Philosophy of Science
Problem Set #4

Ozaner Hansha

May 12, 2021

Question 1

Part a: The issue examples are:

- Important, Tractable, Not neglected: Thinking about how to avoid unaligned super A.I. (e.g. conferences, papers, education, etc.).
- Important, Not tractable, Neglected: Explosion of the sun, heat death of universe, etc. Huge cosmic events such as these are important, in that they determine humanity's long term future, but are untractable due to their scale, and negelected mostly because they are so far into the future/impossibly untractable.
- Important, Tractable, Neglected: Some might argue that serious change towards climate change is just such a problem. It is important in maintaining the habitability of the planet for humans, tractable in that it is certainly possible to aggressively push towards less carbon emission, but neglected in that actual large scale changes and predicted emissions rate are far and away higher than anything required to avoid catastrophe, despite widespread knowledge of the issue, due to lack of political will on the largest contributors (i.e. corperations).

Part b: Climate change, while tno entirely neglected, seems to pose the greatest existential risk to humans in the short term. It also has the crucial property of not being tied to a single event or action, unlike a super AI or astronomical event. Such events are concrete and thus easy to gather support/political will around. But invisible threats, climate change being the prototypical example of, pose a much greater threat than other existential risks due not to its severity or probability, but the difficulty in taking actions towards it, as its effects and responsibility are spread out.

Part c: An existential risk is an event that could occur that has an appreciable chance to bring something/group of things to extinction (i.e. wipe them out). The term is most synonyms with human existential risks but, applies to other groups as well. For example, climate change poses an existential risk to the great coral reefs.

Part d: Imagine climate change has a 10% probability of existential catastrophe, while a sudden large asteroid impact has a 90% chance of existential risk. While the probabilities here might not be realistic, the asteroid impact occuring is certainly far more serious than climate change. Desptie this however, the actual probability of the asteroid impact occuring is far less likely than climate change, which is all but assured at this point. We can phrase this mathematically:

$$\begin{aligned} P(\text{Climate Change}) &\approx 1 & P(\text{Catastrophe by Climate Change}|\text{Climate Change}) &\approx .1 \\ P(\text{Asteroid}) &\approx 0 & P(\text{Catastrophe by Asteroid}|\text{Asteroid}) &\approx .9 \\ P(\text{Catastrophe by Climate Change}) &= P(\text{Catastrophe by Climate Change}|\text{Climate Change})P(\text{Climate Change}) &&\approx .1 \\ P(\text{Catastrophe by Asteroid}) &= P(\text{Catastrophe by Asteroid}|\text{Asteroid})P(\text{Asteroid}) &&\approx 0 \end{aligned}$$

Part e: One type is the betterment of humanity. Some may argue that humanity living now may be worth more to us than humanity simply existing in the long term future. This is especially true, they'd argue, if humanity now is in a particularly destitute state (as many are), and/or if the future of humanity we are

ensuring the existence of is a destitute one (e.g. one where we cannot live comfortably due to climate change, economic disaster/slavery, etc.)

Another type of phenomena might be explicitly lowering the number of future humans due to the chance of much greater suffering for later generations. For example, if it was found that humanity's future was guaranteed to be bleak, yet could still exist, some may find it better to prioritize doing things to prevent the future of humanity from even existing (e.g. voluntary banning of conception).

Question 2

Part a: Assuming they are at similar levels of climate/environmental management for those 7 centuries, they will likely have extremely exacerbated climate problems. Assuming they inhabit a world like earth, their habitable land is warmer and weather more chaotic. Other than this, or even if their current climate is similar to ours, I don't see much difference than with our current situation. I suppose overpopulation might be an issue, but evidence points to our population leveling off soon, following some sigmoidal curve.

Part b: In such a situation, we would have no reason to believe in any great filters (see part c) prohibiting intelligent life from spreading across the universe or simply existing. Here it seems any hurdles to civilization were just in having them exist in the first place (e.g. life to hard to evolve, rare earth, no planets, etc.) and since we have already past that hurdle, belief in any other hurdle in the future can't be based off the lack of alien species, but some other inference.

Part c: If we conclude from the discovery that complex life (by complex we mean multicellular plant life like allege) is much easier to develop on planets than the rare earth hypothesis supposes, then this means that life of this sort should be common all over the universe. Taking this together with the Fermi paradox, then, points to some great hurdle between galaxy spanning civilizations (or even just space age ones) and hte evolution of intelligent life (which can arise in as few as 100s of millions of years after multicellular life). If there wasn't some such great hurdle then we'd expect to see signs of such civilizations all throughout the universe via radio waves and other signatures.

