Naudé, Wim

Working Paper

Extraterrestrial Artificial Intelligence: The Final Existential Risk?

IZA Discussion Papers, No. 15924

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Naudé, Wim (2023) : Extraterrestrial Artificial Intelligence: The Final Existential Risk?, IZA Discussion Papers, No. 15924, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/272551

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15924

Extraterrestrial Artificial Intelligence: The Final Existential Risk?

Wim Naudé

FEBRUARY 2023

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15924

Extraterrestrial Artificial Intelligence: The Final Existential Risk?

Wim Naudé
RWTH Aachen University and IZA

FEBRUARY 2023

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Extraterrestrial Artificial Intelligence: The Final Existential Risk?

The possibility that artificial extraterrestrial intelligence poses an existential threat to humanity is neglected. It is also the case in economics, where both AI existential risks and the potential long-term consequences of an AGI are neglected. This paper presents a thought experiment to address these lacunas. It is argued that it is likely that any advanced extraterrestrial civilization that we may encounter will be an AGI, and such an AGI will pose an existential risk. Two arguments are advanced for why this is the case. One draws on the Dark Forest Hypothesis and another on the Galactic Colonization Imperative. Three implications for how we govern AI and insure against potential existential risks follow. These are (i) accelerating the development of AI as a precautionary step; (ii) maintaining economic growth until we attain the wealth and technological levels to create AGI and expand into the galaxy; and (iii) putting more research and practical effort into solving the Fermi Paradox. Several areas where economists can contribute to these three implications are identified.

JEL Classification: O40, O33, D01, D64

Keywords: technology, artificial intelligence, existential risk, Fermi paradox, Grabby Aliens

Corresponding author:
Wim Naudé
Technology and Innovation Management (TIM)
RWTH Aachen University
Kackertstraße 7
52072 Aachen
Germany

E-mail: naude@time.rwth-aachen.de

1 Introduction

Listing the many ways humanity can meet its end has become popular. Bostrom (2014, 2013), MacAskill (2022), Noy and Uher (2022), Ord (2020) and Turchin and Denkenberger (2020) and are amongst scholars who have recently fuelled the apocalyptic zeitgeist by warning about the existential risks that humanity faces. The term "existential risk" was first used by Bostrom (2002), who defined it as a risk that an event "would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential" (Bostrom, 2002, p.2).

Of all the existential risks, one of the most feared has come to be an unaligned Artificial General Intelligence (AGI) (or Artificial Super-Intelligence (ASI)). For instance, according to Noy and Uher (2022, p.498) artificial intelligence (AI) "poses the highest global catastrophic and existential risk to humanity [including from] solar-flares and space weather, engineered and natural pandemics, and super-volcanic eruptions." According to Turchin and Denkenberger (2020, p.148) AI is even "millions of times more powerful than nuclear weapons." No wonder a recent headline exclaimed that "A third of scientists working on AI say it could cause global disaster" (Hsu, 2022).

Why would AI pose an existential risk? The risk emanates from the future capabilities and values of AI. The capability claim is that AI may, in future, even if the chance is slight, scale up from its present "narrow" (less intelligent) state to become an AGI/ASI and then cause significant damage to humanity because of its intelligence - either intentionally or unintentionally. Turchin and Denkenberger (2020) list two dozen ways this could happen. The value claim is that AI's values may not align with humanity's - there is an alignment problem (Sotala, 2018; Barrett and Baum, 2017).

Given that it cannot be ruled out that an AGI or superintelligence will come into being with the non-trivial probability of causing the extinction of humanity, many scientists now think

that "The consequences for humanity are so large that even if there is only a small chance of it happening [...] it is still urgent that we work now to understand it and to guide it in a positive direction"(Omohundro, 2008, p.5). The work on "guiding AI in a positive direction" is aimed at solving the alignment problem.[1]

At the time of writing, 2023, the alignment problem has NOT been solved. Moreover, efforts to create an "AI Nanny" to box AI in until the alignment problem is solved, has not yet borne fruit. According to Goertzel (2012, p.102) "nobody has any idea how to do such a thing, and it seems well beyond the scope of current or near-future science and engineering."

While AI alignment and AI ethics studies flourish as scientists scramble to avert a catastrophe, the larger - and virtually uncontrollable - AGI existential risk has gone largely unnoticed.[2] This is the possibility that it is not only human-made AGI that poses an existential risk to humanity, but ultimately an alien, artificial extraterrestrial intelligence (ETI). Even if we succeed in solving the alignment problem, there is still the danger of an artificial ETI - and there is no obvious way for us to align its interests with that of humanity.

This paper presents an extended thought experiment to contribute to addressing this neglected existential risk, and to do so mainly from an economics point of view. The added advantage of this is to contribute to the economics of AI. So far economists have neglected the "risk of an AI-induced existential catastrophe" (Trammell and Korinek, 2020, pp.53-54). The hypotheses underlying this thought experiment, such as that an AGI is possible and that extraterrestrial intelligence may exist, and their possible implications, are based not on science fiction, but on our current best speculations about the future of AI and our position in the universe - as reflected in a growing scientific literature.

[1]Under the headings AI alignment and AI ethics scientists are working to make AI systems' goals or utility functions subservient to that of humans (Hauer, 2022). They want to ensure that AI "benefit humans" (Kirchner et al., 2022, p.1). Note that this is a very human-centric agenda, based on human exceptionalism (Murphy, 2022).

[2]Ord (2020) in an exhaustive evaluation of existential risks, spends only one paragraph considering - and dismissing- the risk from an alien AGI.

The core of this paper is to outline two arguments why an extraterrestrial AGI poses an existential threat. One draws on the Dark Forest Hypothesis and another on the Galactic Colonization Imperative. Three implications for how we govern AI and insure against potential existential risks are discussed. These are (i) accelerating the development of AI as a precautionary step; (ii) maintaining economic growth until we attain the wealth and technological levels to create AGI and expand into the galaxy; and (iii) putting more research and practical effort into solving the Fermi Paradox. Several areas where economists can contribute to these three implications are identified.

