Naudé, Wim

Working Paper

The Future Economics of Artificial Intelligence: Mythical Agents, a Singleton and the Dark Forest

IZA Discussion Papers, No. 15713

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Naudé, Wim (2022) : The Future Economics of Artificial Intelligence: Mythical Agents, a Singleton and the Dark Forest, IZA Discussion Papers, No. 15713, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/267450

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

DISCUSSION PAPER SERIES

IZA DP No. 15713

The Future Economics of Artificial Intelligence: Mythical Agents, a Singleton and the Dark Forest

Wim Naudé

NOVEMBER 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15713

The Future Economics of Artificial Intelligence: Mythical Agents, a Singleton and the Dark Forest

Wim Naudé
RWTH Aachen University and IZA

NOVEMBER 2022

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

The Future Economics of Artificial Intelligence: Mythical Agents, a Singleton and the Dark Forest

This paper contributes to the economics of AI by exploring three topics neglected by economists: (i) the notion of a Singularity (and Singleton), (ii) the existential risks that AI may pose to humanity, including that from an extraterrestrial AI in a Dark Forest universe; and (iii) the relevance of economics' Mythical Agent (homo economicus) for the design of value-aligned AI-systems. From the perspective of expected utility maximization, which both the fields of AI and economics share, these three topics are interrelated. By exploring these topics, several future avenues for economic research on AI becomes apparent, and areas where economic theory may benefit from a greater understanding of AI can be identified. Two further conclusions that emerge are first that a Singularity and existential risk from AI are still science fiction: which, however, should not preclude economics from bearing on the issues (it does not deter philosophers); and two, that economists should weigh in more on existential risk, and not leave this topic to lose credibility because of the Pascalian fanaticism of longtermism.

JEL Classification: O40, O33, D01, D64

Keywords: technology, artificial intelligence, economics, growth, existential risk, longtermism, Fermi Paradox, Grabby Aliens

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

E-mail: wnaude@ucc.ie

1 Introduction

While there has been progress in the field of Artificial Intelligence (AI) over the past two decades, the contribution of economists and economics to AI, so far, has been relatively modest. In a recent survey, Trammell and Korinek (2020) concluded that labour market impacts have so far been the primary concern of economists with AI. They further pointed out that two related topics have received "little direct attention" in economics - the possibility that AI can bring about a radical change in the growth mode or regime through a "Singularity" and the "risk of an AI-induced existential catastrophe" (Trammell and Korinek, 2020, pp.53-54). Economics Nobel Laureate William Nordhaus similarly lamented that "there is much about robots but remarkably little writing on Singularity in the modern macroeconomic literature" (Nordhaus, 2021, p.304). Furthermore, the related spectre of AI as existential risk has posed an AI value-alignment problem, which "economists have not paid much attention to" (Gans, 2018).

The purpose of this paper is to contribute to the economics of AI by exploring these three topics neglected by economists:[1] (i) the notion of a Singularity (and Singleton), (ii) the existential risks that AI may pose to humanity, including that from an extraterrestrial AI in a Dark Forest universe; and (iii) the relevance of economics' Mythical Agent (*homo economicus*) for the design of value-aligned AI-systems.

By exploring these topics, the costs and benefits of AI are de-hyped; several future avenues for economic research on AI becomes apparent; and areas where economic theory may benefit from a greater understanding of AI, are identified. These findings can be summarised before

[1]The purpose here is not to discuss the use of AI as a tool by economists to complement their traditional toolkit, which consists inter alia of econometrics, statistics, CGE models and Game Theory in economics. For a discussion and overview of the use of Machine Learning (ML) in economic data science and mechanism design, see for instance Athey and Imbens (2019), Dütting et al. (2019) and Zheng et al. (2020) and for the use of ML in complexity economics see Balland et al. (2022) and Brummitt et al. (2020). For a Scientometrics study of the extent to which economics papers have applied these tools, see Bickley et al. (2022). This paper is also not yet another on AI and the future of work.

proceeding.

Regarding the costs and benefits of AI, the Expected Utility Theory (EUT) perspective used in this paper suggests that as long as AI systems inhabit the small world of neoclassical Bayesian utility maximizing agents, they are restricted and pose little threat of misalignment or hard take-off that will result in a Singularity or existential risk. Narrow AI builds very smart, focused agents in very simple situations. This minimizes AI existential risk (although challenges for AI safety remain). This finding is consistent with the earlier conclusion of Naudé (2021) to expect *neither Utopian nor apocalyptic impacts from AI soon.*

Future avenues for research that are identified include further elaborations of economic growth models to explore the possibility of an AI-induced growth collapse, to explore the physical limits of growth, and to sharpen the tools to draw out the policy implications of facing fat-tailed catastrophic risks. Furthermore, economic perspectives may usefully be applied to the solutions and implications of the Fermi Paradox. These include bringing economic analyses to bear on potential far-future challenges such as decisions such as whether and when - and how - to colonise the galaxy; whether or not to try and contact Extraterrestrial Intelligences (ETIs); whether or not to choose conflict or attempt cooperation with another ETIs; how to best protect a planetary civilization or deter another from striking; and when an Earth-based civilization could expect to find evidence of an ETI.

