Civilizational filters and distribution of values in the multiverse

Complementary notes on multiverse-wide superrationality

Caspar Oesterheld

One factor that might affect the distribution of values in the multiverse is the possibility of civilizational collapse or extinction.

Given the number of stars in our galaxy and the fact that it is common for stars to have planets, many habitable planets probably exist in the galaxy, some of which are probably older than Earth. If evolution had happened on these other planets in a similar way as it happened on Earth, intelligent life ought to have developed on at least some of those planets. And at least some of those lifeforms ought, in turn, to have built technological civilizations capable of sending detectable signals into space. Given these assumptions and the immense size and age of our galaxy, not only should many such civilizations already have arisen in our corner of the universe; we should be able to detect their signals or even observe them directly (see the Drake equation). And yet, we do not. This is the Fermi paradox – the discrepancy between estimates of the probability that we can somehow observe intelligent life in our galaxy, and our lack of any such observations. Various explanations of the Fermi paradox have been proposed (Webb, 2015). It could be that Earth-like or otherwise habitable planets are actually very rare, or that humans are indeed just the first intelligent civilization to arise in the galaxy. Perhaps we have not existed long enough to detect or be detected by extraterrestrials. Maybe we are bad at detecting alien civilizations, possibly because they quickly develop communication technologies that are more advanced than radio, or because we do not know what to look for – alien life is just too weird. It could be that civilizations are afraid of each other and hence avoid communication, etc.

One interesting class of hypotheses is Robin Hanson's Great Filter (Hanson's TEDx talk provides a good introduction to the topic). The main idea is that the development of intelligent life toward a space-colonizing civilization contains at least one step that is very difficult to pass, thus rendering the number of civilizations colonizing space smaller than a naive calculation may suggest. We do not know how early or late in evolutionary timelines these filters occur. It could be that the formation of simple cells is very unlikely to happen on any given planet, that simple cells very rarely develop into complex eukaryotes, or that complex single cells in most cases do not assemble into still more complex multicellular life forms, and so on. But such filters could also appear at much later stages. For all we know, most intelligent civilizations may self-destruct before becoming visible to other civilizations. Humanity came dangerously close to nuclear war more than once, and other civilizations may not have been quite as lucky. Figure 1 illustrates the concept of filters at multiple stages of the development of a civilization.

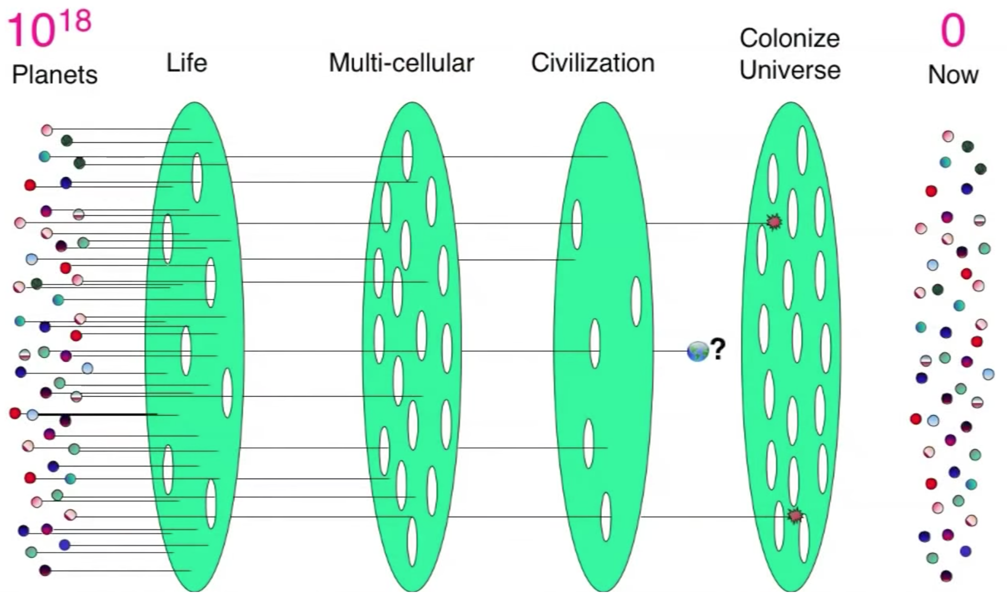

Figure 1: An illustration of the different levels of potential filters, from Robin Hanson's TEDx talk "The great filter".