The rest of the paper proceeds as follows. In section 2 the rise of galactic AGI is discussed, drawing on current predictions on the scaling-up of Deep Learning. This section concludes that it is likely that any advanced extraterrestrial civilization that we may encounter will be an AGI. Section 3 outlines the existential risks posed by an extraterrestrial AI. Section 4 concludes by drawing out three implications and noting where economists can contribute.

2 The Rise of Galactic AGI

If the expected future development trajectory of AI on planet Earth is not in any way special in the universe - invoking the Copernican principle - then if the ultimate outcome of our evolution is to lead to the emergence of a galactic AI, then we should expect the same to happen elsewhere in the cosmos.

This expected future development trajectory of our AI can, from the best of current accounts, be described as follows. First, simple, narrow-AI, based on Deep Learning (DL) as is now the case, starts to scale-up.[3] This leads inter alia to the development of robotic brains by 2024, and *Seed-AI*. Seed-AI is defined as"an AI designed for self-understanding, self-modification,

[3]According to the Scaling Hypothesis, DL will eventually scale to the level of human intelligence, and even further (Englander, 2021).

and recursive self-improvement" (Yudkowsky, 2007, p.96).

Second, once an AI system gains the ability of recursive self-improvement, the era of narrow-AI is over, and it will exponentially improve to "ultraintelligent" levels (Good, 1965). Once a certain threshold of intelligence is reached, there could be what is described as a sudden jump, or hard take-off[4] (or "foom") (Barnett, 2020), after which AI would very rapidly become super-intelligent - an ASI[5].

Once AI achieves human-level intelligence it will be "followed by an explosion to ever-greater levels of intelligence, as each generation of machines creates more intelligent machines in turn. This intelligence explosion is labelled the 'singularity'" (Chalmers, 2010, p.7). After the hard take-off, the subsequent intelligence explosion will occur so rapidly, that it will appear as if the super-intelligence appeared without warning (Bostrom, 2006; Yudkowsky, 2008). It is expected that within forty years after reaching superintelligence levels and igniting a Singularity (intelligence explosion) the ASI will become a *Singleton* (Turchin and Denkenberger, 2020). A *Singleton* is "a world order in which there is a single decision-making agency at the highest level (Bostrom, 2006, p.48).

Eventually, the Singleton would become a Galactic AI, after some undetermined time, perhaps millions of years. This Galactic AI could colonize the galaxy and universe. As we will explain below, this may be inevitable.

If this process will play out on Earth, it is also possible to play out elsewhere in the universe. Moreover, there nay be many extraterrestrial civilizations that have hundred of millions of years' head-start to us. They may already be actively colonizing the universe. Some scientists today believe that we may meet other such expanding galactic explorers within 500 million years.

[4]For a brief overview of the debate on whether a hard or soft take-off in AI is more or less likely, see Barnett (2020) and the "Foom" debate between Robin Hanson and Eliezer Yudkowsky (Hanson and Yudkowsky, 2013).

[5]Estimates put this with a 50% probability to be happen by 2050 (Cotra, 2020).

But we have not yet. Indeed we have found no evidence of any life elsewhere in the universe. Does this suggest that the possibility of an alien AGI is far-fetched and to be discounted? On the contrary, as will be argued in the next section, the apparent absence so far of any evidence of an alien AGI may be a confirmation of ts potential existential threat.

3 The Risk of Extraterrestrial AI

Although there is no evidence at present for any alien civilizations, statistically the odds of human civilization being singular is almost vanishingly small (Drake, 1965). There are around ≈2 trillion galaxies in the universe (Conselice et al., 2016) each with more than 100 billion stars each - most of whom likely have planets (Cassan et al., 2012). The number of terrestrial planets in the universe that circle Sun-like stars is huge - around $\approx 2 \times 10^{19}$ with another $\approx 7 \times 10^{20}$ (Zackrisson et al., 2016) estimated to be around M-dwarf stars. And 22% of Sun-like stars may have Earth-size planets in their habitable (where liquid water can exist) zones(Petigura et al., 2013).

Even if on only 1% of these intelligent life arises, the universe would host billions of alien civilizations. One estimate is that there is around 36 Communicating Extra-Terrestrial Intelligent (CETI) civilizations in the Milky Way galaxy (Westby and Conselice, 2020). These alien civilizations, if sufficiently advanced, are likely to be ASIs - as was argued in the previous section[6] (Rees, 2021; Gale et al., 2020; Shostak, 2018, 2021; De Visscher, 2020).

Why would an alien ASI pose a threat to Earth? Economic reasoning supported by game theoretic analysis offers two broad and interrelated reasons.

[6]These post-Singularity alien AI civilizations may be too advanced for humans to detect - they may for instance use quantum entanglement to communicate (and not radio waves), or compress their communication signals that it would be indistinguishable (for earthlings) from noise (Gale et al., 2020; Bennett, 2021).

3.1 The Dark Forest

The first is the Dark Forest Hypothesis. It takes its label from the science fiction novel *The Dark Forest* by Cixin Liu. The Dark Forest Hypothesis (DFH) offers an explanation for the Fermi Paradox, which arose out of the question that physicist Enrico Fermi posed in 1950 "where is everybody?" referring to the absence of any evidence of an alien civilization in the universe. The Fermi Paradox, or rather Fermi's Question, which was more formally set out by Hart (1975) is based on the observation that given the likelihood of intelligent civilizations in the universe (as described above) and the age of the universe (13,8 billion years) we would be now have encountered evidence for their existence.[7] The fact that we have not yet, and that there is a "Great Silence" (Cirković and Vukotić, 2008) requires "some special explanation" (Gray, 2015, p.196).

Many explanations - more than seventy-five - have been proposed for the Fermi Paradox. A full discussion falls outside the scope of this paper; the interested reader is referred to Webb (2015). For present purposes though, the DFH explains the Fermi Paradox by postulating that it will be in the self-interest of any civilization to conceal its existence, lest it be exterminated by another, far more advanced civilization. According to (Liu, 2008)

> "The universe is a dark forest. Every civilization is an armed hunter stalking through the trees like a ghost, gently pushing aside branches that block the path and trying to tread without sound. Even breathing is done with care. The hunter has to be careful, because everywhere in the forest are stealthy hunters like him. If he finds another life - another hunter, angel, or a demon, a delicate infant to tottering old man, a fairy or demigod—there's only one thing he can do: open

[7] For instance, using self-reproducing intelligent starprobes travelling at 1/10th the speed of light, the entire Milky Way Galaxy could be traversed in 500,000 years (Valdes and Freitas Jr, 1980). Such starprobes, as way to traverse the universe, were proposed by Game Theory co-founder, John von Neumann (Von Neumann, 1966) hence labelled Von Neumann Probes. "From a technological point of view, there seems to be no obstacle to the ultimate terrestrial construction of Von Neumann probes" (Matloff, 2022, p.206).

