Bartleby's Gale: Upon the Simulation Argument

U. Kanad Chakrabarti*

March 26, 2021

The Simulation Argument posed by philosopher Nick Bostrom goes something like this: there may come a point in humanity's development, if we haven't by then already gone extinct, when our successors (posthumans) will have sufficient computing power to simulate arbitrarily complex phenomena (much as scientists currently simulate systems such as climate change or the global economy).[1] A posthuman civilization may find it interesting to simulate its human ancestors. By way of analogy, current humans watch historical dramas, play videogames and look at figurative paintings. These can be seen as species of simulations: viewers imagine what it might be like to be within the depicted situation. Although posthumans may choose not to ancestor-simulate (for ethical reasons, i.e. simulation may cause needless suffering to conscious simulated agents), we — currently-existing humans — without knowing anything about posthumans' ethical or decision-making framework, cannot exclude the possibility that we are ourselves simulated.

The Argument is interesting because it allows several disparate ideas to be jointly discussed: existential risk, the function of museums, and media/gaming theory.[2] This essay considers the existential risk element from an aspect not often treated in Argument-related literature; that is, in light of writing on postcolonial critical theory.

Extinction is an end state, and there are many ways to get there. A more general term is "existential catastrophe", which includes the outright elimination of all human life, but also allows for other scenarios. For instance: a vastly reduced population that

*Note: This essay, currently in draft form, will be published in late 2021 in *Shifter Magazine*, https://shifter-magazine.com/. The author wishes to thank the editors, Avi Alpert and Rit Premnath, for their helpful comments.

[1]The Simulation Argument is more precisely stated in Nick Bostrom, "Are You Living in a Computer Simulation?" in *Philosophical Quarterly* 53, no. 211 (2003): 2-3, https://www.simulation-argument.com/simulation.pdf.

[2]Simulation of human lives has shown up repeatedly in cinema, for instance in Rainer Werner Fassbinder's *World on a Wire* (1973), Andrei Tarkovsky's *Solaris* (1972), based upon Stanislaw Lem's 1961 novel, or more recently in *The Matrix* (1999). For points related to videogames, see Kanad Chakrabarti, "The Game Engine as Transhumanist Sandbox", in *Flat Journal*, 2020, https://flatjournal.com/work/the-game-engine-as-transhumanist-sandbox/

leaves society in a primitive state, indefinitely, as a result of a civilizational collapse — a millennial stagnation. Or, civilization, while maintaining present or higher levels of technological sophistication, actually results in lived experiences that are much worse for many people—an *unrecoverable dystopia*. [3]

In terms of specific sources of risk: thermonuclear war would probably not prompt global extinction, in part because the Southern Hemisphere is nuclear-free, and countries like New Zealand are sufficiently isolated to survive a nuclear winter without a total loss of human life. [4] Nor does climate change seem likely to cause extinction, although it would clearly result in enormous suffering. [5] A higher-probability existential risk comes from a malevolent AI (possibly human-assisted), which would seek to consume all available resources to maximize its encoded reward function.[6] A pandemic, or certain types of nanotechnology, might also be engineered, by humans or by a malevolent AI, as weapons.[7]

Misaligned AI is estimated to pose a 10% chance of leading to extinction over the next century, making it the main contributor to the total risk from anthropogenic causes of extinction (estimated at 17%). By comparison, non-anthropogenic events, such as a supervolcano eruption or asteroid strike, are estimated at a probability of 1-in-10,000.[8] The discussion above has liberally used "we," "us", "humanity," and so forth, but, as argued by Donna Haraway and others, it is abundantly clear that no such singular position exists.[9] Hence, how to weigh (a), the urgent and geographically uneven needs of present and nearby generations, against (b), the risk that all future generations could be extinguished? In his 2015 Oxford Union address, philosopher Derek Parfit provided a valuation framework. He pointed out that we should not differentiate between helping a starving child in front of us and helping one on another continent.[10] Empathy with respect to geographically distant suffering is arguably analogous to concern for human suffering in the distant future.[11] If we believe that human lives—now and in the fu-

[3] Toby Ord, *The Precipice: Existential Risk and the Future of Humanity* (New York: Hachette, 2020): 36-42, 153-161.

