Colonization, Consciousness, and Longtermism

Prepublication draft.
Forthcoming in *Outer Space and Philosophy* (Routledge)

Émile P. Torres

1. What Is Longtermism?

Axiological longtermism is, roughly speaking, the thesis that the value of our actions in the present largely derives from their effects in the far future. Deontic longtermism is, also speaking roughly, the claim that what we ought to do is whatever makes the far future go best (Greaves and MacAskill 2021). This is based on considerations from (a) the field of physical eschatology, or the study of the future evolution of the cosmos, including our own solar system, and (b) modern cosmography, or "the science that describes and maps the general features of the universe" (Boeyens and Levendis 2008). According to (a) and (b), humanity or our posthuman descendants could survive on Earth for another ~1 billion years, and if we spread beyond our solar system, our lineage could persist for at least another ~10^40 years, at which point protons are expected to decay (although we do not know for sure that this will happen) (Adams 2008).

Furthermore, given that there are upwards of 400 million stars in the Milky Way galaxy and up to 10^23 stars within our future light cone—that is, the region of the universe that we could theoretically access if traveling at the speed of light —the future human population could be enormous (Ord 2021, 27). There are two possibilities here, depending on whether future people are biological or digital beings, the latter of whom would live almost entirely in virtual-reality computer simulations. On the standard account, these simulations would be run on planet-sized computers powered by Dyson swarms, thus enabling a greater population density than if our descendants were biological beings residing on terraformed exoplanets or in free-floating spacecraft like O'Neill cylinders. According to Toby Newberry, the Milky Way galaxy could contain some 10^36 biological and 10^45 digital people (Newberry 2021). On Nick Bostrom's count, there could be 10^58 digital people in the universe as a whole, although he adds that the number is probably much greater (Bostrom 2014). In an earlier paper, Bostrom estimated some 10^23 biological humans per century in the Virgo Supercluster alone, which could also house roughly 10^38 digital people per century (Bostrom 2003).

Hence, the longtermist thesis is based on the idea that if all the consequences of our actions count equally, then the potential bigness of the far future implies that these far-future effects will be the primary determinant of the value of our actions in the present. On a totalist utilitarian view, this means that the actions we ought to take are those that positively influence the far future rather than near term. “For the purposes of evaluating actions,” Greaves and MacAskill write, “we can in the first instance often simply ignore all the effects contained in the first 100 (or even 1000) years, focusing primarily on the further-future effects. Short-run effects act as little more than tie-breakers” (Greaves and MacAskill 2019).

2. Overview of the Argument

In this chapter, I want to take a closer look at the longtermist vision of the far future and highlight several issues that, as far as I know, have not been taken seriously enough in the longtermist literature. My first claim is that colonizing the universe beyond our solar system will *almost certainly* require us to become or create digital beings, as interstellar travel looks to be impossible for biological humans. In other words, to produce “astronomical” amounts of “value” in the far future, we will need to colonize the universe, but to colonize the universe, we will need to create what I will call a “Digital World,” i.e., a population of beings that are digital rather than biological in nature. This yields two potential challenges:

First, a Digital World would be, in ways that are not often appreciated, radically different from the world we are familiar with. For example, it is not clear that digital “beings” would be countable, as the calculations above assume: they would more likely take the form of hive minds, collective intelligences, distributed selves, and protean entities that continually share bits and pieces of themselves with other entities in their environment. This underlines the problem of “cluelessness,” whereby judging the goodness or rightness of our actions in terms of their consequences is difficult or impossible because we are, in a deep and fundamental sense, unable to anticipate these consequences. Put differently, longtermists frequently defend their position using expected value calculations, which involve assigning probabilities to known outcomes having determinable values and then taking the average of these probability-weighted values. But if we cannot even begin to imagine the outcomes that might obtain once a Digital World has been inaugurated, we cannot use expected value calculations to guide our actions. The Digital World, which is necessary for space colonization, would be so alien from our own world that we are clueless in a very profound way.

