Week 2: Population Axiology and the Landscape of Existential Risk

January 27, 2023

Population axiology is the study of which outcomes are better or worse than others, when these outcomes involve changes in who exists. Population axiology is important, especially for our purposes, because it's much more common than you might think for our actions to affect who it is that exists in the future. This goes double when we're explicitly trying to affect the (far) future.

We can think of population axiology as being concerned with two central questions:

(i) How do we compare outcomes which involve the same number of numerically distinct people?

(ii) How do we compare outcomes which involve different numbers of people?

Here are three extreme answers to (i):

I Anonymity: differences in identity make no difference.

II Full incomparability: differences in identity always render two outcomes incomparable.

III Extreme narrow person-affecting view: only differences in wellbeing levels for those who exist in both outcomes matter for the ranking.

Question: Do any of these answers to (i) seem plausible? Do any suffer from serious problems? Are there more moderate answers which might be more successful?

When it comes to (ii), much of the discussion in the philosophical literature has centred around how to avoid Parfit's "Repugnant Conclusion": the claim that it would be worse for there to be a large number of excellent lives than for there to be some (arbitrarily) larger number of mediocre lives. This is harder than it seems, as evidenced by the Mere Addition Paradox. We can broadly carve up the space of different-number population axiologies into four categories (they're not all mutually exclusive):

I Deny the Mere Addition Principle; say that A^+ is worse than/not better than A [Average/variable value/asymmetric/some rank-discounted/some egalitarian views.]

II Mess with the fixed population principles; deny that Z is better than A^+ [lexical/some anti-aggregative views].

III Deny a structural feature of the argument. [Non-transitive/option set dependent views.]

IV Accept the Repugnant Conclusion. [Total Utilitarianism/Prioritarianism.]

Question: Which of these approaches seems the most promising?

Avoiding the Repugnant Conclusion isn't as easy as denying one of the premises of the Mere Addition Paradox. This is an important argument, but there are other, stronger arguments. Some highlights: Arrhenius 2011, Nebel 2019, Spears and Budolfson 2021. Some of the potential pitfalls of views which avoid the Repugnant Conclusion:

I Violating non-sadism conditions.

II Violating plausible fixed-population conditions (Non-Elitism/Non-Extreme Priority)

III Endorsing a lexical structure of wellbeing (denying Finite Fine-Grainedness)

IV Denying ex ante pareto principles

V Pairwise Cyclicity

Question: Might any of these options be ok on closer inspection?

What does all this mean for us?

Population axiology is highly **controversial**. But the case for Longtermism is, supposedly, pretty strong. One common misconception is that the case for Longtermism rests solely on reducing extinction risks (thereby vastly increasing the number of expected future lives). But this is only one important effect we might have on the future. We can categorise three main important effects we might be able to have on the far future:

E1 Reducing extinction risk, thereby increasing the expected number of future lives.

E2 Improving the average quality of future lives (from good to better).

E3 Reducing the expected number of future unhappy lives (either by eliminating them entirely or by replacing these with good lives).

Question: Do concerns about the asymmetry mean we shouldn't treat E1 effects too seriously? Are there other reasons to be sceptical?

Question: Does the non-identity problem mean we shouldn't treat E2 effects too seriously? Are there other reasons to be sceptical?

Question: Are there reasons to be sceptical that it would be morally important to produce E3 effects?

Question: What do you think the views we've discussed would say about E1/E2/E3 effects?

The Risk Landscape

Existential catastrophe via	*Chance within next 100 years*	
Asteroid or comet impact	~ 1 in 1,000,000	
Supervolcanic eruption	~ 1 in 10,000	
Stellar explosion	~ 1 in 1,000,000,000	
Total natural risk	**~ 1 in 10,000**	
Nuclear war	~ 1 in 1,000	
Climate change	~ 1 in 1,000	
Other environmental damage	~ 1 in 1,000	
'Naturally' arising pandemics	~ 1 in 10,000	
Engineered pandemics	~ 1 in 30	
	Unaligned artificial intelligence	~ 1 in 10
Unforeseen anthropogenic risks	~ 1 in 30	
Other anthropogenic risks	~ 1 in 50	
Total anthropogenic risk	**~ 1 in 6**	
Total existential risk	**~ 1 in 6**	

Figure 1: Estimates of existential risks from Ord 2020

Question: Are there any missing risks?

Question: Do the risk estimates look right? (Remember they are estimates of existential, non-recoverable risk).

Question: Do you think there might be a substantially greater chance of any of these risks leading to a non-existential global catastrophe?

Question: Would a non-existential global catastrophe indirectly lead to a higher probability of an existential catastrophe?

Question: For each risk, how much do you think we could reduce the probability, proportionally, if someone were to effectively deploy $1bn of extra funding to that end?