[4] Ord, *The Precipice*: 98-100.

[5] Ord, *The Precipice*: 102-112 and related notes.

[6] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014): chapter 8.

[7] Ord, *The Precipice*: 124-138, 158-161.

[8] Ord, *The Precipice*: 169.

[9] Donna Haraway, *A Manifesto for Cyborgs: Science, Technology, and Socialist Feminism in the 1980s* (1985): 21, https://sites.evergreen.edu/politicalshakespeares/wp-content/uploads/sites/226/2015/12/Haraway-Cyborg-Manifesto-2.pdf.

[10] "Derek Parfit – Full Address", (YouTube, posted 10 October 2015), https://www.youtube.com/watch?v=xTUrwO9-B_I.

[11] This statement is subject to certain assumptions, such as whether the value of future lives is qualitatively or quantitatively the same as that of current lives. A related question is: does it make sense to speak of extinction today as being bad for unborn future generations? Those generations, never having been born, will never experience the suffering, if any, associated with extinction. For more on Population Ethics, see Ord, *The Precipice*, 46-48, Appendices A and B.

ture—will likely be composed of more flourishing than misery, then a human extinction that eliminates the possibility of quattuordecillions of un-lived human life-years, will also mean the non-realization of their potential wellbeing. [12]

Walter Benjamin, writing his *Theses on the Philosophy of History*, allegorizes the protagonist of a Klee print (*Angelus Novus*, 1920) in flight amidst a maelstrom of progress. [13] The Angel is surrounded by the ruins of history. In this late work, from 1940, Benjamin rejected the very notion that progress in human existence, individually or collectively, occurs over "homogeneous, empty time": the Judeo-Christian calendar's time of years, decades, or centuries. [14] This calendrical notion of progress, subsequently inherited into both Enlightenment- and Marxist-inflected theories of history, was understandably rendered incredible in 1930s Europe. Instead, Benjamin, on the eve of his suicide, would revert to the eschatological mysteries of Judaism: "For every second of time was the strait gate through which the Messiah might enter."[15]

During the Cold War, progress for humanity-as-a-whole became even more illusory as newly-independent states, for a variety of entangled reasons, including foreign interference and corrupted elites, found that decolonization translated neither to emancipation nor equality. [16] Thus the scene Benjamin describes in the Ninth Thesis may be read today as the ruins of the Global South: the sinking islands of the Pacific, or the unbearably hot cities and flooding deltas of the tropics. Other violences abound too: Israel's splintering and panoptic settler colonialism; or, the hybrids of extractive enterprises and war-machines that flourish in relatively new, often arbitrarily demarcated, states that are still trying to establish institutional stability and basic political consensus.[17]

In analyzing this tension between specific human experiences, particularly of the violated and voiceless, versus the broadly articulated *telos* of the species, one is faced with what McKenzie Wark describes as the post-1960s preference for "the local, the different, the marginal or the specific to the abstract, the global or the universal."[18] Wark's diagnosis (delivered in an ecological-ethnographic context) mirrors another case of competitive coexistence, this time in post-human(-ism) studies, an awkward portmanteau that cov-

[12] A range of 10^{16} to 10^{54} human life-years (of potential future wellbeing) is given in Nick Bostrom, "Existential Risk Prevention as a Global Priority", in *Global Policy* 4, no. 1 (February 2013): 18, https://www.existential-risk.org/concept.pdf

[13] Walter Benjamin, "Theses on the Philosophy of History", in *Illuminations* (New York: Harcourt Brace Jovanovich, 1968): 257-258.

[14] Benjamin, "Theses": 260-261.

[15] Benjamin, "Theses": 264.

[16] One diagnosis of possible causes for the unachieved promise of postcolonial flourishing is offered by David Scott, in this interview with Stuart Hall: https://bombmagazine.org/articles/david-scott/

[17] See Eyal Weizman, "The Politics of Verticality," (23 April 2002), https://www.opendemocracy.net/en/article_801jsp/ . Also, J.-A.Mbembé and Libby Meintjes, "Necropolitics", in *Public Culture* 15, no. 1 (Winter 2003): 11-40, https://www.muse.jhu.edu/article/39984.