Second, the longtermist vision of the future doesn’t just depend on the creation of a Digital World. It also (ostensibly) requires the digital “beings” that spread into the universe to be conscious. A universe without consciousness would be a valueless universe, or so many longtermists would argue. This yields two potential problems, one metaphysical and the other epistemological: (i) it might be that digital consciousness is not possible. It could very well be that functionalism in the philosophy of mind is false and the only type of matter that can give rise to conscious mental states is biological. (ii) Even if a form of functionalism is true, it could be that the *particular* digital beings that we create are not, in fact, conscious—they could be philosophical zombies that, as such, behave intelligently but have no qualitative inner life. If we were to send such beings into the universe to colonize the 10^23 stars within our future light cone, the longtermist project will have catastrophically failed. It thus matters *greatly* that we have robust tests for artificial consciousness to ensure that our digital progeny are actually conscious. One might expect longtermists to have addressed this issue at length, given what is at stake, but so far as I know they haven't. My tentative claim below will be that there may be no robust tests for consciousness in artificial systems—that is, no way to know with sufficient confidence that the colonization process, which may be irreversible once it is started, will in fact increase the total value in the universe. This places a giant question mark over the entire longtermist project: to colonize space, we need digital consciousness, but there may be no way to ever know if the digital beings create are genuinely conscious.

The rest of this chapter will elaborate on these claims. As the longtermist ideology increasingly shapes the world we live in, providing justificatory cover for billionaires like Elon Musk to pursue large projects to make humanity multi-planetary, it is crucial that longtermists address such issues. If there are flaws in the longtermist vision, then current projects guided by longtermism will have taken resources away from other cause areas like global poverty and animal welfare for nothing. In what follows, I will examine my three main theses, namely, that space colonization requires a Digital World, the Digital World would be profoundly different from our world, and confirming with a high degree of certainty that digital beings are conscious could be extremely difficult.

2. The Digital World

Let's begin by briefly looking at why creating a Digital World is almost certainly a necessary condition for space colonization beyond our solar system, if not within it. The main reason is simple: space travel poses serious psychological and physiological risks to biological human beings. The former concerns the fact that extended periods locked in the confines of incommodious spacecraft with a small number of companions could have detrimental psychological effects. As Campa et al. write, the "extreme, confined environments" of spacecraft "are likely to increase the incidence of potentially hazardous psychological effects due to confinement in a very unfamiliar environment and the loss of regular daily contact with the work and home environment on Earth" (Campa et al. 2019).

The latter arises from the effects of microgravity and space radiation, such as solar particle effects, solar wind, and galactic cosmic rays. On Earth, humans are protected from these hazards by our planet's atmosphere and magnetosphere, beyond which radiation doses can be roughly 100 times more than terrestrial exposure. This radiation can damage cells, cause cancer, and have deleterious cognitive effects (see Campa et al. 2019). With respect to microgravity, this has been linked to losses in muscle mass and bone mineral density, along with other forms of tissue atrophy (Li et al. 2018).

Technological advancements might enable us to devise new ways of protecting biological humans from these hazards. One can imagine such advancements making the colonization of our solar system possible. However, while the average distance between Earth and Mars is 140 million miles, the closest star to our Sun is roughly 25 trillion miles and the closest galaxy is 2.5 million light years, meaning that it would take 2.5 million years to reach it if one were traveling at the speed of light. The vastitude and inhospitable conditions of space make interstellar and intergalactic travel all but impossible for biological humans.

In contrast, these would not pose serious challenges for digital beings. As the longtermist Anders Sandberg writes, such beings would be

> ideally suited for colonising space and many other environments where biological humans require extensive life support. … Besides existing in a substrate-independent manner where they could be run on computers hardened for local conditions, emulations could be transmitted digitally across interplanetary distances. One of the largest obstacles of space colonisation is the enormous cost in time, energy and reaction mass needed for space travel: emulation technology [a reference to uploaded minds] would reduce this (Sandberg 2014).

Digital transmission would of course require that the destination had already been colonized, as transmission requires a transmitter and a receiver. But digital beings would also be ideally suited for traveling via spacecraft, perhaps propelled by solar sails, into deep space. Furthermore, they would be functionally immortal and hence able to travel for millions or billions of years to far-away galaxies.