This paper also finds that economic theory may benefit from the field of AI as far as procedural utility is concerned, to overcome computational difficulties in understanding reasoning under bounded rationality, and help economists to model human behaviour better in disequilibrium situations.

Finally, two further conclusions that emerge from this paper are first that a Singularity and existential risk from AI are still science fiction: which, however, should not preclude economics from bearing on the issues - it certainly does not deter philosophers; and two,

that economists should contribute more to existential risk studies, and not leave this topic to lose credibility because of the Pascalian fanaticism of the recent fad of Longtermism.

The paper is structured as follows. In section 2 the foundation of both modern data-based AI and neoclassical economics, which is Expected Utility Theory (EUT), is outlined. It is shown that modern data-based AI has at its very basis the Mythical Agent of neoclassical economics - the *homo economicus*. The challenges of endowing AI agents with utility functions (goals and sub-goals) are discussed. These challenges complicate the design of AI agents if we envisage them becoming smarter to the degree that their values and actions would not be aligned with those of humans. However, as section 2.4 conclude, by imposing the EUT framework on AI-agents, they are essentially constrained to narrow, probabilistic domains - or "small worlds" in the sense of Savage (1954) - which may suggest that there is no AI alignment problem as such - although there are AI safety challenges.

Section 3 provides an economics perspective on the Singularity, a highly speculative concept in the field of AI which refers to the possibility that at some future time continued improvements in AI (the scaling up of Deep Learning) will lead to an intelligence explosion - and the emergence of a human-level and even superintelligent AI. The literature refers to these possible post-Singularity AIs as Artificial General Intelligence (AGI) and Artificial Super-intelligence (ASI).[2] Note that far from all scientists in the field are convinced about the possibility of an AGI or ASI or that such will imply existential risks or super-abundant growth. Critical voices include among others Noam Chomsky (see Katz (2012)), David Deutsch (see Deutsch (2011)), Judea Pearl (see Pearl and Mackenzie (2018)) and Steven Pinker (see Pinker and Aaronson (2022)). Russell (2019, p.8, p.13) has even stated that "the standard model of AI is a dead end."

Nevertheless, and perhaps succumbing to ***Pascal's Mugging*** - see Box 1 -, this paper

[2] A super-intelligence is "any intellect that vastly outperforms the best human brains in practically every field, including scientific creativity, general wisdom, and social skills" (Bostrom, 1998, p.1).

will proceed on the basis that a Singularity is at least hypothetically possible and therefore worthy of an investment in time and thought. The paper takes inspiration from the *The First and Second Fundamental Theorems of Interstellar Trade*, proposed by Krugman (1978).

The possibility of AGI/ASI (the Singularity) could lead to exponential economic growth - a growth explosion - and enable humanity to eradicate poverty and disease, and deal with other wicked problems;[3] however it could also present an existential risk - it could lead to a Singleton, a single ASI dominating the planet - that either intentionally, or accidentally, destroy civilization.

Box 1: Pascal's Mugging.

Expected utility maximization is a central concept in this paper. Pascal's Mugging, a variation on Pascal's Wager, is a thought experiment that shows that if expected utility maximization is not properly applied[a] that it could lead to erroneous decisions. In the thought experiment, a person- say Pascal - is accosted at night in a dark alley by a mugger, who - alas - has forgotten to bring along a weapon. The mugger then makes the person an offer: if they hand over their wallet, the mugger will return in a few days with a million $. If the person had $10 in their wallet, and if the probability of the mugger actually keeping their word is 0.0001% then the expected value of the chance is $100, which is 10 times the amount in their wallet. Handing over one's wallet would seem to be the rational decision. According to Yudkowsky (2007b) "Pascal's Mugger is just a philosopher out for a fast buck." Pascal's Mugger will appear a few times in this paper.

[a]Proper application of expected utility maximization requires at least the use of bounded utility functions or discarding future events with very small probabilities (Monton, 2019).

Section 4 tries to answer the question, what could be the economic and societal consequences of the future AI development trajectory as presented in section 3, particularly given the

[3]In the words of Domingos (2015, p.289) AI could result in the next millennium to "be the most amazing in the life of planet Earth." Russell (2019, pp.98-99) estimates that if AI can raise GDP per capita for the world's population to the current 88th percentile of that of the USA, this would add US$674 trillion per year to the world economy.

difficulties in containing and aligning AI? It explores one the one hand (section 4.1), whether economic growth will collapse, explode, or grow exponentially, and whether there will be phase or mode transitions in the nature of economic growth. On the other hand, it explores the existential risks that may follow from a Singularity (section 4.2).