[18] See McKenzie Wark, "Friction" in *Public Seminar* (5 September 2016), https://publicseminar.org/2016/09/friction/. See also McKenzie Wark, *Molecular Red* (London: Verso Books, 2016).

ers thinking about the future of human society and humanism-as-discipline. Within post-humanist studies, one finds distinctly pro- and anti-science strands in worldviews, languages, arguments, and objectives. [19] In the work of Rosi Braidotti (to take an example), one perceives that the failure of the Enlightenment (whether incarnated in liberal-technocratic, Communist, or Modernist forms) as an emancipatory project may have given way to shibboleth: reflex-like suspicion of any scientific or philosophical theory that rests on a nowhere/nowhen perspective, thereby making any discussion of species-level futures fraught.[20]

In any event, such academic-epistemic competition in perspectives seems unhelpful and may benefit from a broader historical vantage point. For example, consider Russian Cosmism, a Tsarist-era philosophy that attempted to reconcile the suffering of present-day lived experience with a longer, more œcumenical and technologically-positive narrative. Two Cosmist assertions are relevant: (a) universal justice will not be served until death itself, as the root of all suffering, is eliminated, leading to an universal resurrection; and (b) humanity has an obligation and imperative to colonize the known universe.[21]

As Wark writes in *Molecular Red*, Cosmist ideas root a post-Soviet lineage of relations between the laboring human and the machinic/ecological context, and ultimately can be traced through to the writings of Donna Haraway and Kim Stanley Robinson. Despite their different starting points, these speculative cyborg-feminist theories and astro-ecological science fictions, respectively, can be seen as consistent with humanity's expanding "moral community," which must, in time, include natural systems, non-human creatures, and digital entities. [22] As Haraway writes: "In the fraying of identities and in the reflexive strategies for constructing them, the possibility opens up for weaving something other than a shroud for the day after the apocalypse that so prophetically ends salvation history."[23]

So perhaps the Benjaminian Angel, updated to the contemporary, stands in a current

[19] See Rosi Braidotti, "A Theoretical Framework for the Critical Posthumanities" in *Theory, Culture & Society*, 36, no. 6 (2019) https://journals.sagepub.com/doi/full/10.1177/0263276418771486. Also, this talk by Braidotti clearly illustrates the difference in perspective and motivations: https://youtu.be/gNJPR78DptA

[20] See Thomas Moynihan, "Our Visions of the Future Determine Our Society Today", in *Palladium*, 2020, https://palladiummag.com/2020/05/11/our-visions-of-the-future-determine-our-society-today/

[21] See Kanad Chakrabarti, "Silenus' Cup, Drained by AI" in *Public Seminar*, 2019, https://publicseminar.org/essays/silenus-cup-drained-by-ai/

[22] For moral community, see Ord, *The Precipice*: 18. See also Philip Højme, "Whose Survival? A Critical Engagement with the Notion of Existential Risk," in *Scientia et Fides* 7, no. 2 (2019), http://dx.doi.org/10.12775/SetF.2019.016, and Benjamin Bratton, "Outing Artificial Intelligence: Reckoning with Turing Tests", in *Alleys of Your Mind: Augmented Intelligence and Its Traumas*, Matteo Pasquinelli (ed.) (Centre for Digital Cultures, Leuphana University of Luneburg, 2015), https://meson.press/books/alleys-of-your-mind/: 72-73. The idea of extending certain moral and ethical constructs to potentially sentient/sapient digital entities comes out of a conversation and unpublished notes with/from Nick Bostrom and his collaborators.

