A St Petersburg Paradox for Risky Welfare Aggregation

Zachary Goodsell

November 22, 2020

Abstract

The principle of *Anteriority* says that prospects that are identical from the perspective of every possible person's welfare are equally good overall. The principle enjoys prima facie plausibility, and has been employed for various theoretical purposes. Here it is shown using an analogue of the St Petersburg Paradox that Anteriority is inconsistent with central principles of axiology.

An extremely natural and plausible principle of welfare aggregation is that the overall value of a prospect depends only on what it offers for each possible person. This is the principle of *Anteriority* (McCarthy, Mikkola, and Thomas 2020: 81):

(Anteriority) If each possible person is equally likely to exist in either of two prospects, and for each welfare level, each person is, conditional on their existence, equally likely to have a life at least that good on either prospect, then those prospects are equally good overall.

Anteriority must be rejected, on the grounds that it conflicts with an even more fundamental principle: the principle of *Dominance*. Nevertheless, the attractiveness of Anteriority should not be underestimated.

The positive argument for Anteriority is that it meshes nicely with a plausible welfarist picture of population axiology, according to which overall value supervenes on the distribution of individual welfare. To see why, imagine that you have two prospects that satisfy the antecedent of Anteriority. From the perspective of every possible person's welfare, then, the two prospects look exactly the same: there is the same probability of that person not existing, and conditional on her existence, there is the same probability of her ending up with any particular welfare level on either prospect. If two prospects are exactly the same from the perspective of every possible individual's welfare in this way, then it is hard to see how the distribution of individual welfare throughout the prospects could support any discrimination between them.

Anteriority is a sort of weak *ex ante* Pareto principle for prospects with variable population. Unlike *ex ante* Pareto principles, Anteriority remains neutral between various popular ways of cashing out the welfarist vision. The principle

is prima facie acceptable for Prioritarians (Parfit 1997) as well as Totalists, including those who deny that existence can be better or worse for an individual than nonexistence.

Ex Post Egalitarianism is straightforwardly incompatible with Anteriority (Fleurbaey and Voorhoeve 2013). Adherents of this view might hope to recover a weakened version of Anteriority that is restricted to pairs of prospects whose possible outcomes contain no inequality. Even this weakened version of Anteriority will be impugned by the following argument, given a slight strengthening of an auxiliary assumption.

Despite its apparent virtues, Anteriority must be rejected. The issue is not that it yields a counterintuitive verdict on a particular case, but that it conflicts with an even more fundamental principle of value:

(Dominance) If for each possible outcome, the first of two prospects is at least as likely as the second to yield an outcome at least as good as that outcome, then the first prospect is at least as good as the second (in this case we say the first *dominates* the second). If for some possible outcome the first prospect is strictly more likely than the second to yield something at least that good, then the first is strictly better (and in this case we say it *strictly* dominates).

Dominance is a fundamental assumption of decision theory that has garnered widespread acceptance. The principle is also extremely intuitively compelling: the value of a prospect, one would have thought, is solely a result of the values of the outcomes it might yield, and the probability with which it yields them, and these considerations are unanimously in favour of a prospect that dominates.

Given some widely accepted assumptions about the structure of individual welfare, Anteriority and Dominance conflict. (The conflict is closely related to puzzles arising from the possibility of unbounded value (Broome 1995; Russell and Isaacs, forthcoming) and of infinite populations (Vallentyne and Kagan 1997; Bostrom 2011). Neither possibility is assumed here.) Our first assumption concerns how welfare levels might be distributed in the world:

(Mix-and-Match)

(a) There are infinitely many possible people, and for any way of assigning welfare levels to any finite number of them, there is an outcome in which those people are the only people to exist and they all have their assigned welfare levels.

(b) For any sequence of positive real numbers that sum to one, and for any way of assigning outcomes to those numbers, there is a prospect that yields each of those outcomes with probability equal to the corresponding number.

Both parts of Mix-and-Match are typically assumed without note (Nover and Hájek defend part (b) (2004: 246–47)). Three more principles codify very plausible assumptions about the structure of value:

(Transitivity) The relation *being equally good overall as* is transitive.