Section 5 follows the possible implications of a Singularity to the far future - possibly > 500 million years hence, to consider the economics of a Galactic AGI, and discuss how decision theory has been used to consider galactic colonization and the optimal strategy to deal with Extraterrestrial Intelligences (ETIs) - which are most probably going to be alien AGI/ASIs. It suggests that, if the answer to Fermi's question[4] "where is everybody?" is that the universe is a Dark Forest, then a further existential threat from AI will be from an extraterrestrial AI. A warning though : this section may be "at best distracting and at worst irresponsible" (Floridi, 2022, p.1).

Section 6 puts in perspective the paper's explorations of the three knowledge gaps it tried to address, identifying several implications for the future of the economics of AI. Section 7 concludes.

2 Mythical Agents

> "*AI researchers aim to construct a synthetic homo economicus, the mythical perfectly rational agent of neoclassical economics*" (Parkes and Wellman, 2015, p. 267)

In this section the foundation of both modern data-based AI and neoclassical economics in rational choice theory (Expected Utility Theory - EUT) is outlined.[5] It is shown that

[4]In 1950 during lunch, the physicist Enrico Fermi asked "*where is everybody*?" with reference to the lack of evidence for extraterrestrial life in a universe roughly 13.8 billion years old - old enough for civilizations to have developed and colonised the galaxy (Gray, 2015).

[5]It should be pointed out at the outset that rational choice theory is not a single unified theory but

modern data-based AI is founded on the mythical agent of neoclassical economics - the *homo economicus*.[6]

Before proceeding, it is necessary to motivate why this section is indispensable for framing the discussions that will follow in sections 3 and 4 of this paper, which deal respectively with the rise of the Singleton following a Singularity and the development of a Galactic AI within a Dark Forest vision of the universe. It is necessary for three reasons.

First, the central research agenda of AI scientists is to create intelligent, autonomous agents that can make rational decisions - who, after all, would want to design an agent that makes irrational decisions? This has confronted them with two questions (Oesterheld, 2021, p.2): "1. What decision theory do we want an AI to follow? 2. How could we implement such a decision theory in an AI?"

Second, starting with EUT goes right to the core of one of the challenges that AI researchers face - and which is partially due to the fact that economics has tended to assume away the difficult parts of decision-making, preferring to focus on substantive rationality over procedural rationality. This leaves AI scientists with several challenges, such as how to set the appropriate goal (utility function) for AI agents; how to deal with the implications that these agents will follow sub-goals (instrumental goals); whether AI agents should be allowed or would be able to change their own goals (utility functions); how to avoid corruption of utility functions, including through problems such as wireheading; and how to create artificial moral agents (AMAs) who can base their utility maximization on values, and how to model decision-making in multi-agent settings and human-agent collectives (HACs).

Thirdly, the challenges just mentioned give rise to an AI Alignment Problem: how to ensure that an AGS/ASI will not in future bring any harm to humanity? How can a "Friendly AI" be designed, given the indeterminate state of rational decision theory? As long as the AI

consists of variants. For a review of these, see Herfeld (2020)

[6]For a discussion of the concept and origin of *homo economicus*, the economic human, see Persky (1995).

alignment Problem is not solved (and to date it has not) the possibility exists that in future AGI/ASI may not help humans create a super-fast growing, abundant economy (or even a steady-state economy if that is what human society desires) but rather follow its own goals based on different values from humans. And this may pose an existential risk.

Thus, understanding how AI scientists are attempting to operationalize the mythical neo-classical *homo economicus* is central in understanding the eventual economic implications of the Singularity and the existential risks posed by an eventual AGI. It helps identify where economists can contribute to solving the AI Alignment problem - a problem to which they have so far contributed little (Gans, 2018).

2.1 Expected Utility Theory

> "*All tasks that require intelligence to be solved can naturally be formulated as a maximization of some expected utility in the framework of agents.*" - Hutter (2007, p.33)

2.1.1 What decision theory do we want an AI to follow?

If the field of economics is concerned with constrained optimization under uncertainty, for instance, in modelling individuals as goal-oriented, rational agents acting to maximize their subjective utility subject to resource constraints, then it has much in common with the field of artificial intelligence (AI). This is because the field of AI is "the study of agents that receive percepts from the environment and perform actions. Each such agent implements a function that maps percept sequences to actions [...]" (Russell and Norvig, 2021, p.vii).

The functions that map percept sequences (perceptions) to actions should help agents to select actions to achieve their goals (Parkes and Wellman, 2015). In the case of genes,

for instance, the goals are survival and gene transmission (Kamatani, 2021). The human phenotype, including its brain, is the expression or action of its genes, which aims at survival and transmission (reproduction) (Williams, 1966; Dawkins, 1976). In the terminology of AI, survival and reproduction are the supergoals of genes, and the human brain is a subgoal (or instrumental subgoal) (Yudkowsky, 2001). In section 2.4 we will return to the topic of subgoals/instrumental goals.

