is a problem, you can change the aggregation metric and give more weight to smaller deviations, for instance, by using a function that gives more weight to small deviations (a concave transformation of the distance values in the vector).

Let us now see what the account says about the Mere Addition Paradox, if we consider all possible theories, i.e., all possible rankings of A, A+, and Z. For simplicity, let us again use the toy example with one specific instance of the Mere Addition Paradox, where there are only three specific alternatives to consider, A, A+, and Z.

Fig. 8

	A+ is better than A	Z is better than A+	A is better than Z
0	T1: T2: T3: ***A+*** **A+** Z ***A*** Z **A+** Z **A** **A**	T3: T4: T5: **Z** **Z** A **A+** A **Z** A **A+** **A+**	T5: T6: T1: **A** **A** A+ **Z** A+ **A** A+ **Z** **Z**
1	T7: T8: T9: **A,** Z **A,** **A+** **A+,** **A+**, Z **A** Z	T10: T11: T9: **Z,** A A, **A+** **Z,** **A+**, A **A+** **Z**	T12: T13: T9: **A, Z** A+ **A**, A+ **A, Z** A+, **Z**
2	T6: T4: T5: **A** Z **A** **A+** **A** Z Z **A+** **A+**	T2: T6: T1 **A+** A **A+** **Z** **A+** A A **Z** **Z**	T4: T2: T3: **Z** A+ **Z** **A** **Z** A+ A+ **A** **A**

The top-ranked theories in terms of overall distance to all conditions are the theories with only one violation, a swap: T1, T3, and T5 (overall distance = 2), followed by all theories with at least one tie: T7, T8, T9, T10, T11, T12, and T13 (overall distance = 3), and bottom ranked we have theories with two swap-violations: T2, T4, and T6 (overall distance = 4). This result can be generalized to theories that provide *uniform* violations of the conditions: if the theory entails a certain violation in one case (say, A1+ < A1), then it entails the same kind of violation in all cases (Ai+ < Ai, for all i).

So, we have reduced the initial 13 possibilities to 3 –that is always something— but the remaining top-ranked ones are very different (each alternative gets one top-position, one medium, and one bottom). This means that all population axiologies that judge there to be a tie between some of the populations in the Mere Additions are ruled out. In particular, it means that we have ruled out a person-affecting view, according to which adding new people – moving from A to A+ –does not make an evaluative difference. We have also ruled out a view according to which population A is not better than Z, but only equally as good as Z.

Can we break the tie among the remaining three theories? If not, it is unclear how we can be guided to act by these theories. We can't break it by applying the Kemeny-method again, for that will give us the same set of rankings back. Nor can we break it by applying the majority rule, since it leads to a cyclical ordering. (Note that the three rankings comprise a Condorcet-set.)

But closeness is not the only factor that is relevant when we assess a violation. First of all, some violations are *intuitively worse* than others. For example, a violation of Avoidance of RC that says that Z is better than A even if Z has not more total wellbeing than A seems worse than a violation that says that Z is better than A when Z has more total wellbeing because it is much bigger and the wellbeing of its members is almost crossing the ceiling for being just barely worth living. Similarly, a violation of NAE in which the well-off people are dragged down to the level of being barely worth living, like in A+ compared to Z, is worse than a violation in which one population is a Pareto-improvement of another (all people are at least as well off and some are better off). This means that even if two theories have the same proportion of C-violations, one theory can be preferable to the other because its violations are intuitively not as bad as the ones of the other theory. This suggests the following principle

> If T1's C-violations are more severe than T2's, then T1 is in that respect worse than T2 , other things being equal.

Second, some violations are *farfetched* or *unrealistic*, because they involve populations that could exist in worlds that are very far from the actual world. It seems intuitively less worrisome if the violations of the theory involve populations that are very farfetched. This might in part depend on the fact that our intuitions can be said to be less reliable when the target is some very unusual or farfetched scenario that cannot happen in realistic worlds. It might also depend on the fact that it is less problematic if a theory gives the wrong result in farfetched scenarios than in realistic scenarios.[15] As an example of a farfetched violation, consider violations of Avoidance of RC that involve Z-populations that are of such an astronomical size that they are almost not physically possible. So, two theories can have the same proportion of C-violations, but one is preferable to the other because its violations are more farfetched or more unrealistic. This suggests that

> If T1's C-violations are less farfetched than T2's, then T2 is in that respect worse than T1, other things being equal.