Such considerations strongly imply that, to realize the longtermist aim of colonizing the accessible universe, we will need to create a Digital World. Even if biological humans figure out a way of reaching and living on Mars, it appears impossible that we could survive interstellar, much less intergalactic, distances as biology-based beings.

3. Beyond the Event Horizon

The creation of a Digital World, though, would mark a radical departure from the world in which we currently live. Verner Vinge's notion of the "technological Singularity" may be useful here, as the circumstances of existence in a Digital World may be hidden behind an "event horizon" that prevents us from knowing anything significant about them.[1] There are two aspects of this Digital World that make it inscrutable to us: one concerns the nature of the "beings" who would exist in it, while the other concerns the environments in which they are embedded. Focusing on the first, consider James Hughes' observation that digital beings would "be able to copy, share, and sell [their] memories, beliefs, skills, and experiences." They may have the ability to "selectively adopt personalities for specific purposes." Some might choose to publicly broadcast their inner thoughts and feelings, while "others will choose to spend a lot of time in someone else's life—like climbing into John Malkovich's head for weeks instead of 15 minutes at a time." The Digital World could thus inaugurate a radical new "post-individual" age, whereby bits and pieces of minds are duplicated, combined, modified, and discarded in an ever-fluctuating mosaic of ontological chaos. Hughes thus contends that "the most dramatic challenges to our social and philosophic world will probably come from hive minds and distributed selves," with collective intelligences arising from the merger of individuals and the boundaries between individuals coming to overlap in complicated, continually evolving ways (Hughes 2004).

This is one reason that calculations of future "people" are dubious: it is not at all clear that individuatable "people" would exist in a Digital World. Rather, such "beings" would more likely take the form of hive minds, collective intelligences, distributed selves, and protean entities whose boundaries are porous, thus enabling them to potentially redesign "themselves" from moment to moment. Perhaps each star would come to host a single "super-individual" around it, rather than trillions of individual people living in simulated worlds, thereby reducing the total number of "people" in the future by many orders of magnitude. If we can't begin to imagine what the Digital World might be like, then talk of the "expected value" of the far future is questionable. In other words, our decision-theoretic predicament is one of uncertainty rather than risk, where "uncertainty refers to situations under which either the outcomes and/or their probabilities of occurrences are unknown to the decision-maker," while "risk refers to decision-making situations under which all potential outcomes and their likelihood of occurrences are known to the decision-maker" (Park and Shapira 2017). Since we cannot imagine potential outcomes of a Digital World, we cannot assign probabilities to them. Our ignorance of the future is much deeper and more fundamental than longtermists seem to assume.

4. Digital Consciousness

Let’s be clear about what the longtermist vision would entail, given the above considerations: longtermists imagine our descendants colonizing space and building planet-sized computers on which to run virtual-reality worlds full of trillions of digital people. But since colonizing space requires digital beings, we can recognize two stages of the Digital World unfolding: in the first, digital beings would interact with the physical universe, control spaceships, and eventually build giant computers powered by Dyson swarms. In the second, the resulting simulations would be populated by digital beings with simulated bodies, interacting with each other in these virtual-reality environments. The first stage is necessary for the second, while the second is what would enable “astronomical” amounts of “value” to be generated.

None of this would matter though, if the resulting digital beings—especially those populating these vast computer simulations—were not conscious. Consider Hilary Greaves and William MacAskill’s conception of “human,” which they take “to refer both to *Homo sapiens* and to whatever descendants with at least comparable moral status we may have, even if those descendants are of a different species, and even if they are non-biological” (Greaves and MacAskill 2021). Since possessing a comparable moral status as such almost certainly requires such descendants to be conscious, the desideratum of consciousness is built-into their definition of “humanity.” Put differently, if our descendants were to lack consciousness (and hence comparable moral status), then “humanity” would no longer exist, which is just to say that we will have undergone extinction—a type of extinction that I call “normative extinction” (Torres 2023).[2] Or consider a scenario from Nick Bostrom “in which machine intelligence replaces biological intelligence but the machines are constructed in such a way that they lack consciousness (in the sense of phenomenal experience) […] The future might then be very wealthy and capable, yet in a relevant sense uninhabited: There would (arguably) be no morally relevant beings there to enjoy the wealth” (Bostrom 2013). Hence, it matters greatly that our digital descendants are capable of conscious experience—that there is “something it is like to be” them.