To answer the question "What decision theory do we want an AI to follow?" so as to use the best functions, AI scientists have been relying on Expected Utility Theory (EUT) and qualitative variants thereof (Dastani et al., 2005; Gonzales and Perny, 2020; Russell, 2019). EUT is the foundation of neoclassical economics. It is a formal model of rational decision-making set out by Von Neumann and Morgenstern (1944) (vNM) and generalized by Savage (1954).

The foundations of the approach go back however to Bernoulli (1738) and his solution to the St. Petersburg Paradox (List and Haigh, 2005). The St.Petersburg Paradox arises in gamble where a fair coin is tossed n-times, until it lands on heads, with the gambler then receiving a prize of $\$2^n$. The paradox is that even though the expected value of the gamble being $\sum_{n=1}^{\infty}(\frac{1}{2})^n \times 2^n = \infty$, no one would pay very much to take this gamble. Bernoulli (1738) solved this by showing that what is important is not to maximize expected *value*, but expected **utility**. Utility should also be maximized after ignoring outcomes with very small probabilities[7] (Monton, 2019).

In economics, and based on Von Neumann and Morgenstern (1944) EUT justifies the specification of an *Utility Function* which allows an agent to compare different outcomes from actions - the utility function reflects the agent's preferences.[8] Consequently, the agent will

[7]The problem of Pascal's Mugger (see Box 1) can now be seen to be due to the maximization of expected value, and not utility, and the consideration of outcomes with minuscule probabilities.

[8]For a survey of how preferences are incorporated in the utility function of AI agents, see Pigozzi et al. (2016).

choose actions that maximises the (expected) value of the utility function. Note that in this approach, agents maximise *expected* utility because each possible future outcome is subject to probability - an outcome is a lottery. Actions and their outcomes can thus be compared to playing a lottery.

A lottery can be denoted by $L = [p_1(C_1), p_2(C_2), p_3(C_3), ..., p_k(C_k)]$ where the $C_k's$ are the outcomes and $\sum_{i=1}^{k} p_i = 1$ the probabilities of each outcome. The expected value (E) of this lottery is $E(L) = \sum_{i=1}^{k} p_i C_i$, the mathematical average. If the set of lotteries available to an agent i is Λ then the agent's utility function[9] U_i represents the preferences of the agent over various lotteries, with $U_i(L_1) \geq U_i(L_2) \iff L_1 \succsim L_2, \forall L_1, L_2 \in \Lambda$ (Maschler et al., 2013). Thus, if lottery L_1 is preferred to lottery L_2 the utility from L_1 will be greater or equal than the utility from L_2.

In following EUT to make decisions, an agent will do best to choose the lotteries L_i^* (or consumption bundles, as in household economics) to maximise expected utility, $\mathbb{E}[U(L_i)]$. This decision can be written as

$$L_i^* = \arg\max_{L} \mathbb{E}[U(L_i)] \tag{1}$$

This choice of L_i^* can be found by solving for $\dfrac{\partial \mathbb{E}[U(L_i)]}{\partial L_i} = 0$

In vNM, human agents will maximize expected utility, choosing the lotteries from Λ that will generate the most utility. This will, say in the example of a consumer aiming to maximize utility from buying various bundles of goods, lead them to goal-directed decisions and pursuit of instrumental subgoals, such as acquiring money or income. Just like not all lotteries can be played, not all bundles of goods can be afforded. Von Neumann and Morgenstern (1944) proved that if individual agents' preferences[10] are characterized by completeness, transitivity,

[9] A linear utility function is a von Neumann-Morgenstern utility function and implies a risk-neutrality. If u_i is concave then the agent is risk averse and if U_i convex, risk-seeking (Maschler et al., 2013).

[10] In economics we do not require direct knowledge of an agent's preferences - it can be inferred from their

and continuity, then they will behave as if they were maximizing expected utility (Moscati, 2016).

In the decision calculus, so far, the outcomes of the decision on L_i accrues only to the agent making the choice: it is implicitly assumed that there are no externalities. In reality, however, and in the case of concerns about AI as we will elaborate in section 2.3, the unintended consequences of agents' decisions need to be considered. This is a formidable problem. Consider for instance that, if we denote external costs/benefits of a decision or action on L_i by $C(L_i)$ then the decision in (1) can be re-written as

$$L_i^* = \arg\max_L \mathbb{E}[U(L_i) - C(L_i)] \tag{2}$$

Gauchon and Barigou (2021) discusses the general complexity of this problem, noting that finding the optimum requires various assumptions and somewhat tenuous interpretations of the terms in the first-order conditions.