[23] Haraway, *A Manifesto*: 20-21.

of history where human subjectivity is merely one temporality amongst others, without privilege or preference. After all, digital entities—avatars or emulations with agency far beyond that of today's bots—might experience time a thousandfold faster than humans and see us as doddering nonagenarians.[24]At the other extreme, non-human, systemic actors (such as the planetary life-support system), with increasing claims to legal and ethical personhood, operate on time scales that span many human generations.[25]

Extending Benjamin's allegory, perhaps we can reimagine the Angel as *Trivia*, who was a manifestation of the Roman goddess Diana and guarded the three-way crossroads. This would let us add another path to Haraway's synthesis of cybernetic and Marxist worldviews: a radical future that freely jettisons all that is inessential about the human. This re-conception, sometimes termed *inhumanism*, comes in several flavors: one version posits the human as an information-processing entity capable of learning and teaching its peers. In this view, articulated by Peter Wolfendale, the uniqueness of the human lies in language, tool-use, the ability to mentally simulate environments, and neural (or algorithmic) plasticity.[26] Any entity, irrespective of substrate (physical manifestation), with these capabilities, is arguably human. Inhumanism thus can be seen to build upon the project of Haraway, Bostrom *et al.*, by striving to eliminate carbon bias—current humans' prejudice in favor of biological, fleshy substrates that look like us, which in some cultures is partially inherited from quasi-religious commitments centered on Adamic origin myths.[27].

In what can be seen as a limiting case, the posthuman, having done away with the corporeal, libidinal and carnal self, and therefore with the currently obvious markers of identity, may go on to incorporate everything around it, in what would, in effect, be the fulfillment of Cosmism's second imperative. One possibility for posthuman's "common task" could be a computational megastructure, built from planets and asteroids, surrounding the sun (or any star) and harvesting all the solar energy currently lost to space.[28]

[24]Nick Bostrom and Eliezer Yudkowsky, "The Ethics of Artificial Intelligence" (draft, 2011): 11-12, https://nickbostrom.com/ethics/artificial-intelligence.pdf.

[25]See Vincent Ialenti, *Deep Time Reckoning* (Cambridge, MA: The MIT Press, 2020), which covers the canonical example of long-term nuclear waste storage. How do we think about danger that persists for up to 100,000 years?

[26]Peter Wolfendale, "The Reformatting of *Homo Sapiens*", in Angelaki, 24 (2019): 1, 55-66, https://doi.org/10.1080/0969725X.2019.1568733 . For another version, that specifically takes aim at un-scientific, Marxist-inflected definitions of the human, and suggests as an alternative the human as an engineered thing that continually re-makes itself, see Reza Negarestani, "The Labor of the Inhuman", parts 1 & 2, in *e-flux*, 52, February 2014,https://www.e-flux.com/journal/52/59920/the-labor-of-the-inhuman-part-i-human/ and http://worker01.e-flux.com/pdf/article_8979538.pdf .

[27]Reza Negarestani's project, in significant part, is to free our understanding of intelligence from the detritus of religion and the body: Reza Negarestani, *Intelligence and Spirit* (Falmouth: Urbanomic, 2018): 57-62, 504-507

[28]See Thomas Moynihan "Can Intelligence Escape its Terrestrial Past?: Anticipations of Existential Catastrophe & Existential Hope from Haldane to Ćirković", in *Cosmos and History: The Journal of Natural and Social Philosophy*, Vol. 16, no. 1, 2020: 96-98.

The notion of a great computer in the sky brings us to the major existential risk not discussed above, known as the Fermi Paradox. It goes as follows: given the number of stars, many much older than the Earth, associated habitable planets in the universe, and (presumably likely) energy-gathering megastructures, intuition and mathematical reckoning suggest that the skies should be visibly teeming with life. Yet, for almost a century, we have detected nothing. Is this because, by the time a civilization advances to the point where it could contact us, it has already destroyed itself? Or do we inhabit a Hobbesian universe where each species alternates between being hunter and prey, and hence finds it safest to remain invisible.[29]