(Pairwise Anonymity) Switching out someone for another person with the same welfare level doesn't change the overall value of an outcome.

(Addition Comparability) There is some outcome Ω with a finite population, and some welfare level x, such that Ω is either better or worse overall than an outcome in which two people at welfare level x are added to the people in Ω, and everybody else's welfare stays the same. (Refuting the restriction of Anteriority to prospects whose outcomes contain no equality will require the further assumption that everybody in Ω has welfare level x.)

Transitivity is usually taken to be obvious. Pairwise Anonymity is mostly uncontroversial, at least as applied to finite-population outcomes as it will be here. Vallentyne and Kagan (1997) show that problems arise for a natural strengthening of Pairwise Anonymity in infinite-population cases, but this does not make the failure of the weaker Pairwise Anonymity in finite-population cases very plausible. Addition Comparability is also seemingly acceptable: surely you would discover that things are worse than you thought if there turn out to be two more people than you thought, and their lives are horrifically terrible. Notice that Addition Comparability does not require the controversial assumption that some lives are better or worse than nonexistence for the people who live them.

The inconsistency between Anteriority and Dominance can now be demonstrated. Here is one way to do so, where the outcome Ω and the welfare level x are witnesses to Addition Comparability. First, take two infinite sequences of possible people $i_1, i_2, \ldots$ and $j_1, j_2, \ldots$, such that these sequences have no members in common with each other or with the population of Ω. Then, for each number n, let v_n be that outcome in which exactly i_1 through i_n are added to the population of Ω, all at welfare level x. Similarly, let w_n be that outcome where j_1 through j_n are added at welfare level x instead. Let $v_n^\frown w_n$ be an outcome where, in addition to the people in Ω, those from v_n and w_n also exist at welfare level x.

Now consider the following sequence of prospects (A) through (F). I write the prospects as tables, whose top row specifies the probability with which the outcome below it will come about, given that prospect:

(A)

$1/2$	$1/4$	$\ldots$	2^{-n}	$\ldots$
v_2	v_4	$\ldots$	v_{2^n}	$\ldots$

(B)

$1/4$	$1/4$	$1/8$	$1/8$	$\ldots$	$\ldots$	$2^{-(n+1)}$	$2^{-(n+1)}$	$\ldots$	$\ldots$
v_2	v_2	v_4	v_4	$\ldots$	$\ldots$	v_{2^n}	v_{2^n}	$\ldots$	$\ldots$

(C)

$1/4$	$1/4$	$1/8$	$1/8$	$\ldots$	$\ldots$	$2^{-(n+1)}$	$2^{-(n+1)}$	$\ldots$	$\ldots$
v_2	w_2	v_4	w_4	$\ldots$	$\ldots$	v_{2^n}	w_{2^n}	$\ldots$	$\ldots$

(D)

$1/4$	$1/4$	$1/8$	$1/8$	$\ldots$	$\ldots$	$2^{-(n+1)}$	$2^{-(n+1)}$	$\ldots$	$\ldots$
Ω	$v_2^\frown w_2$	Ω	$v_4^\frown w_4$	$\ldots$	$\ldots$	Ω	$v_{2^n}^\frown w_{2^n}$	$\ldots$	$\ldots$

(E)

1/4	1/4	1/8	1/8	...	...	$2^{-(n+1)}$	$2^{-(n+1)}$	...	...
Ω	v_4	Ω	v_8	...	...	Ω	$v_{2^{n+1}}$	...	...

(F)

1/2	1/4	...	2^{-n}	...
Ω	v_4	...	v_{2^n}	...

Dominance entails that (A) and (B) are equally good. By Pairwise Anonymity, v_n and w_n are equally good for each n, so (B) and (C) must be equally good by Dominance. Anteriority implies that (C) and (D) are equally good. Applying Pairwise Anonymity and Dominance again yields that (D) and (E) are equally good. (E) and (F) must be equally good by Dominance, because 1/4, 1/8 and so on add up to 1/2 exactly. The welfare level x was chosen, using Addition Comparability, so that v_2 is better or worse than Ω. By Dominance, then, (A) and (F) are not equally good, contradicting Transitivity.