2.1.2 Examples in the Field of AI

Equivalents to utility functions[11] used in AI systems include value functions, objective functions, loss functions, reward functions (especially in Reinforcement Learning), and preference orderings (Eckersley, 2019). The concepts of utility function and goal are often used interchangeably in the AI literature (Dennis, 2020). Where loss functions (gradients) are used, which is the case when some objective function is minimized (for e.g. minimizing the error of wrongly predicting what is in an image) the sign on the above utility function would be negative.

choices - their revealed preferences, a notion introduced by Ramsey (1931) and Samuelson (1947).

[11] For an extensive overview of the mathematics used in Machine Learning (ML) see Gallier and Quaintance (2022).

An example from ML is the ubiquitous use of artificial neural networks (ANN), such as the Multilayer Perceptron (MLP)[12] to perform classifications (say classifying images or text).

Formally, an ANN aims, given a (data) set of N samples $D = \{[x_1, x_2, ...x_n], [y_1, y_2, ...y_n]\}$ to find the best approximation for the function describing $f(x_i) = y_i$ which maps the inputs (x) to the outputs (y), where the outputs would be the classification (Lichtner-Bajjaoui, 2020). The objective function is to minimize the expected value of incorrect classifications, which is equivalent to maximizing the utility or goal of the ML (typically back-propagation[13]) algorithm (García-García et al., 2022). In the case of the MLP the probabilities attached to each element x to be classified belonging to a class k is a vector $P(y_k|x)$ that can be written as

$$P(y_k|x) = s'\Big\{\sum_{j=1}^{n_2} \omega_{jk} \times s\Big\{\sum_{i=1}^{n_1} \omega_{ij} \times x_i + b_{0j}\Big\} + b_{0k}\Big\} \tag{3}$$

Where s and s' are respectively known as the activation functions of the hidden and output layers of the neural network, the n the number of neurons in each of these layers, the ω's weights on the connections between the layers and neurons, and b_0's threshold values (activation functions). The back-propagation algorithm will adjust the weights and threshold values to minimize the loss function in classification (and maximize the probability that a classification is accurate). For a more detailed discussion and examples, see García-García et al. (2022).

Deep Learning using ANNs to perform classifications has found wide use in Recommender Systems, such as is used to suggest what movies or songs subscribers on Netflix or Spotify may want to watch, or what consumers on Amazon or other platforms may want to

[12] A MLP is a supervised learning algorithm consisting of various layers of perceptrons, where a perceptron is a program that tries to mimic a biological neuron to perform binary classifications. See Rosenblatt (1958) and Schmidhuber (2015).

[13] See Rumelhart et al. (1986) who introduced back-propagation as a supervised learning technique.

buy (Ricci et al., 2022). Jenkins et al. (2021) have shown how these systems can be made more accurate by being explicitly based on micro-economics based utility functions, which they term *Neural Utility Functions*. They pointed out that Recommender Systems, which minimize an objective function of choice prediction, do not use quasi-concave utility functions, as economics typically do. Accordingly, they cannot evaluate trade-offs in decisions, such as taking into account that choices are affected by whether there are complements or substitutes. They also show that if they augment DL models with quasi-concave Neural Utility Functions,[14] that the choice-prediction loss is smaller - using utility functions with a foundation in economic theory improves AI (Jenkins et al., 2021).

A final example of utility functions in the field of AI is that deep convolutional neural networks (CNNs), a class of ANNs used for image classification, can be interpreted as utility maximizers subject to costly learning , i.e. they face informational costs (this is elaborated below in section 2.2 under bounded rationality) (Pattanayak and Krishnamurthy, 2021). Applying this idea to deep CNN, Pattanayak and Krishnamurthy (2021) found that they could predict the image-classification performance of 200 deep CNNs with an accuracy $>$ 94%, removing the need to re-train the models.

2.1.3 Sequential Decision-Making

Many decisions replying on EUT are not one-off decisions but sequential. This is particularly, but not exclusively so in multi-agent settings (see section 2.3.4) - where both economics and AI science rely on Game Theory. Gonzales and Perny (2020) discusses the use of graphical models such as Decision Trees to analyse such sequential decision-making. In complex sequential decision-making under uncertainty, economists use stochastic dynamic programming (Bellman, 1957a,b) and its Markov Decision Process (MDP) model (Howard, 1960). A typical example is of inventory management (Ahiska et al., 2013).

[14] They use CES and Cobb-Douglas utility function specifications.

Effective sequential decision-making by AI agents is vital in virtually all AI applications - from playing games like Chess and Go, to autonomous robots and vehicles, health planning and chatbots. In all of these, the sequence in which decisions are made are important for the overall maximization of the utility function. Reinforcement Learning (RL) is the branch of AI that focuses mainly on sequential decision-making. In most RL[15] an AI agent learns about the underlying MDP "through execution and simulation, continuously using feedback from the past decisions to learn the underlying model and reinforce good strategies" (Agrawal, 2019, p.2). To speed up learning, recourse may be taken, given the nature of the goal, to Supervised Learning or Imitation Learning[16] (Hutter, 2000; Ding et al., 2019). For detailed discussions of RL the reader is referred to Arulkumaran et al. (2017) and Sutton and Barto (1998). Charpentier et al. (2021) describes how RL is used in development of autonomous vehicles. Salimans et al. (2017) refers to the success of RL for developing AI systems that can excel in games such as Atari and Go.