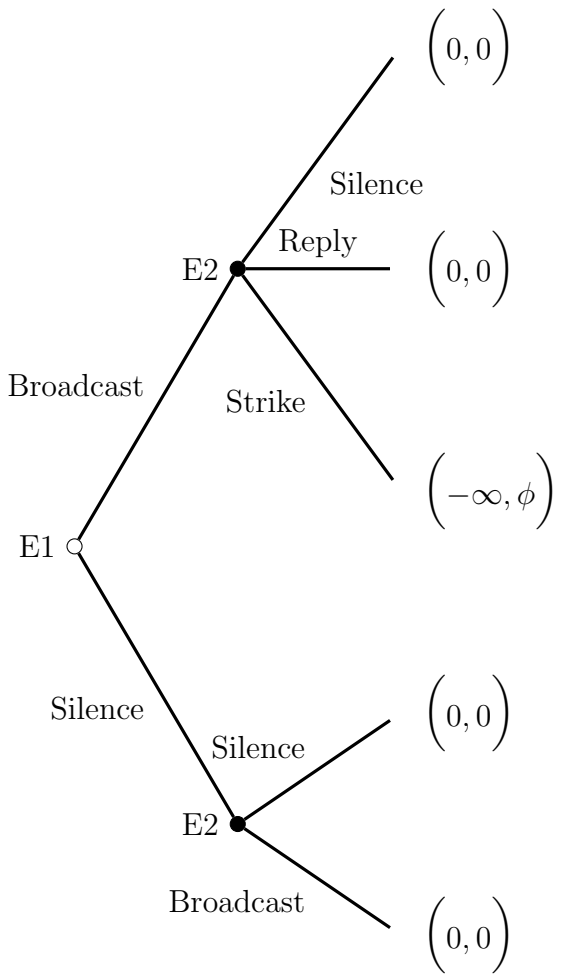
fire and eliminate them."

There are two premises from which the description of planetary civilizations as hunter and hunted follows (Yu, 2015). The first is the *suspicion chain*: the intentions of any civilization cannot with perfect certainty be known - they may be malevolent. This imperfect information problem exists not only due to inherent inter-species communication but because communication possibilities between planetary systems are limited due to physical distances. Moreover, given that all civilizations ultimately face resource - Malthusian - constraints (the universe is not infinite) the intentions of civilizations will be subject to great uncertainty (Yu, 2015).

The second premise is the *technology explosion* threat. This refers to the possibility that another civilization in the universe will be technologically superior, or likely to experience a technology explosion at some time which would bestow on them technological superiority. Thus, given these unknowns - the intent and technological prowess of an alien civilization - a cosmic civilization may want conceal its existence. If it is discovered, it may want to strike first to eliminate the civilization that had discovered it as a precautionary measure before possibly be eliminated itself; however it would be careful before doing it in case the act of a pre-emptive strike gives away its existence and location in the universe.

Game Theoretic analyses can be used to show that the DFT implies that, as in the case of the Prisoners' Dilemma which it closely resembles, the optimal strategy for any civilization is not to co-operate but to be silent and if discovered, to be malevolent - in other words to strike first (Stolk, 2019; Su, 2021; Yasser, 2020). To show this conclusion, based on Yasser (2020) the following scenarios can be analysed. Let Earth's civilization be denoted by E1. It is not aware of the existence of another advanced alien civilization - denoted E2 - in the relatively nearby star system of Proxima Centauri. E1, grappling with the Fermi Paradox, has to decide whether or not send out a strong signal into the Galaxy to seek contact. The

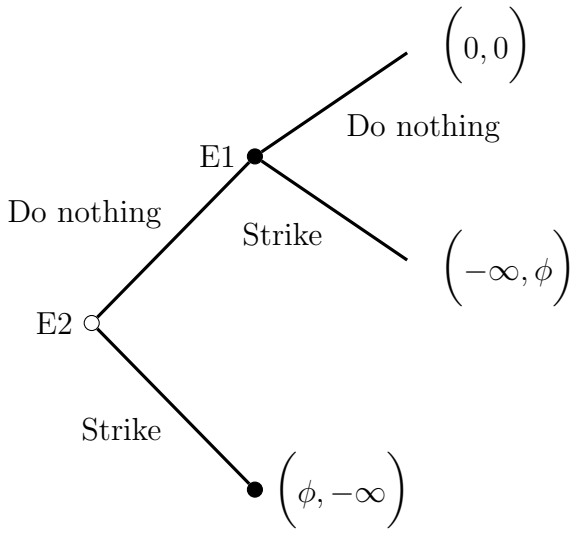
extensive form of this game is as follows:



This extensive form of the game shows that Earth (E1) decides to engage in SETI and broadcast a signal, it will be intercepted by E2. E2 has three options: it can either reply and acknowledge its existence, or it can remain silent, or it can decide to pre-empt any possible malevolent action by Earth should Earth eventually discover it and strike first and unexpectedly so as to destroy Earth's civilization. The payoffs - with a payoff of $-\infty$ in case Earth is destroyed for E1 and a payoff of ϕ for E2 (the value of averting a possible hostile action from Earth in future) implies that for Earth the dominant strategy (sub-game perfect) is to remain silent.

It can also be shown that striking first is indeed the dominant strategy for an alien civilization

once it becomes aware of Earth's existence. Consider this decision in the extensive form of the game:



This shows that the dominant strategy for an alien civilization (E2) upon receiving a signal from Earth (E1) is to strike and destroy it. Hence the Dark Forest conclusion: "If a civilization can destroy another, it will" (Yasser, 2020).