How, then, can we be sure that the digital beings that we become or create will in fact be conscious? Given that the entire longtermist vision of the far future hangs on this question, one might assume that a considerable amount of ink has been spilled reflecting on it. Yet, so far as I know, no longtermist has systematically examined it to date. The issue’s importance is underlined by the claim that once a colonization explosion has commenced, there may be no do-overs, which means that we will need a robust method of identifying consciousness in artificial systems before the *very first* generation of digital beings is launched into space. The stakes could not be higher, since if this initial generation of beings is not conscious, and if they proceed to replicate throughout space, an existential catastrophe will have occurred. By “existential catastrophe,” longtermists mean any failure to fulfill our “longterm potential” in the universe. Since fulfilling this “potential” requires flooding the universe with value, and hence consciousness, the avoidance of an existential catastrophe crucially depends on our ability to detect consciousness in non-biological entities before colonization begins.

One response could be to say that if the digital beings that colonize space are brain emulations that replicate the functional organization of one or more actual human brains, then they would almost certainly be conscious. There are several problems with this. First, we do not know if consciousness is an “organizational invariant,” that is, a property that emerges from systems with the right functional organization, independent of their material substrate (Chalmers 2011). Some form of “biological naturalism” could be true instead, which would mean that emulated brains would not be conscious, even if they were to convincingly reproduce human-level intelligent behaviors (see Pigliucci 2014). Second, even if the digital beings that initiate space colonization begin as brain emulations, it is entirely possible that they would quickly morph into new, alien minds as a result of the phenomena discussed in the previous section. If such beings have access to their code, they might also recursively self-improve, resulting in minds that are radically different than ours. We may, therefore, be much less confident that the resulting beings are conscious, even if they were conscious at some point in their earlier history. Third, whether or not the digital beings that colonize space are conscious, how sure can we be that the digital beings who end up populating the virtual-reality environments envisaged by longtermists will be conscious? Would these also be brain emulations? Again, this is unlikely because of the “post-individual” phenomena explored by Hughes, the possibility of self-improvement, and so on.

Hence, we would need a “consciousness test” that could enable us, with an extremely high degree of certitude, to affirm that the beings who populate the Digital World are in fact conscious. Have any such tests been proposed? In her book *Artificial You: AI and the Future of Your Mind*, Susan Schneider explores three potential tests for consciousness: the chip test, the AI Consciousness Test (or ACT Test), and a test based on Integrated Information Theory (ITT). I cannot dwell on the details of each test here; suffice it to say that, while these are probably the best tests thus far delineated, all have serious limitations. Consider the ACT Test, for instance. This is based on the idea that making sense of certain ideas or scenarios requires one to have had conscious experiences. Imagine that a digital system is asked about the possibility of an out-of-body experience, life after death, switching bodies with another system, or whether it would prefer being shut off for 300 years in the future or having been shut off for 300 years in the past (a time bias). Making sense of such questions would, Schneider argues, require the system to have experienced conscious states. In her words, "these scenarios would be exceedingly difficult to comprehend for an entity that had no conscious experience whatsoever." Yet, she adds that, while passing this test may be sufficient for a digital system to be considered conscious, it is not necessary, as one can imagine systems that fail but are nonetheless conscious (Schneider 2019, 57). Furthermore, this test might not apply to digital beings that achieve superintelligence, as they could find ways to cheat that we might not be able to detect. The ACT Test is thus limited to "some kinds of AIs, not all AIs" (Schneider 2019; for further criticisms, see Udell 2021).

The point is that we have no good way to determine whether the digital systems that we create are conscious. This problem is even more acute when we consider the possibility of hive minds, collective intelligences, distributed selves, and the protean entities mentioned earlier. Yet the entire longtermist project depends on us being overwhelmingly confident that digital beings would in fact be conscious—indeed, it requires not just that we have a test for the first generation of digital beings, but that these digital beings have tests of their own to ensure that the entities populating the computer simulations they build throughout the accessible universe are also conscious. The lack of a robust test for artificial consciousness thus places a giant question mark over the entire longtermist project.