There are many ways to respond to this puzzle, but investigating each option would take us too far afield. Notice, though, that the reasoning establishing the equal value of any adjacent pair of prospects from (A) through (F) seems unimpeachable except, perhaps, in the case of (C) and (D). For example, if you prefer (B) to (C), then either you regard v_n better than w_n for some n, which would be quite strange since for all I have said the people who would exist in either outcome may be exactly similar, or you think that v_n and w_n are equally good for any n, in which case it is completely obscure what feature of (B) is supposed to make it better than (C). The same goes for the comparison of (D) and (E), and that (A) and (B) and (E) and (F) are equally good is surely undeniable. Once it is granted that v_2 is not equally good as Ω, it is almost undeniable that (A) and (F) are also not equally good. The equal value of (C) and (D) seems by far the weakest link in the paradoxical chain of reasoning. So Anteriority must go.

Let us now consider a case wherein a choice between prospects (C) and (D) arises.

> Planetary Prospects:
> God will toss a fair coin until it lands tails, and she will record the number of tosses. Then, she will choose one of two buttons fairly at random, and press it. One of the buttons is red and the other is blue. As things stand, neither button does anything when pressed: your choice affects what they do when pressed. You are to modify one of the buttons so that when it is pressed 2^n Jovians are created on Jupiter, where n is the number of God's tosses. Just before you make your choice, the representative of Saturn will act similarly: she will modify the blue button so that should it be pressed, 2^n Saturnians will be created. Neither modification will affect the other, and any people created will have welfare level x.

Choosing to modify the red button yields prospect (C), and the blue button prospect (D), assuming that we would have outcome Ω were God to abstain from pressing. It was established above that (C) is better or worse than (D), so

this choice is a weighty one. It is weighty even though no possible person gets a greater chance of existence, or a greater chance of a better life, given either choice.

Planetary Prospects may be used to cast doubt on patterns of reasoning employed elsewhere in population axiology. Parfit (1984: 419) observes that a certain version of Averagism implies that whether it is good to have children now depends on the average welfare and number of the ancient Egyptians. This consequence seems absurd. He writes: 'research in Egyptology cannot be relevant to our decision whether to have children'. By the same token, research into how things will go on Saturn ought to be irrelevant to what we should do on Jupiter, assuming that there will be no interaction between the planets. But Planetary Prospects shows that this thinking is wrong. Counterintuitively, information that has nothing to do with what you can affect *can* make a difference to the comparative value of the prospects available to you. Specifically, the information that the representative of Saturn chose to modify the blue button is relevant to which button you should modify. Parfit's Egyptology objection to Averagism is therefore rendered indecisive.

(Superficial adjustments to the case show exactly that information about what went on in ancient Egypt could be relevant to the values of prospects now. Specifically, make it so that God's coin tossing and button pressing all occurred in the past, and so that your choice is between 2^n children at welfare level x being created now if God pressed the red button, and those same children being created if God pressed the blue button. Furthermore, make it so that Ramesses II faced a similar situation in ancient Egypt thousands of years ago, and that any Egyptian children created in this way were cordoned off so as to not affect the rest of history.)

Given that Anteriority is to be rejected, it is reasonable to ask whether some weaker principle in the vicinity can replace it. One natural candidate is the restriction of Anteriority to finite-population prospects. To be precise, let a *finite-population prospect* be a prospect on which there are only finitely many possible people who have a chance of existing. *Finite Anteriority* is the restriction of Anteriority to finite-population prospects:

(Finite Anteriority) If each possible person is equally likely to exist in either of two *finite-population* prospects, and for each welfare level, each person is conditional on their existence equally likely to have a life at least that good on either prospect, then those prospects are equally good overall.

Since most of the classical problems in population axiology arise in finite-population cases, Finite Anteriority will, in those cases, do any theoretical work that one could hope to use Anteriority for. Moreover, Finite Anteriority is consistent with all of the previously employed principles besides Anteriority (the proof is too involved to include here).