Filters at earlier levels would mainly make civilizations rare in general. Late filters, on the other hand, change the distribution of values – value systems that more commonly lead to extinction, collapse, or stagnation have less power than those who lead to space colonization. The specific implications for the distribution of values depends on the respective filter. Here are some tentative example considerations:

- If the main filter is the use of weapons of mass destruction (WMD) in warfare, it will favor the survival of more peaceful civilizations, and hence render their values more important. Civilizational traits that correlate with peacefulness may include democracy, using (or not using) WMDs to deter conflict, capitalism, having fewer internal value differences or greater tolerance thereto, a world government, global authoritarianism, etc. These traits may, in turn, correlate with particular values.
- If the filter is some risk whose mitigation requires large-scale coordination, it will favor civilizations that are good at (causal) coordination problems. Such risks may include global warming, resource depletion, and existential risk from research and technology (like gray goo, black holes created in particle colliders, nuclear weapons tests igniting the atmosphere, etc.). Similarly, civilizations that manage to progress carefully will likely also be favored over ones that cannot contain their scientific curiosity or economic growth.
- If a mere lack of interest in colonizing space is a significant filter, then adventurous civilizations with a preference for colonizing space will have more influence than those who, for whatever reason, decide not to venture beyond their home planet.
- If the filter is something that civilizations have hardly any influence on, such as asteroid

impacts, supernovae etc., it is unlikely to favor any particular value system[1].

There are further civilization-level events that cannot be used as explanations of the Fermi paradox, but that nevertheless have similar implications for how much power different value systems hold:

- Some global catastrophes may not destroy a civilization, but still change their values significantly. As before, this gives power to civilizations that can avoid value-changing catastrophes, but it also gives power to the value systems that result from such catastrophes.
- Uncontrolled artificial intelligence (AI) does not work well as an explanation of the Fermi paradox, because an uncontrolled AI could still colonize space on its own and thereby represent a spacefaring civilization (if only as a rogue relic of the one that created it). Nevertheless, a failure to control artificial intelligence is one way for a civilization to lose a lot of its power. As with the other technological risks mentioned above, this AI filter would select for careful and well-coordinated civilizations. Similarly, because AI arms races would make AI control especially difficult, it favors the survival of peaceful civilizations, just as with the WMD warfare filter. A society with a large fraction of whole brain emulation citizens (Hanson, 2016) may also be better at controlling AI (cf. Salamon and Muehlhauser 2011; Bostrom, 2014), resulting in these societies having a higher influence than civilizations that directly create *de novo* AI, i. e. artificial intelligent systems not modeled after humans or other animals.

Needless to say, these are all speculative claims which we should not expect to have strong practical implications without further examination. Instead, they should be seen as examples of the kind of considerations that are necessary in order to properly estimate the distribution of values across the multiverse.

References

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies.* OUP Oxford.

Hanson, Robin (2016). *The Age of Em: Work, Love, and Life when Robots Rule the Earth.* Oxford University Press.

Webb, Stephen (2015). *If the Universe Is Teeming with Aliens ... WHERE IS EVERYBODY?: Seventy-Five Solutions to the Fermi Paradox and the Problem of Extraterrestrial Life.* Springer.

[1]Civilizations that develop protection mechanisms (like a colony on another planet) faster are less vulnerable to some of these risks, but these differences appear negligible. They could be significant in regions of the multiverse where, say, asteroid impacts are especially common, but such regions are probably too hostile for higher lifeforms to evolve in the first place.