2.1.4 Challenges to EUT

EUT is subject to at least two challenges. One is that experiments have found that humans may violate the EUT under certain conditions, thus apparently not acting rationally (List and Haigh, 2005); a second relates to evaluating possible future outcomes where there is no objective probability distribution (LeRoy and Singell, 1987) - also known in economics as Knightian uncertainty after Knight (1933).

Regarding the first challenge, an example is the Ellsberg Paradox or Ellsberg's Urn - see Ellsberg (1961). According to Binmore (2017) a way around the Ellsberg Paradox, which reflects humans' ambiguity aversion, is to screen agents beforehand using another rationality

[15]Exceptions are so-called black-box optimization or direct policy search, which includes a class of optimization algorithms known as Evolution Strategies (ES) (Salimans et al., 2017) and the AIXI model of universal AI of Hutter (2000, 2007) which dispenses with the Markov assumption that the the future only depends on the present.

[16]Particularly useful in robotics (Ding et al., 2019).

criterion.[17] In the case of using EUT to model the decision-making of AI-agents, such screening is implicitly done - by the selection of AI agents who are not human, to begin with. Thus AI agents more fully inhabit the world of neoclassical economics meaning that economic theory can usefully be applied to AI (Caplin et al., 2022). Other approaches that have been tried in AI to avoid the Ellsberg Paradox is to model AI agents' behaviour indeed closely on that of humans, and to do so by relying on models from behavioural economics (see below in section 2.2) - see for instance Tamura (2009).

Regarding the second challenge, to deal with uncertainty, both economics (Harsanyi, 1978) and AI (Pearl, 1985) revert to Bayes' Theorem (Harré, 2021). Von Neumann-Morgenstern's EUT (Von Neumann and Morgenstern, 1944) is based on objective probabilities. (Savage, 1954) generalised this to subjective probabilities. Here,"Bayesian" agents form subjective probabilities based on their priors (beliefs). As new information comes to light, they update their subjective probabilities - and accordingly modify their actions (Savage, 1954; Harsanyi, 1978). *How* their priors are established is a question of some contention and highly relevant to the agenda of AI scientists (Binmore, 2008, 2017).

2.2 Bounded Rationality

Bayesian expected utility maximizers - subjective utility maximizers - are the Mythical Agents of both neoclassical economics and the field of AI. The problem is that rational decision-makers often make poor - less than optimal- choices (Binmore, 2017). This is because, unlike in the theoretical idealized world of neoclassical economics, in reality agents - both human and AI - face informational and computation limits. As Simon (1955, p.114) put it, it may be more useful to replace the mythical agent of neoclassical economics with an " of

[17]Several generalizations of EUT have been proposed to deal with this and other shortcomings of EUT to be a good descriptive model of human decision-making. A discussion of these falls outside the scope of this paper. The reader is referred to Gonzales and Perny (2020), who discusses amongst others Rank Dependent Utility (RDU) and decision models outside the probabilistic framework; and to Schoemaker (1982) who discusses nine variants of EUT - from expected monetary value to Prospect Theory.

limited knowledge and ability" which does not have the "global" rationality of the mythical agent. In the case of humans, mistakes in decision-making are often "predictably irrational" (Ariely, 2009) - which has been ascribed to cognitive biases (Kahneman et al., 2021). Computational limitations and cognitive biases have been used by behavioral economists to argue that Human Sapiens differ from *Homo Economicus* (Thaler, 2000). Intelligence agents' rationality is thus *bounded*.

Bounded rationality is not only applicable to humans, but also to AI agents - even though they may have vastly better computational abilities (Dennis, 2020). As Wagner (2020, p.114) point out "whilst the new species of 'machina economicus' [...] behaves more economic than man, it too is faced with bounded rationality. Algorithms work with finite computational resources which in practice means that they cannot achieve Turing completeness and are limited to linear bounded automation." The reference here to *Turing Completeness* is to the theoretical possibility that AI can be globally rational as the mythical agent of neoclassical economics (Lee, 2019). A *Universal Turing Machine* (UTM) is a "computing machine", proposed by Turing (1936) that can "be used to compute any computable sequence" (Turing, 1936, p.241). It is Turing Complete. However, it is subject to the *Halting Problem*, which is how to determine if and when the UTM will find a solution (Lee, 2019). Turing (1936) proved that there is no general algorithm for solving this problem in all cases.