Should Finite Anteriority be accepted by those initially inclined towards Anteriority? The answer to this question turns in part on whether motivation can be found for Finite Anteriority that does not extend to Anteriority proper. Af-

ter all, if an argument for Finite Anteriority also works in favour of Anteriority, there must be something wrong with the argument.

The project of motivating Finite Anteriority without overgeneration is a difficult one, but not without promise. Many initially plausible principles begin to break apart when applied in full generality, but often the core idea need not be abandoned. The above argument against Anteriority in its full generality should be taken as a warning for those who endorse Finite Anteriority, that utmost care must be taken in this vicinity.

There is a related difficulty for Finite Anteriority that also warrants consideration. It is a problem even for the fixed-population restriction of Finite Anteriority. To generate the problem, suppose that Totalism is true for finite populations, which is to say that the overall value of an outcome is given by the sum of real-valued welfare levels of its inhabitants. Now consider the prospects (I) and (II), each of which involves the certain existence of only Ann and Bob (this example is inspired by one of Seidenfeld, Schervish and Kadane (2009: 333)):

(I)

	1/2	1/4	...	2^{-n}	...
Ann	2	4	...	2^n	...
Bob	2	4	...	2^n	...
Total	4	8	...	2^{n+1}	...

(II)

	1/4	1/4	1/8	1/8	...	...	$2^{-(n+1)}$	$2^{-(n+1)}$	...	...
Ann	4	2	8	2	...	...	2^{n+1}	2	...	...
Bob	2	4	2	8	...	...	2	2^{n+1}	...	...
Total	6	6	10	10	...	...	$2 + 2^{n+1}$	$2 + 2^{n+1}$	...	...

By Finite Anteriority (and even the restriction of Finite Anteriority to fixed population cases), (I) and (II) are equally good. Since 1/4, 1/8 and so on add up to exactly 1/2, both people are in both prospects faced with a half chance of welfare level 2, a quarter chance of welfare level 4 and so on. On the other hand, given Totalism for outcomes with finite populations, (II) strictly dominates (I) overall and is thus strictly better.

There are numerous ways out of this puzzle, for those attracted to Finite Anteriority. One is to deny Totalism for fixed-population comparisons. Another is to deny that there is a sequence of welfare levels with the required structure, for instance by positing that there are finite bounds on how good or bad a life can get. Nevertheless, it is clear that even extensive restrictions of Anteriority do not automatically get us out of trouble. Those of us who still hope to employ Anteriority-like principles in the foundations of population axiology have a lot of work to do.[1]

1. This paper has been greatly improved by comments from and discussion with Weng Kin San, Jake Nebel, Jeff Russell and an anonymous reviewer for *Analysis*.

References

Bostrom, Nick. 2011. "Infinite Ethics." *Analysis and Metaphysics* 10:9–59.

Broome, John. 1995. "The Two-Envelope Paradox." *Analysis* 55 (1): 6–11.

Fleurbaey, Marc, and Alex Voorhoeve. 2013. "Decide As You Would With Full Information! An Argument Against Ex Ante Pareto." In *Inequalities in Health: Concepts, Measures, and Ethics,* edited by Ole Norheim, Samia Hurst, Nir Eyal, and Dan Wikler. Oxford University Press.

McCarthy, David, Kalle Mikkola, and Teruji Thomas. 2020. "Utilitarianism with and Without Expected Utility." *Journal of Mathematical Economics* 87:77–113.

Nover, Harris, and Alan Hájek. 2004. "Vexing Expectations." *Mind* 113 (450): 237–249.

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

———. 1997. "Equality and Priority." *Ratio* 10 (3): 202–221.

Russell, Jeffrey Sanford, and Yoaav Isaacs. Forthcoming. "Infinite Prospects." *Philosophy and Phenomenological Research.*

Seidenfeld, Teddy, Mark Schervish, and Joseph Kadane. 2009. "Preference for Equivalent Random Variables: A Price for Unbounded Utilities." *Journal of Mathematical Economics* 45:329–340.

Vallentyne, Peter, and Shelly Kagan. 1997. "Infinite Value and Finitely Additive Value Theory." *Journal of Philosophy* 94 (1): 5–26.