5. Conclusion

In conclusion, fulfilling the longtermist project requires the colonization of space beyond our solar system, which in turn requires the creation of a Digital World. This Digital World will likely be so different from our current world that we are in a position of decision-theoretic uncertainty rather than risk. Furthermore, since the entire longtermist vision depends on the possibility of digital consciousness, if non-functionalist theories like biological naturalism are true—and they might be—then this vision cannot be fulfilled. But even if functionalism were true, we would still encounter the epistemological problem of being able to determine with a very high degree of certainty that the particular digital beings that we create are in fact conscious, and that the digital beings that they create in giant computer simulations are also conscious, and so on. The challenges that such considerations pose to the longtermist project are, I believe, much more formidable than longtermists have previously recognized.

Bibliography

Adams, Fred C. "Long-Term Astrophysical." *Global Catastrophic Risks*, 2008, 33.

Boeyens, Jan CA, and Demetrius C Levendis. “Elements of Cosmography.” *Number Theory and the Periodicity of Matter*, 2008, 183–208.

Bostrom, Nick. “Astronomical Waste: The Opportunity Cost of Delayed Technological Development.” *Utilitas* 15, no. 3 (2003): 308–14.

———. “Existential Risk Prevention as Global Priority.” *Global Policy* 4, no. 1 (2013): 15–31.

———. *Superintelligence*. Oxford University Press, 2014.

Campa, Riccardo, Konrad Szocik, and Martin Braddock. “Why Space Colonization Will Be Fully Automated.” *Technological Forecasting and Social Change* 143 (2019): 162–71.

Chalmers, David J. “A Computational Foundation for the Study of Cognition.” *Journal of Cognitive Science* 12, no. 4 (2011): 325–59.

Greaves, Hilary, and William MacAskill. “The Case for Strong Longtermism.” *GPI Working*, 2019.

Greaves, Hilary, and William MacAskill. “The Case for Strong Longtermism.” *GPI Working*, 2021.

Hughes, James. “The Illusiveness of Immortality.” *Death and Anti-Death*. 3 (2005).

Li, Kai, Chao Yang, Hongyu Zhang, Feng Wu, Hailong Wang, Hongqing Cao, Zihan Xu, Bai Ding, Yinghui Li, and Zhongquan Dai. “Screening and Identification of Novel Mechanoresponsive MicroRNAs in Rat Femur under Simulated Microgravity.” *Acta Astronautica* 153 (2018): 166–73.

Newberry, Toby. “How Many Lives Does the Future Hold.” *Global Priorities Institute Technical Report*, 2021.

Ord, Toby. “The Edges of Our Universe.” *ArXiv Preprint ArXiv:2104.01191*, 2021.

Park, K Francis, and Zur Shapira. “Risk and Uncertainty.” *The Palgrave Encyclopedia of Strategic Management*, 2017.

Pigliucci, Massimo. “Mind Uploading: A Philosophical Counter-Analysis.” *Intelligence Unbound: Future of Uploaded and Machine Minds, The*, 2014, 119–30.

Sandberg, Anders. “Ethics of Brain Emulations.” *Journal of Experimental & Theoretical Artificial Intelligence* 26, no. 3 (2014): 439–57.

Schneider, Susan. *Artificial You: AI and the Future of Your Mind*. Princeton University Press, 2019.

Torres, Émile P. *Human Extinction: A History of the Science and Ethics of Annihilation*. Routledge, 2023.

Udell, David B. “Susan Schneider’s Proposed Tests for AI Consciousness: Promising but Flawed.” *Journal of Consciousness Studies* 28, no. 5–6 (2021): 121–44.

Yudkowsky, Eliezer. 2007. Three Major Singularity Schools. Machine Intelligence Research Institute. https://intelligence.org/2007/09/30/three-major-singularity-schools/.

[1] See Yudkowsky 2007.

[2] Normative extinction would occur if *Homo sapiens* has successors, but these successors have lost some normatively important capacity that excludes them from the extension of "humanity," meaning that "humanity" no longer exists. It contrasts with demographic, phyletic, terminal, final, and premature extinction, all of which have their own unique ethical and evaluative implications. See Torres 2023.