The existence of such a great filter poses a problem for humans since we are right at the stage where they would seem to have to apply, given the fermi paradox. If so many other planets, that we now assume to have been able or had evolved complex life, failed to produce civilizations capable of being detected from afar, our own chances may be just as slim.

Part d: If one were to take the Doomsday argument seriously, then we should be able to predict the mean of the human population distribution (across past, present, and future) given our current population figures. The argument holds that given any point (really range) of years we would expect ourselves to be born in, the mean is the most likely, with the probability lowering as we stray farther from it. If, then, the last generation of humans is upon us, due to them being essentially immortal and sterile, then doomsday argument logic tells us humanity will survive for much longer (technically indefinitely) than predictions without immortal humans. The odds we'd be born in such a time (the mean is always near the present since it always has most of the human population due to the basically always increasing population) are high considering the distribution ends at this last generation.

Part 2f: First let us add the assumption that as time goes on more humans will have more happy years. In other words, humanity's happiness isn't decreasing unboundedly nor is bounded (i.e. a converging series of happiness). We can make this concrete by saying that human happiness is linearly correlated with the number of years it exists (this should be a low ball considering more happy humans are created with each generation):

$$H = Y$$

Strategy A has the following:

$$
\begin{aligned}
E[H] &= E[Y] && \text{(years mean happiness)} \\
&= \sum_{n=0}^{\infty} n \underbrace{(1-.05)^n}_{\text{alive for } n \text{ years}} \cdot \overbrace{.05}^{\text{extinct the next year}} && \text{(Probabiility of existence using Strategy A)} \\
&= 19 && \text{(no need to show my work here)}
\end{aligned}
$$

Strategy B has the following:

$$
\begin{aligned}
E[H] &= E[Y] && \text{(more years, more happiness)} \\
&= \sum_{n=0}^{\infty} n \underbrace{(1-.15\cdot.99^n)^n}_{\text{alive for } n \text{ years}} \cdot \overbrace{.15\cdot.99^n}^{\text{extinct the next year}} && \text{(Probabiility of existence using Strategy B)} \\
&\approx 139.959 && \text{(no need to show my work here)}
\end{aligned}
$$

The math is clear, strategy B results in a much longer expected period for the existence of humanity, and thus more happiness. Indeed we'd expect a larger difference since we only assumed a linear relationship between years and happiness, not accounting for rising levels of human quality of life as technology progresses and other factors.

Question 3

Part b: The person-affecting principle seems to be at odds with the non-identity problem. Consider the example of a mother waiting 3 months to have a healthy baby, rather than doing it immediately and having an un healthy one. One cannot simultaneously hold the person-affecting view and claim that this choice was wrong, in any moral sense at least. This is because no matter the choice, no person (other than maybe the mother making the choice) is harmed. If the mother waits, the baby who would have never been born will never exist, and it is impossible to harm them. The same goes vice versa.

Part c: The future beneficence principle (FBP) states that if an action can be taken to greatly improve the well-being of future generations, even if it changes the identity of that future generation, than it should be taken.

In the case of the mother (or say a large group of mothers), this would rule that they should wait, to prevent a generation of unhealthy children with no recourse for a better life. The identity of those children changed, i.e. those are different childrem that have an improved well-being, but the principle allows for that.

If the principle didn't allow for that, that is if it only perscribed improving the future of our progeny without changing their identities, then only actions that improve those who are already born/those that do not affect when and under circumstances people are conceived. But this is impossible, pretty much any action will vary when and how people are conceived, e.g. building a new bridge for my already born children will make the time for someone to drive home from work faster and thus they would conceive a child with their spouse eariler than if the bridge had not otherwise been built leading to a different zygote being formed, etc.

When we delve into it, it seems that without allowing for the change in the future generation's identities, the future beneficence principle cannot prescribe anything without being contradictory. "Do whatever you would have done if you had not read this principle, lest the identities of future generations be altered." It's hardly a convincing statement.

Part d: If one denied that robots could even have "lives" of the sort that we should supposedly strive to make better by the FBP (e.g. dualist, robots don't have souls, etc.) then a future of artificial replacement is *not* supported by the future beneficence principle.

Part g: If humans could merge with AI, and be identically capable of whatever arbitrary moral goodness the artificial entities referenced in Shiller's argument have, then this would have no bearing on the argument. All it would change is that we wouldn't call for human extinction and replacement, just replacement.

If it was impossible, then it strengthens the part of Shiller's argument that we should have humans go extinct rather than be improved.