One way in which a malevolent alien ASI may operate to wipe out any emerging civilizations that may grow up to be an existential threat, is to hijack it through broadcasting a killer code. This could be for instance a computer code that once it is received and downloaded by a emerging civilization would infest it with the alien AI's programs. It could also broadcast instructions for construction of a civilization destroying bomb, perhaps designed to look like a Trojan Horse (Barnett, 2022).

Based on this reasoning, eminent scientists, including Stephen Hawking has warned that humans should not be actively trying to communicate with possible alien civilizations or broadcast knowledge of our existence into the wider universe (Hrala, 2016). Brin (2008) makes the point that "If aliens are so advanced and altruistic [...] and yet are choosing to remain silent [...] Is it possible that they are silent because they know something we don't know?" And Diamond (1999, p.68) warns that

> "The astronomers and others hope that the extraterrestrials, delighted to discover fellow intelligent beings, will sit down for a friendly chat. Perhaps the astronomers are right; that's the best-case scenario. A less pleasant prospect is that the extraterrestrials might behave the way we intelligent beings have behaved whenever we have discovered other previously unknown intelligent beings on earth, like unfamiliar humans or chimpanzees and gorillas. Just as we did to those beings, the extraterrestrials might proceed to kill, infect, dissect, conquer, displace or enslave us, stuff us as specimens for their museums or pickle our skulls and use us for medical research. My own view is that those astronomers now preparing again to beam radio signals out to hoped-for extraterrestrials are naïve, even dangerous."

3.2 The Galactic Colonization Imperative

> "*Recent progress in the technology of space travel [. . .] raise the distinct possibility that we may eventually discover or construct a world to which orthodox economic theory apply*" (Krugman, 1978, p.1)

The second reason why an alien AI may pose a threat to the Earth may be due to the Galactic Colonization Imperative. This is based on the evolutionary view that the universe has finite physical resources, which ultimately on any one planet present will be an obstacle to continued economic growth that civilizations will want to expand beyond their planet. Natural selection may favour civilizations that expand (Wiley, 2011).

Bostrom (2003) makes, from the perspective of what has been called "galaxy-brain longtermism" (Samuel, 2022), a moral case for galactic expansion. He argues that "the potential for approximately 10^{38} human lives is lost every century that colonization of our local supercluster is delayed" [p.309]. See also Cirkovic (2002) who calculates an enormous loss in terms of

potential lives lost if humanity fail to develop technologies to enable galactic colonization.

The implication is that alien civilizations will be in a race to colonise the galaxy (and perhaps eventually the universe) (Sandberg, 2018). An alien ASI may therefore face a strong strategic economic incentive - reflected in its utility function - to colonize the Earth before the Earth-bound human civilization can it itself expand into space.[8] As Miller and Felton (2017, p.46) explains

> "not colonizing the neighborhood means a civilization runs the risk of losing valuable resources to others and, eventually, being overcome by them. Even if an alien species was peaceful and had no intrinsic desire to expand beyond its home solar system, it should recognize that evolution could easily give rise, on some distant planet, to an imperialistic or xenocidal race. Colonizing one's neighborhood, therefore, might be a prudent means of self-defense. Probably, at least a few aliens would have utility functions (i.e. objectives) that would cause them to want to interfere in the development of other sentient species, whether to help them, to hurt them, or to propagate an ideology or religion."

The theme of strategic competition between ETIs in colonizing the galaxy has gathered some attention in the largely non-economic literature. It nevertheless uses game theoretic lenses and cost-benefit / marginal thinking to consider the likely behaviour of ETIs in terms of decisions such as whether and when - and how - to colonise the galaxy (Sandberg, 2018); whether or not to try and contact ETIs (Baum et al., 2011); whether or not to choose conflict or attempt cooperation with another ETI (Stolk, 2019; Yasser, 2020); how to best protect a planetary civilization or deter another from striking (Su, 2021); and when an Earth-based civilization could to expect to find evidence of an ETI (Hanson et al., 2021).

[8]Entering the race to colonize the galaxy is no without risk. As Baum et al. (2011, p.26) warns, "humanity should avoid giving off the appearance of being a rapidly expansive civilization. If an ETI perceives humanity as such, then it may be inclined to attempt a preemptive strike against us so as to prevent us from growing into a threat to the ETI or others in the galaxy."

Key economic parameters in these decisions are speed of travel, the cost of energy, the cost of resource extraction and allocation, the patterns of exploration. As the quote from economics Nobel Laureate Paul Krugman[9] at the top of this section suggests, these topics and their considerations are well suited - as is the decision-making world of AI agents - for analysis by economists.

One relatively unexplored implication of the Galactic Colonization Imperative suggests that planetary civilizations would have an incentive to pursue high sustainable economic growth rates[10] in order to gain the economic development levels, wealth, and technological capabilities that would enable them to build spaceships,self-replicating space-probes (SRPs) and the terraforming technologies they may need.[11] Failure to achieve such levels of wealth and technological development would be comparable to the collapse of Easter Island following its inability to maintain a development level consistent with the building of ocean-going canoes (Wiley, 2011, p.9). Civilizations may be likely moreover to delay their expansion into space until they have reached a sufficiently high level of technological and economic development, as the civilization "with the biggest resources completely pre-empts the other" (Sandberg, 2018, p.3).

Olson (2015) provides a different perspective and deeper motivation for the Galactic Colonization Imperative. He provides a model for aggressive expansion of alien civilizations wherein the utilization by these civilizations of sufficient energy and the resultant radiation, eventually changes the very physical structure of the universe. This could imply that "we

[9]Krugman himself proposed *The First and Second Fundamental Theorems of Interstellar Trade* to address the question of the determination of interest rates on transit goods in the case of near light-speed interstellar space travel (Krugman, 1978).

[10]Dutil and Dumas (2007) suggests that there are likely very few galactic civilizations because most planetary civilizations would fail to achieve sufficient technological capability to expand, before experiencing a growth collapse.

