# Week 5: Extinction Risk and the Asymmetry

February 16, 2023

Many people find the following combination of judgements plausible:

*The Asymmetry*

(i) It is not better (or there is no moral reason) to create happy people, but

(ii) It is worse to (or there is moral reason not to) create unhappy people.

The second conjunct (ii) is fairly uncontroversial, but (i) has been much discussed. It's often supported by a famous quote by Jan Narveson (1973, p. 80): "we are in favour of making people happy, but neutral about making happy people". For our purposes, we're mainly interested in three things. First, if (i) is true, to what extent does this undermine Longtermism or change Longtermist priorities? Second, is (i) true?

**Question**: What's your first-pass answer to these two questions?

Some arguments for Longtermism appeal to the vast amount of value that would be wasted if humanity goes prematurely extinct, thus leaving the total population across history much smaller (e.g., Bostrom 2003). These arguments look wrong-headed if the Asymmetry is correct.

For other arguments it's less clear what the upshot of the Asymmetry would be. Some Longtermists argue that there are things we can do to increase the average quality of future lives while keeping the (expected) number roughly the same. They might also argue that we should attempt to prevent risks of there being a future in which many people have bad lives.

**Question**: Does the Asymmetry undercut either of these other sorts of interventions? Does it support any of them? Does it suggest any change in priorities relative to those of, say, a Total Utilitarian?

There's a substantial literature on whether (i) is true. (Broome 2004; Broome 2005) gives an important argument against the axiological version of (i). It goes roughly like this: Broome thinks that we should accept (i) only if we believe (like Narveson) that adding happy people is a neutral matter. Neutrality might be interpreted as *equal goodness*. That is, adding someone in the "neutral range" (which might be all the wellbeing levels of lives worth living) is equally as good as adding nobody.

Broome identifies a problem with this view. Suppose wellbeing levels 10 and 20 are in the neutral range. Suppose we start with some population $A$, and add Bee ($B$ for short) at level 10, giving us $A+B_{10}$. According to the equal goodness interpretation, $A \sim A + B_{10}$. But also $A \sim A + B_{20}$. By transitivity, we can conclude that $A + B_{10} \sim A + B_{20}$. But Broome says that is obviously false: the second population is better. He concludes that we should accept the equal goodness interpretation of neutrality.

**Question**: Is Broome drawing the right conclusion? Might there be a way to resist it?

There is another way to understand neutrality. It could be understood as value incommensurability, rather than equal goodness. Since the incommensurability relation is non-transitive, we can have $A+B_{20} \succ A+B_{10}$, even though $A$ is incommensurable with both of these populations.

However, Broome identifies another problem with the incommensurability interpretation: this sort of neutrality is *greedy*. Suppose the population $A$ consists of two people, $P$ and $Q$, who are initially at wellbeing levels 10 and 20 respectively. That is, $A = P_{10} + Q_{20}$. According to the incommensurability interpretation, $A$ is incommensurable with both $P_{10} + Q_{20} + B_{10}$ and $P_{10} + Q_{20} + B_{20}$. Applying a principle of impartiality, it's easy to see that

(*1) $P_{10} + Q_{20} + B_{10} \sim P_{10} + Q_{10} + B_{20}$

(*2) $P_{10} + Q_{20} + B_{20} \sim P_{20} + Q_{20} + B_{10}$

It follows that

(*3) $P_{10} + Q_{20} \parallel P_{10} + Q_{10} + B_{20}$
(*4) $P_{10} + Q_{20} \parallel P_{20} + Q_{20} + B_{10}$

Broome thinks that these conclusions show that the incommensurability interpretation doesn't work. He says:

> *Incommensurateness is not neutrality as it intuitively should be. It is a sort of greedy neutrality, which is capable of swalling up badness or goodness and neutralizing it. This is implausible [and ...] will turn out to have some practical implications that are very implausible indeed.*
>
> (Broome 2004, p. 170)

**Question**: Do you think Broome is right?

Other authors have pushed back on Broome's objection to the Asymmetry, notably Frick (2017) and Rabinowicz (2009). Frick is happy to accept the incommensurability interpretation of neutrality, and in fact argues that this interpretation makes (i) *more* plausible, since doing otherwise would amount to giving zero weight to the wellbeing of new persons. Rabinowicz argues that the greediness argument establishes that additions of good lives must be able to count for or against other values, but does not establish that additions of good lives must sometimes make an outcome better.

I (Francis n.d.) have provided an alternate version of Broome's greediness objection which I think does something to undermine these responses. It's an attempt to show that the "greediness-type" conclusions really are unacceptable. We start with perhaps one person $P$ at a slightly negative wellbeing level, say $-1$. Let $T$ be some collection of a trillion people. We can add the $T$ people at the top of the neutral range, say level 100, resulting in the population $P_{-1} + T_{100}$. Finally, we can imagine bringing everyone in $T$ down to a mediocre level in order to slightly improve $P$'s life, and push her over the threshold of a life worth living, leaving us with the population $P_1 + T_1$. Now consider the following argument:

(P1) $P_{-1} + T_{100} \succ P_1 + T_1$
(P2) $P_1 + T_1 \succ P_{-1}$
(C) Hence $P_{-1} + T_{100} \succ P_{-1}$

The thought behind (P1) is: it would be absurd to prioritise tiny benefits to $P$ over massive benefits to each of the trillion in $T$, just because $P$ is *slightly* worse off. The thought behind (P2) is: a situation where everyone has a life worth living (even if only barely) must be better than a situation where everyone has a life worth not living. If I've got a bunch of good stuff on the left, and a bunch of bad stuff on the right, I don't need to know how much is in each direction in order to know that the stuff on the left is better.

**Question**: Is this a good argument? Where might it go wrong? How do you think Frick or Rabinowicz might respond?

**Further Issues**

(1) We've mostly been talking about the *axiological* asymmetry. How might things change when it comes to deontology? Can arguments against (i) be reconstructed in this case?

(2) Frick (2017) posits a final value of humanity in order to explain why we have some reason to avoid extinction (which is painless/bad for no one) even though we have no reason to increase the number of happy people. Is this a plausible view? Might it generate reasons to avert extinction of arbitrarily large strength, as the expected lifetime of humanity increases?

(3) Even supposing there are persuasive arguments against (i), it might be too central to our intuitive judgements to give up. For example, almost nobody acts as though we have moral reasons to have children that are even close to as strong as our moral reasons to save lives. Is (i) just too compelling to abandon?