One solution to the Fermi Paradox may be that, while sufficiently-advanced civilizations have, in fact, already created megastructures, these planetary-scale structures have, owing to their material composition, become indistinguishable (to us) from their "natural" environment. [30] "They" are, as it were, hiding in plain sight. A fictional example comes from Stanislaw Lem's *Solaris*: a planet-wide Ocean that exhibits signs of unified consciousness: that is morphologically neither organic nor inorganic; but that can activate planetary tides powerful enough to stabilize Solaris' orbit around its two suns.[31] As if in keeping with Edouard Glissant's "right to opacity", the being remains unresponsive to the cosmonauts' repeated probes, until they perpetrate an act of colonial violence by firing an intense blast of radiation at the Ocean. [32] Subsequently, the cosmonauts become haunted by terrifyingly familiar simulations, like waking dreams, fully-rendered, 3-D and tactile, drawn from their own subconscious, guilt-ridden memories. Similar to the psychological trauma that Frantz Fanon would diagnose amongst French soldiers in Algeria, we find that the brutality inflicted upon the Other tends to boomerang back upon the colonizer. [33]

Returning briefly to *why* a computational megastructure, of vast yet bounded intelligence, might exist at all: it could be an entity whose *raison d'être* is to run simulations of worlds, one of which, under the Simulation Argument, we might currently inhabit.[34]. Such an idea, albeit clad in the garb of contemporary futurology, has religious precedent. For instance, in his exegesis tracing Aristotle's thought on creation and contingency as transmitted through the Jewish and Islamic traditions, Giorgio Agamben writes: "The

[29] A comprehensive survey of the Fermi Paradox can be found in Milan Ćirković *The Great Silence*, (Oxford: Oxford University Press, 2018). There is a plausible resolution of the Paradox in Anders Sandberg, Eric Drexler and Toby Ord, "Dissolving the Fermi Paradox" (2018), https://arxiv.org/abs/1806.02404. See also *Collapse V* (Falmouth: Urbanomic Press, 2012), particularly in the contributions of Nick Bostrom, Milan Ćirković, Jack Cohen and Ian Stewart.

[30] Ćirković *Silence*, 134-137.

[31] Stanislaw Lem, *Solaris*, (London: Faber and Faber, 1970), 18-23.

[32] Edouard Glissant, *Poetics of Relation*, (Ann Arbor: University of Michigan Press, 2010): 121.

[33] Frantz Fanon, *The Wretched of the Earth*, (New York: Grove Press, 1963-2004): see Chapter V: Series A, Cases 4 & 5.

[34] For the intersection of the Simulation Argument with philosophical and theological perspectives on creation, see Eric Steinhart, "Theological Implications of the Simulation Argument", in *Ars Disputandi*, No. 10 (2010), https://www.tandfonline.com/doi/abs/10.1080/15665399.2010.10820012

prince of the *falasifa* himself, Avicenna, conceived of the creation of the world as an act in which the divine intelligence thinks itself".[35]Another example is the *Avatamsaka (Flower Garden) Sūtra* in East Asian Buddhism.

In any event, the distant future will hopefully be unrecognizable to us. It may be populated by being(s) that look and act nothing like us, even if some of them are our descendants.[36] In one hopeful possibility, having eradicated racial, gender, or species bias, the future may be filled with literally unimaginable promise. Yet we are unlikely to get there if we cannot transit safely through the gale of present emergencies, the omnipresent "now-time", much of which seems to stem from tribalism, greed, and violence, indelibly encoded human tendencies common to postcolonial state and colonizer alike.[37] More worryingly, there is a persistent and unavoidable risk that current decisions, even if taken in good faith, will be inadequate and unworthy of the future, imbued as they would be with normative and ethical prejudices arising from parochial histories, religions and theories - inheritance of a flawed species living in all-too-specific places and times.

[35]Giorgio Agamben, "Bartleby, On Contingency", in *Potentialities*, (Stanford: Stanford University Press, 1999): 246.

[36]Nora Khan, "Towards a Poetics of Artificial Superintelligence" in *after us*, Issue 1, No. 2, 2015, http://www.aft3r.us/towards-a-poetics-of-artificial-superintelligence/.

[37]Benjamin develops the idea of *jetztzeit*, a "now-time" or immediate moment, albeit in a mystical-revolutionary, Messianic sense, in the Fourteenth Thesis, see Benjamin, "Theses": 261.