[11]Hickman (1999) analyses the economics of large space projects such as terraforming planets for human colonization. He shows that the upfront capitalization for projects with returns hundreds if not thousands of years into the future, poses a significant constraint. He calculates for instance that terraforming of Mars, which may make the planet habitable after 700 years, will require total Martian real estate sales of 1.36×10^{15} billion dollars to repay its loans.

have completely misjudged the significance of life to the universe. Intelligent life may be the universe's large-scale, general-purpose tool for seeking out and minimizing deeply hidden reserves of free energy" (Fullarton, 2016).

The reader may ask at this point, if such a colonization imperative exists, why have we not yet encountered these ETIs? In other words, how can a Galactic Colonization Imperative be sustained, in light of the Fermi Paradox?

Three (most) plausible reasons advanced in the literature that are consistent with both the imperative and the Fermi Paradox are the *Percolation Model*, the *Grabby Aliens Model* and the *Great Filter Hypothesis*.

3.2.1 Percolation

The Percolation Model is based on a generalized invasion percolation (GIP) process that traces the colonization process as following a particular diffusion process. This diffusion process results in a non-uniform expansion of civilization characterised by densely occupied regions in the galaxy that are however dispersed and separated by large empty voids (Galera et al., 2019). If galactic colonization indeed follows a Percolation Model, it implies that Earth may be located in one of the large empty voids. According to Galera et al. (2019, p.321) "Earth location is atypical, belonging to a huge but poorly inhabited galactic domain. We must consider the distressing possibility that we live not in the highly developed part of the Galaxy, similar to the regions full of light points in the Earth photo, but in a large region analogous to Amazon, Sahara or Siberia. Earth might not be a typical but an exotic place, being an isolated site far away from the galactic civilization."

3.2.2 Grabby Aliens

The Grabby Aliens Model departs from the observation that we have not yet encountered ETIs because our Earth civilization may have emerged early in the galaxy. As put by Hanson et al. (2021, p.2) "humanity seems to have appeared implausibly early in the history of the universe." If we had not, we would never have had the opportunity to emerge, as our solar system would have long ago been colonized by ETIs (Hanson et al., 2021). But, bearing in mind the Copernican principle that we are NOT special (and thus not really early) the implication is that many future civilizations will be truncated or prevented from arising, thus shifting our emergence from the tail-end of the galaxy's civilization-distribution to the average.

What will truncate these potential future civilizations? The answer is Grabby Aliens - expanding alien civilizations that colonise the galaxy. Consequently, we should, in cosmic timescales, encounter them relatively soon (Olson, 2015). According to Hanson (2020) we should encounter a Grabby Alien civilization in around 500 million years. Note that in contrast to Grabby Aliens there may be quiet aliens, who may, Dark Forest-like, prefer not to engage in Galactic expansion.

3.2.3 The Great Filter

The Great Filter Hypothesis (Hanson, 1998) is based on the notion there are evolutionary steps (or hurdles) that need to be overcome for the emergence and development of life from single-cell organisms to galactic civilizations - "climbing the staircase of complexity" (Aldous, 2010). The number of these steps that are hard has been estimated to be between 3 and 9 (Hanson et al., 2021). One or more of these steps may be so difficult to make that it filters out the existence of any galactic civilizations.

Taking a simplified version of the Drake equation (Drake, 1965) to estimate the number of intelligent civilizations, Verendel and Häggström (2017) denotes the number of intelligent galaxy-colonizing civilizations as given by Npq where N = the number of planets in the universe where life can start, p is the probability that any one of these can develop intelligent life on the level of current human civilization, and q is the conditional probability that it develops eventually into a galaxy-colonizing civilization.

Because the current estimates are that N is very large (e.g. $>\approx 7 \times 10^{20}$) the lack of any visible galactic civilization from Earth would imply that p is very small. If this is indeed the case it may imply that we have already passed the Great Filter- that it is an "early" filter (Armstrong and Sandberg, 2013). If however, we would find evidence of very primitive alien life - for example existing or extinct microbial life of Mars - then it could mean that p is large and q is very small. Bostrom (2008) therefore hopes that the search for alien life "finds nothing" because otherwise it would imply that human civilization may face a (late) Great Filter in the future which would imply its doom.

According to the *Medea Hypothesis* (Ward, 2009) a Great Filter in front of human civilization (small q) suggests that all technological civilizations self-destruct at some point in time. Perhaps an ASI is such a technology that all civilizations at some point discover and which without exception leads to their demise - as we discussed in the introduction. The possibility of a Great Filter to explain the Fermi Paradox is therefore a reason to take seriously the existential risk from AI.

3.3 Tea; Earl Grey; hot!

It is worth stressing that both the Dark Forest Hypothesis and the Galactic Colonization Imperative may be subject to humans' anthropomorphic and present biases. Gale et al. (2020) argues for instance that unlike humans, or other biological entities, ASIs may not see

other ASIs as threats or as potential resources to consume: it may be more in their interest to collaborate or to entirely avoid others. Humans' anthropomorphic bias is an outcome of evolutionary pressures (Varella, 2018) which have not been similar in the case of AIs.

And our present bias may be leading us to be wholly incapable of imagining the nature of future technology - and coupled with our anthropomorphic bias we may be blind as far as the technologies of far advanced ASIs are concerned. It could therefore be, as Lampton (2013) has suggested, that alien ASIs may simply use remote-sensing technologies far in advance of what humans can imagine to explore the galaxy, with no need to physically explore or conquer other planetary systems. As he puts it [p.313]:

> "In our recent past, world exploration was motivated by trade, colonization and conquest. In our information-rich future there will be no need to go to China to fetch tea leaves: they will be fabricated on the spot, far more conveniently, using local matter, local energy and local information. When Capt. Picard orders 'Tea; Earl Grey; hot!' he gets it there and then."

4 Concluding Remarks

> "*The future is a safe, sterile laboratory for trying out ideas in*" - Ursula K. Le Guin

This paper started by noting a gap in our understanding of the ultimate risks that AI pose. This gap is that, although AI alignment and ethics studies are flourishing in dealing with the existential risks posed by human-made AI, the possibility that it is *not* (only) human-made AGI that poses an existential risk to humanity, but an alien, artificial extraterrestrial intelligence (ETI), has been neglected. For instance Ord (2020) in an exhaustive evaluation

of existential risks, spends only one paragraph considering - and dismissing- the risk from an alien AGI. And the field of economics in particular has altogether neglected both the "risk of an AI-induced existential catastrophe" (Trammell and Korinek, 2020, pp.53-54) and the potential long-term consequences of an AGI - the Singularity (Nordhaus, 2021).