With these extra principles at hand we *might* be able to break the tie. Perhaps all theories tied for distance to the conditions have equally unrealistic violations, but one theory stands out as having less severe violations than the others. To have a greater chance of breaking ties, the simple ranking-distance approach must be revised. We could merge closeness with the other factors and go for an 'element-weighted' Kemeny-measure, according to which the alternatives get weighted so that a more realistic violation increases the distance, and a more severe violation increases the distance. Mathematically this can be done in many different ways, but in order to validate the principles we listed about farfetchedness and severity these weights must make the distance function increasing for both farfetchedness and severity. If we move beyond the toy-example and consider cases where the conditions supply rankings of many pairs of alternatives and the theories order all these alternatives, we have a greater chance to find differences between the theories in terms of the kinds of violations they imply. Of course, nothing guarantees that we will find enough relevant differences between the theories; it depends on which set of theories we consider.

We also have a problem of comparing the severity of a violation of a condition across theories. From which perspective should we carry out these comparisons? One option is to be subjective and just take the perspective of the moral agent. However, one might think that how severe a violation is not (wholly) up to each

[15] I am indebted to Gustaf Arrhenius and Hilary Greaves for this point.

agent to decide.[16] Furthermore, in order to compare all theories, we will have to do some trade-off between the *different* kinds of violations; one theory may have less realistic but much more severe violations than another. How should we trade off these features of violations against each other?

Even if these problems can be solved, we may still be stuck with ties where all considered theories have the same overall distance to the conditions. A partial remedy can be to consider other theoretical virtues, such as simplicity and parsimony. Furthermore, we can consider the credences we have in the conditions. Perhaps we have more credence in two of the three conditions, which would speak in favour of the theory that satisfies those two conditions.[17] More specifically, we could weight the distance between a theory and a condition by its credence.

8. Concluding remarks

This is as far as I have come in my thinking (not that far admittedly). I am unsure about how to answer all the questions surrounding how to construct a satisfactory weighted Kemeny-measure. This may provoke a very disconcerting thought: have we embarked on yet another wild goose chase, leading to another impossibility theorem, this time at a higher level?I can't show you that we need not worry about this. But note that there has been a lot of theorizing on weighted Kemeny-measures and there seems to be no known, *very general*, impossibility theorem that the researchers on ranking-distance stumble on.

In my particular application, I need to sort out the trade-off between different features of violations, but perhaps we can give people quite a lot of leeway on how to do this. Except for some general constraints, it is up to the decision-maker to decide on the trade-off between farfetchedness and severity. If the decision-maker is unsure about how to do this in all relevant cases, we can ask her to assign the alternatives some *set* of weights so that we at least get a *partial* trade-off ordering: x is more distant than y if it is more distant on all weight assignments.

There are further issues to be addressed, for recall that the discussion in this paper was premised on some simplifying assumptions. Which questions do we have to face if we lift these assumptions?

Full comparability. If we relax this assumption, we need to be able to compare *gaps* with swaps and ties. Which comes closer to a certain strict ranking? On the one

[16] Thanks to Hilary Greaves for pressing me on this issue.

[17] To determine how much credence we have in a theory we might need to know how the theory explains the value ordering. This means that we need to go beyond the minimalist framework that identifies theories with their orderings.

hand, it seems closer to a tie than a swap, since it agrees with a tie that the ranking is not reversed. On the other hand, it seems to take us further away from both swaps and ties, since a gap denies the comparability of the alternatives in question.[18]

Disjunctive conditions. I have assumed that the conditions provide strict rankings of pairs of alternatives (A+ should be better than A, Z should be better than A+, and A should be better than Z.) But what should we do when the condition provides a *disjunction* of rankings, for example A+ is *either* better than *or* equally as good as A? It seems reasonable to first determine the distance between the theory and each disjunct and then choose the *shortest* of those distances as a measure of how close the theory is to the disjunctive condition. After all, to satisfy a disjunctive condition is to satisfy one of the disjuncts. For the example above this means that a theory that says that A+ is worse than A is at a distance of 1 from satisfying the disjunctive condition, for it only takes one change (from A+ < A to A+ = A) to satisfy one of the disjuncts.

Closeness to transitivity. How do we measure closeness to the transitivity condition? This is actually not a problem for the Kemeny-measure approach. We can ask how many changes it takes to transform a target ranking into a transitive ranking. So, for instance, a violation of this kind, x > y, y > z, x = z, will be closer to the transitivity condition than a violation of this kind x > y, y > z, x < z. The former requires one change (from x = z to x > z) and the latter two (from x < z to x = z and then to x > z).