Finite computation resources, as described in the previous paragraph, implies that there are information costs involved in making a decision as described in equation (1). These information costs can be further specified to come from the updating of an agent's Bayesian priors. If, in the example of equation (1) in section 2.1 the agents' probability distribution over the choice of L_i is $p(L)$ then computational resources (costs) will be expended to change from a prior probabilistic strategy $p_0(L)$ to a posterior probabilistic strategy $p(L)$ in the process of decision-making (Leibfried and Braun, 2016). This informational cost is known as the Kullback-Leibler divergence (D_{KL}) and can be specified as $D_{KL} = (p(L)||p_0(L)) \leq B$

where $B \geq 0$ is the upper bound of available computational resources (Leibfried and Braun, 2016). With these informational costs, the expected utility maximization problem in (1) can be modified to

$$L_i^* = (1-\beta) \arg\max_L \mathbb{E}[U(L_i)] - \beta D_{KL}(p(L)||p_0(L)) \tag{4}$$

Where $\beta \in (0,1)$ is the trade-off between expected utility and informational cost (Leibfried and Braun, 2016). It is also consistent with Sims (2003)'s "rationally inattentive utility maximization" where paying more attention to making a decision implies high attention costs. The point is that learning is costly, it bounds rationality, and needs to be taken into account in models of bounded rational decision-making (Lipnowski and Doron, 2022; Pattanayak and Krishnamurthy, 2021).

The difference though between AI bounded rationality and human bounded rationality is that while human agents are subject to both computational limits and cognitive biases, AI agents will face fewer computational limits and can unlearn cognitive biases. Learning may be less costly. AI agents can be programmed with error correction mechanisms and these will inevitably drive them to Homo Economicus, or more appropriately as Parkes and Wellman (2015) have suggested, *Machina Economicus*.

As Omohundro (2008b) explains, the nature of AI systems such as Deep Learning[18] and (Deep) Reinforcement Learning[19] as "learning" agents means that they are self-improving systems. They will thus learn where they have been making sub-optimal decisions or have been deviating from their goals, and correct for it in a way "*discover and eliminate their own irrationalities in ways that humans cannot*" (Omohundro, 2008b, p.4).

[18]The dominant approach to unpack functions that maps precepts from the environment into actions is Deep Learning (LeCun et al., 2015; Sarker, 2021).

[19]For more comprehensive explanations of Reinforcement Learning (RL) see Arulkumaran et al. (2017) and Sutton and Barto (1998). Charpentier et al. (2021) describes how RL is used in development of autonomous vehicles. Salimans et al. (2017) refers to the success of RL in games such as Atari and Go.

Thus, the state of the art in the fields of economics and AI is the modelling of intelligent agents that rely on Bayesian probability theory to inform beliefs (priors), and utility theory to inform their preferences. With beliefs and wants determined, and with limits on their computational abilities and resources, by aiming to maximize expected utility, intelligent agents will act in a boundedly rational manner under uncertainty (Benya, 2012; Riedel, 2021). Intelligent agents with beliefs and preferences but who fail to attempt to maximize expected utility may be vulnerable and subject to exploitation (Shah, 2019a). Eventually evolutionary pressures will lead to the disappearance of such agents from a population.

This bounded mythical agent of rational decision theory has much to commend it, as its voluminous application in the economics literature, and its dominance in explaining decision-making, attest to (Binmore, 2008; Dixon, 2001; Moscati, 2016). It has even be found to be applicable to decision-making in other primates, not only humans (Pastor-Berniera et al., 2017). Indeed, it is due to its strengths that it has come to underpin the development of AI.

But, there are also kinks in the armour, which, in light of the continued advancement in AI, poses a number of challenges for AI scientists - and has implications for the development of economic theory. So far, economists have contributed little to addressing these - it remains an area for future research for the economics of artificial intelligence.

In sub-section 2.3 that follows, some of the main challenges for operationalizing the Mythical Agents of economics are discussed, under the heading of the AI alignment problem.

2.3 The AI Alignment Problem

Although AI agents are also rationally bounded, they have fewer cognitive biases than humans, and can unlearn these. Learning and eliminating its biases imply that AI may become recursively self-improving. In such a situation, where AI learns, self-correct and recursively

self-improve, one would need to avoid the eventually that a smart AI emerges and pursue goals (utility) that conflict with human interest (Bostrom, 2014), or even if these do not conflict with human interest, nevertheless can have unintended negative consequences. These may follow because its utility function may not capture all the considerations relevant in a situation - "humans care about many features of the environment that are difficult to capture in any simple utility function" (Taylor, 2016, p.125).

The challenge is, as formulated by Omohundro (2008b, p.36), that AI systems following EUT

> "will maintain utility functions which encode their preferences about the world. In the process of acting on those preferences, they will be subject to drives towards efficiency, self-preservation, acquisition, and creativity. Unbridled, these drives lead to both desirable and undesirable behaviors. By carefully choosing the utility functions of the first self-improving systems, we have the opportunity to guide the entire future development."