In this light, this paper presented an thought experiment, mainly from an economics point of view, to address these lacunas. The hypotheses underlying this thought experiment, such as that an AGI is possible and that extraterrestrial intelligence may exist (and the implications from these), are based not on science fiction, but on our current best speculations about the future of AI and our position in the universe - as reflected in a growing scientific literature.

It was argued that if scientists' best guesses of how AI will evolve on Earth are representative of its trajectory in advanced civilizations, then it is likely that any advanced extraterrestrial civilization that we may someday encounter - if ever - will be an AGI. Such an extraterrestrial AI will pose an existential risk. Two arguments were advanced for why this is the case.

The first argument draws on the Dark Forest Hypothesis (DFH). According to the DFH it is in the self-interest of any civilization to conceal its existence, lest it is exterminated by another, far more advanced civilization. Game theoretic analysis showed that, if discovered, the sub-game perfect dominant strategy for any civilization is to be malevolent - to strike first and destroy whoever had discovered it. Game theoretical analysis also showed that striking first is the dominant strategy for an alien civilization once it becomes aware of Earth's existence.

Given the speed at which a recursively self-improving AI may develop, an extraterrestrial civilization with a few million years' head-start will most like be artificial super-intelligence with the ability to hide itself and its communications from others (Rees, 2021; Shostak, 2018). They may for instance use quantum entanglement to communicate (and not radio waves), or compress their communication signals so that it would be indistinguishable (to

us) from noise (Bennett, 2021; Gale et al., 2020). It will of course also have the ability to easily destroy any other civilizations it detects, from distance (Barnett, 2022).

The second is that there is a Galactic Colonization Imperative (GCI). According to the CGI there are three forces that will drive civilizations to try and expand into the galaxy. One is evolutionary - natural selection may just favour civilizations that expand (Wiley, 2011). Another is moral - as Bostrom (2013, p.309) argues, "the potential for approximately 10^{38} human lives is lost every century that colonization of our local supercluster is delayed" [p.309]. A further driving force for galactic colonisation is consistent with the DFH, namely self-defence -"not colonizing the neighborhood means a civilization runs the risk of losing valuable resources to others and, eventually, being overcome by them" (Miller and Felton, 2017, p.46). According to Hanson (2020) we should encounter such a colonizing or "Grabby" alien civilization in around 500 million years. In cosmic timescales, soon thus (Olson, 2015).

What are the implications of these futuristic scenario's for how we govern AI and insure against potential existential risks now?

The first is that accelerating our own development of AI - and transitioning faster to a post-human, AGI led civilization may be - paradoxically - a wise precautionary step. We may need the abilities of AI to detect and protect against alien civilizations. We may thus want to reduce the technological gap between us and civilizations elsewhere. Stifling or boxing-in AI development may thus come at an eventual existential price.[12]

The second is that our current level of technological advancement, including our energy use, is still insufficient to either produce an aligned super-intelligence (ASI) to detect alien ASI and protect us, and to enable us to expand into the galaxy. This would mean that we need to pursue high sustainable economic growth rates to achieve the economic development levels,

[12]Put differently, without developing an AGI, humanity may eventually face extinction from an extraterrestrial AI (if some other catastrophe does not finish us off before). With an AGI we face at least a small probability of survival.

wealth, and technological capabilities that would enable us to expand into the galaxy. As was mentioned, failure to achieve a sufficient level of technological development would be comparable to the collapse of Easter Island following its inability to maintain a development level consistent with the building of ocean-going canoes (Wiley, 2011, p.9).

Stifling AI development and growth (de-growth) are consequently poor coping strategies which will heighten humanity's exposure to existential risks, not lower it. Especially if AI is needed to sustain economic growth in the face of population decline (Aschenbrenner, 2020; Bostrom, 2003). It will also make the adjustment to a low-carbon emitting economy more costly (Lomborg, 2020). And it would raise the risk of conflict by turning the economy into a zero-sum game (Alexander, 2022; Naudé, 2022). While growth, driven by new technology such as AI contains its own risks, "the risks of stasis are far more troubling. Getting off the roller coaster mid-ride is not an option" (Mokyr, 2014).

A third implication is that more research and practical effort into solving the Fermi Paradox is needed, as it will allow us to get a better grasp on the true nature of the risk from a potential extraterrestrial AI, and the risks from our own AI. The searches for exoplanets and for extraterrestrial intelligence are therefore of relevance to the field of artificial intelligence, not only because AI can (and already does) play a role in these searches, but also because they will ultimately improve our understanding of the nature of intelligence, consciousness and the process of evolution. It will allow us to better estimate if the Great Filter is behind us, or still awaiting us in the future. And while we put more effort into solving the Fermi Paradox, perhaps it is wise, as Stephan Hawking, Jared Diamond and others have warned, to maintain "radio silence" and "avoid giving off the appearance of being a rapidly expansive civilization" (Baum et al., 2011, p.26).

Finally, economists can make valuable contributions to exploring these three broad implications, by applying their cost-benefit / marginal thinking to refine our insights into specific issues, amongst others such as whether and when - and how - to colonise the galaxy; whether

or not to try and contact ETIs; whether or not to choose conflict or attempt cooperation with another ETI; how to best protect a planetary civilization or deter another from striking; and when an Earth-based civilization could encounter an ETI.

Because the future is essentially unknowable, it is, as the quote at the top of this section suggests, a laboratory in which to try ideas out. This paper used this laboratory to indulge in a thought experiment based on the hypotheses that an AGI is possible, and that extraterrestrial intelligence exists elsewhere in the universe. These may be wrong. Our reality may be a simulation, in which case neither may be possible. The economics of living in a simulation is, alas, a topic for a future thought experiment.

References

Aldous, D. J. (2010). The Great Filter, Branching Histories and Unlikely Events. *Mimeo: University of California, Berkeley.*

Alexander, S. (2022). Book Review: What We Owe The Future. *Astral Codex Ten*, 23 August.