Closeness to the universal domain condition. This can be measured by the proportion of cases that the theory applies to, at least when we consider realistic cases.

Even if we can avoid another impossibility theorem, we can wonder whether it is worth trying to work out the best weighted Kemeny-measure. We started with the observation that decision-makers are in a hurry and we ended with yet another theoretical puzzle (This is a typical outcome when philosophers try to be practically relevant). Why think this puzzle is easier to solve than deciding which condition to drop?

I think the options are not exclusive. When we try to work out the Kemeny-metric and how to apply it to the paradoxes, we simultaneously consider how worrisome the violations of the conditions are. This evaluation process can make us reassess the plausibility of some condition(s); perhaps they were overshooting:

[18] If, as is claimed by Rabinowicz and Hájek (2022), we can talk about x and y being incomparable but x being almost better than y, then this kind of gap is closer to x being better than y, than to y being better than x.

some violations now seem acceptable. So, engaging in working out the closeness metric can give us ideas about which condition to drop, even if we fail to find a satisfactory closeness metric. Additionally, this work can also give us reason to change our credence distribution in the conditions, which can help us if we want to go for the hedging option. So, a failure to find a closeness metric can have instrumental value for the other approaches to addressing impossibility theorems.

In any case, we are not completely empty-handed as things stand, for we have established the following principles:

- Closeness Dominance
- Closeness Equality
- Inclusion
- Identity
- A greater proportion of violations takes a theory further away than a lesser proportion of violations, other things being equal.
- Swaps take a theory further away than ties, other things being equal.
- Violations that are more severe make a theory worse, other things being equal.
- Less farfetched violations make a theory worse than more farfetched ones, other things being equal.

Together, these principles will give us some limited guidance on how to rank theories. We can already, at least, rule out certain theories. We can already tell decision-makers *not* to use certain theories. This is progress of some sort.

References

Arrhenius, G. (forthcoming). Population ethics: The Challenge of Future Generations. Oxford University Press.

Arrhenius, G. (2021). 'Population paradoxes without transitivity'. In in Oxford Handbook of Population Ethics, Arrhenius, G, Bykvist, K, Campbell, T, and Finneron-Burns, E. eds., Oxford University Press.

Blackorby, C., Bossert W., and Donaldson D., (2005) Population Issues in Social Choice Theory, Welfare Economics, and Ethics: 39. Cambridge University Press

Brennan, G. (2015) 'Liberty, Preference Satisfaction, and the Case against Categories, in Weighing and Reasoning'. In Themes from the Philosophy of John Broome, Hirose I. and Reisner, A. (eds.), Oxford University Press.

Brennan, G. and Baurmann, M. (2006) 'Majoritarian inconsistency, Arrow impossibility and the continuous interpretation: A context-based view'. In C. Engel and L. Daston (eds.), Is There Value In Inconsistency? Baden-Baden, Germany: NOMOS, 93–118

Bykvist, K. (2022) 'Evaluative uncertainty and population ethics', in Oxford Handbook of Population Ethics, Arrhenius, G, Bykvist, K, Campbell, T, and Finneron-Burns, E. eds., Oxford University Press.

Campbell, D. E. and Kelly, J. S. (1994) 'Trade-off Theory'. Papers and Proceedings of the Hundred and Sixth Annual Meeting of the American Economic Association, The American Economic Review 84:2, pp. 422–426.

Cook, W. D. and Seiford, L. M. (1978) 'Priority Ranking and Consensus Formation'. Management Science 24:16, pp. 1721–1732.

Duddy, C. and Piggins A. (2012) 'The proximity condition'. Social Choice and Welfare 39: 2/3, pp. 353–369.

Greaves, H. and T. Ord (2017). 'Moral uncertainty about population axiology'. Journal of Ethics and Social Philosophy 12:2, pp. 135–167.

Hamming, R. W. (1950). "Error detecting and error correcting codes". The Bell System Technical Journal. 29 (2): 147–160.

Kemeny, J. G. (1959). 'Mathematics without Numbers' Daedalus 88:4, pp. 577–591.

MacAskill, W., Bykvist K., and Ord T. (2020). Moral Uncertainty. Oxford University Press.

Rabinowicz, W. and Hájek, A. H (2022). 'Degrees of commensurability and the repugnant conclusion', Noûs 56:4, pp. 897–919.

Riedener, S. (2021) Uncertain Values: An Axiomatic Approach to Axiological Uncertainty, Philosophical Studies 177, 483–504..