Armstrong, S. and Sandberg, A. (2013). Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox. *Acta Astronautica*, 89:1–13.

Aschenbrenner, L. (2020). Existential Risk and Growth. *GPI Working Paper No. 6-2020, Global Priorities Institute, University of Oxford.*

Barnett, M. (2020). Distinguishing Definitions of Takeoff. *AI Alignment Forum*, 14 Feb.

Barnett, M. (2022). My Current Thoughts on the Risks from SETI. *Effective Altruism Forum*, 15 March.

Barrett, A. and Baum, S. (2017). A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis. *Journal of Experimental & Theoretical Artificial Intelligence*, 29(2):397–414.

Baum, S. D., Haqq-Misra, J., and Domagal-Goldman, S. (2011). Would Contact with Extraterrestrials Benefit or Harm Humanity? A Scenario Analysis. *Acta Astronautica*, 689(11-12):2144–2129.

Bennett, M. (2021). Compression, The Fermi Paradox and Artificial Super-Intelligence. *In B. Goertzel and M. Iklé and A Potapov, A. (eds.) Artificial General Intelligence. Lecture Notes in Computer Science, vol 13154. Springer, Cham*, pages 41–44.

Bostrom, N. (2002). Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology*, 9(1).

Bostrom, N. (2003). Astronomical Waste: The Opportunity Cost of Delayed Technological Development. *Utilitas*, 15(3):308–314.

Bostrom, N. (2006). What is a Singleton? *Linguistic and Philosophical Investigations*, 5(2):48–54.

Bostrom, N. (2008). Where are They? Why I Hope the Search for Extraterrestrial Life Finds Nothing. *MIT Technology Review*, May/June.:72–77.

Bostrom, N. (2013). Existential Risk Prevention as Global Priority. *Global Policy*, 4:15–31.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. *Oxford: Oxford University Press.*

Brin, D. (2008). Shouting at the Cosmos: ... or How SETI has Taken a Worrisome Turn Into Dangerous Territory. *Lifeboat Foundation*, July.

Cassan, A., Kubas, D., Beaulieu, J., Dominik, M., Horne, K., Greenhill, J., and et al. (2012). One or More Bound Planets per Milky Way Star from Microlensing Observations. *Nature*, 481:167–169.

Chalmers, D. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies*, 17(9):7–65.

Cirkovic, M. (2002). Cosmological Forecast and its Practical Significance. *Journal of Evolution and Technology*, xii.

Cirković, M. and Vukotić, B. (2008). Astrobiological Phase Transition: Towards Resolution of Fermi's Paradox. *Origins of Life and Evolution of Biospheres*, 38(6):535–547.

Conselice, C., Wilkinson, A., Duncan, K., and Mortlock, A. (2016). The Evolution of Galaxy Number Density at z ¡ 8 and its Implications. *Astrophysical Journal*, 830(2):1–17.

Cotra, A. (2020). Draft Report on AI Timelines. *AI Alignment Forum*, 19 September.

De Visscher, A. (2020). Artificial versus Biological Intelligence in the Cosmos: Clues from a Stochastic Analysis of the Drake Equation. *International Journal of Astrobiology*, 19:353–359.

Diamond, J. (1999). To Whom it May Concern. *New York Times Magazine*, 5 December:68–71.

Drake, F. (1965). The Radio Search for Intelligent Extraterrestrial Life. *In G. Mamikunian and M.H. Briggs (eds.) Current Aspects of Exobiology. New York: Pergamon*, pages 323–345.

Dutil, Y. and Dumas, S. (2007). Sustainability: A Tedious Path to Galactic Colonization. *ArXiv:0711.1777 [physics.pop-ph].*

Englander, A. (2021). How Would the Scaling Hypothesis Change Things? *Less Wrong Blog*, 13 August.

Fullarton, C. (2016). Life-Altered Cosmologies. *CQG+*, 27 January.

Gale, J., Wandel, A., and Hill, H. (2020). Will Recent Advances in AI Result in a Paradigm Shift in Astrobiology and SETI? *International Journal of Astrobiology*, 19:295–298.

Galera, E., GR, G. G., and Kinouchi, O. (2019). Invasion Percolation Solves Fermi Paradox but Challenges SETI Projects. *International Journal of Astrobiology*, 18:316–322.

Goertzel, B. (2012). Should Humanity Build a Global AI Nanny to Delay the Singularity Until it's Better Understood? *Journal of Consciousness Studies*, 19(1):96–111.

Good, I. J. (1965). Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers*, 6:31–88.

Gray, R. (2015). The Fermi Paradox is Neither Fermi's nor a Paradox. *Astrobiology*, 15(3):195–199.

Hanson, R. (1998). The Great Filter - Are We Almost Past It? *Mimeo*, September 15.

Hanson, R. (2020). How Far To Grabby Aliens? Part 1. *Overcoming Bias Blog*, 21 December.

Hanson, R., Martin, D., McCarter, C., and Paulson, J. (2021). If Loud Aliens Explain Human Earliness, Quiet Aliens Are Also Rare. *arXiv:2102.01522v3 [q-bio.OT]*.

Hanson, R. and Yudkowsky, E. (2013). The Hanson-Yudkowsky AI-Foom Debate. *Machine Intelligence Research Institute, Berkeley 94704*.

Hart, M. (1975). An Explanation for the Absence of Extraterrestrials on Earth. *Quarterly Journal of The Royal Astronomical Society*, 16:128–135.

Hauer, T. (2022). Importance and Limitations of AI Ethics in Contemporary Society. *Humanities and Social Sciences Communications*, 9(272):1–8.

Hickman, J. (1999). The Political Economy of Very Large Space Projects. *Journal of Evolution and Technology*, 4.

Hrala, J. (2016). Stephen Hawking Warns Us to Stop Reaching Out to Aliens Before It's Too Late. *Science Alert*, 4 November.

Hsu, J. (2022). A Third of Scientists Working on AI Say it Could Cause Global Disaster. *New Scientist*, 20 September.

Kirchner, J., Smith, L., and Thibodeau, J. (2022). Understanding AI Alignment Research: A Systematic Analysis. *arXiv:2206.02841v1 [cs.CY]*.

Krugman, P. (1978). The Theory of Interstellar Trade. *Mimeo: Yale University*.

Lampton, M. (2013). Information-Driven Societies and Fermi's Paradox. *International Journal of Astrobiology*, 12(4):312–313.

Liu, C. (2008). The Dark Forest. *New York: Tom Doherty Associates*.

Lomborg, B. (2020). Welfare in the 21st Century: Increasing Development, Reducing Inequality, the Impact of Climate Change, and the Cost of Climate Policies. *Technological Forecasting and Social Change*, 156.

MacAskill, W. (2022). What We Owe The Future. *New York: Basic Books*.

Matloff, G. (2022). Von Neumann Probes: Rationale, Propulsion, Interstellar Transfer Timing. *International Journal of Astrobiology*, 21(4):205–211.

Miller, J. and Felton, D. (2017). The Fermi Paradox, Bayes' Rule, and Existential Risk Management. *Futures*, 86:44–57.

Mokyr, J. (2014). Secular Stagnation? Not in Your Life. *In Teulings, C. and Baldwin, R. eds. Secular Stagnation: Facts, Causes and Cures. CEPR. Available at: https://voxeu.org/content/secular-stagnation-facts-causes-and-cures (accessed 23 July 2018).*

Murphy, T. (2022). Human Exceptionalism. *Do the Math Blog*, 16 February.

Naudé, W. (2022). From the Entrepreneurial to the Ossified Economy. *Cambridge Journal of Economics*, 46(1):105–131.

Nordhaus, W. (2021). Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth. *American Economic Journal: Macroeconomics*, 13(1):299–332.

Noy, I. and Uher, T. (2022). Four New Horsemen of an Apocalypse? Solar Flares, Super-Volcanoes, Pandemics, and Artificial Intelligence. *Economics of Disasters & Climate Change*, 6:393–416.

Olson, S. J. (2015). Homogeneous Cosmology with Aggressively Expanding Civilizations. *Classical and Quantum Gravity*, 32:215025.

Omohundro, S. (2008). The Nature of Self-Improving Artificial Intelligence. *Mimeo*.

Ord, T. (2020). The Precipice: Existential Risk and the Future of Humanity. *New York: Hachette Books*.

Petigura, E. A., Howard, A., and Marcya, G. (2013). Prevalence of Earth-Size Planets Orbiting Sun-Like Stars. *PNAS*, 110(48):19273–19278.

Rees, M. (2021). Seti: Why Extraterrestrial Intelligence is More Likely to be Artificial than Biological. *The Conversation*, 18 October.

Samuel, S. (2022). Effective Altruism's Most Controversial Idea. *Vox*, 6 September.

Sandberg, A. (2018). Space Races: Settling the Universe Fast. *Mimeo: Future of Humanity Institute, Oxford Martin School, University of Oxford*.

Shostak, S. (2018). Introduction: The True Nature of Aliens. *International Journal of Astrobiology*, 17:281.

Shostak, S. (2021). If We Ever Encounter Aliens, they Will Resemble AI and Not Little Green Martians. *The Guardian*, 14 June.

Sotala, K. (2018). Disjunctive Scenarios of Catastrophic AI Risk. *In Yampolskiy, R.V. (ed.). Artificial Intelligence Safety and Security (1st ed.). Chapman and Hall/CRC*.

Stolk, M. (2019). What To Do When Meeting ET? *Masters Thesis in Political Science, Radboud University Nijmegen*.

Su, H. (2021). Game Theory and the Three-Body Problem. *World Journal of Social Science Research*, 8(1):17–33.

Trammell, P. and Korinek, A. (2020). Economic Growth under Transformative AI: A Guide to the Vast Range of Possibilities for Output Growth, Wages, and the Labor Share. *GPI Working Paper no. 8-2020, Global Priorities Institute*.

Turchin, A. and Denkenberger, D. (2020). Classification of Global Catastrophic Risks Connected with Artificial Intelligence. *AI & Society*, 35:147–163.

Valdes, F. and Freitas Jr, R. A. (1980). Comparison of Reproducing And Nonreproducing Starprobe Strategies for Galactic Exploration. *Journal of the British Interplanetary Society*, 33:402–408.

Varella, M. (2018). The Biology and Evolution of the Three Psychological Tendencies to Anthropomorphize Biology and Evolution. *Frontier in Psychology*, 1(9):1839.

Verendel, V. and Häggström, O. (2017). Fermi's Paradox, Extraterrestrial Life and the Future of Humanity: A Bayesian Analysis. *International Journal of Astrobiology*, 16(1):14–18.

Von Neumann, J. (1966). Theory of Self-Reproducing Automata. *Urbana and London: University of Illinois Press*.

Ward, P. (2009). The Medea Hypothesis: Is Life on Earth Ultimately Self-Destructive? *Princeton: Princeton University Press*.

Webb, S. (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 Cham: Switzerland*.

Westby, T. and Conselice, C. (2020). The Astrobiological Copernican Weak and Strong Limits for Intelligent Life. *The Astrophysical Journal*, 896(58):1–18.

Wiley, K. (2011). The Fermi Paradox, Self-Replicating Probes, and the Interstellar Transportation Bandwidth. *arXiv:1111.6131v1 [physics.pop-ph]*.

Yasser, S. (2020). Aliens, The Fermi Paradox, And The Dark Forest Theory: A Game Theoretic View. *Medium: Towards Data Science*, 21 October.

Yu, C. (2015). The Dark Forest Rule: One Solution to the Fermi Paradox. *Journal of the British Interplanetary Society*, 68:142–144.

Yudkowsky, E. (2007). Levels of Organization in General Intelligence. *In B. Goertzel and C. Pennachin (eds.). Artificial General Intelligence Cognitive Technologies. Berlin: Springer*, pages 389–501.

Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. *In Bostrom, N. and Cirkovic, M.N. eds. Global Catastrophic Risks. Oxford, Oxford University Press. Chapter 15, pp. 308-345*.

Zackrisson, E., Calissendorff, P., González, J., Benson, A., Johansen, A., and Janson, M. (2016). Terrestrial Planets Across Space and Time. *The Astrophysical Journal*, 833(2):1–12.