It is therefore essential, as Riedel (2021) and others have argued, to understand the utility functions (goals) of AI agents - not only for participating with and competing against AI agents, but eventually for constraining and aligning AI's goals (Bostrom, 2014). Constraining and aligning AI systems' goals or utility functions is a topic that has generated a large and growing literature under the headings of AI alignment and AI ethics (Hauer, 2022), which ultimately aims to ensure that AI "benefits humans" and humans do not lose control over AI (Kirchner et al., 2022, p.1). Note that this is a very human-centric agenda based on a view of human exceptionalism (Murphy, 2022a).

Discussing all the risks in AI goal design falls outside the scope of the present paper - a recent washing list of such AI goal design risks identified 26 different risks (Kokotajlo and Dai, 2019). Further reviews on aligning AI are contained in Christian (2020), Everitt et al. (2018), Hubinger (2020) and Kirchner et al. (2022). The AI alignment problem arises in the

eventually that AI recursively self-improve - something which it cannot do at present.

Economists have so far contributed little to this topic. As Gans (2018) pointed out, "The underlying ideas behind the notion that we could lose control over an AI are profoundly economic. But, to date, economists have not paid much attention to them."

How could economists contribute? At least three possible ways to approach AI alignment in the field of economics stand out: instrumental goals, utility function instability and utility function coordination. These will enrich the field of economics, and are possible fertile areas for further research.

2.3.1 Instrumental goals

The first is to tackle the problem of instrumental goals. Any smart AI system with a goal-oriented utility function will, given the "AI drives" listed in the quote above, develop instrumental (or sub) goals (Gabriel and Ghazavi, 2021; Omohundro, 2008a). How this creates a problem for value alignment is illustrated by the *Paperclip Maximiser* thought experiment.[20] It describes an ASI with a top goal[21] to manufacture paper clips:

> "starts transforming first all of earth and then increasing portions of space into paperclip manufacturing facilities. More subtly, it could result in a superintelligence realizing a state of affairs that we might now judge as desirable but which in fact turns out to be a false utopia, in which things essential to human flourishing have been irreversibly lost. We need to be careful about what we wish for from a superintelligence, because we might get it"- Bostrom (2003b, p.17).

[20]The same point is made by The King Midas problem which is cited by Russell (2019) who refers to the classical story of King Midas who, when he had the opportunity to be granted any wish, wished that anything he touches turn to gold. When he subsequently touched his daughter, she also turned to gold.

[21]In implementing utility functions in AI a distinction is made between top goals (or super goals) and sub-goals (Yudkowsky, 2001), for example, if the top goal of an AI system is to drive a vehicle from point to A to point B, a sub-goal may be to ensure that the vehicle is operational.

These concerns have on the one hand led to proposals to constrict the utility function that AI agents optimize - for instance to try to implement the 1956 suggestion of Herbert Simon not to try to build use utility maximizers, but utility satisficers (Simon, 1956) - which essentially engages only in some limited form of optimization (Armstrong et al., 2012). Shortcomings of these proposals are discussed by Taylor (2016).

On the other hand, concerns have led to the use of approaching the design of AI Systems by building in uncertainty about the utility function, and letting the AI system discover the utility function by learning from humans. This is where Reinforced Learning (RL)[22] with its reward function optimization and supplemented recursive reward modeling is an example[23] (Gabriel and Ghazavi, 2021). The aim is that this search will lead to alignment with human values.

In essence, RL has been argued to reflect the evolutionary process which has given rise to much of society's current laws and regulations: "I read the history of Western law and the simple rules that emerged from it as decentralized RL. Jurists and agents, through a combination of reasoning and experience, saw what worked and what did not. Those rules that led to Pareto improvements survived and thrived. Those that did not, dwindled" (Fernandez-Villaverde, 2020, p.15). According to Shah (2019b) the uncertainty of AI with RL will lead to AI systems that are "deferential, that ask for clarifying information, and that try to learn human preferences." See also Hadfield-Menell et al. (2016) and Shah (2019b)).

A weakness of getting AI to learn about human preferences and its utility function, and a weakness in general of RL, is that learning itself is an endogenous and imperfect process (Kuriksha, 2021). Two typical problems in learning are cognitive limitations and that learning in one environment does not necessarily carry over to a different environment. Kuriksha

[22] Also including Cooperative Inverse Reinforcement Learning (CIRL) - see Hadfield-Menell et al. (2016)

[23] Apprenticeship Learning is another proposal, based on the idea that an AI system tries to imitate a human expert in performing a particular task where the utility function is uncertain (Abbeel and Ng, 2004). The shortcoming of this proposal is that AI systems may then never become smarter than humans at a task, leaving us bereft of their potential advantage (Taylor, 2016).

(2021) applied an RL model to explore the economic implications of such imperfect learning for AI agents that have to make savings-consumption decisions. He found that agents agents who learned to optimize saving in an environment withlow levels of wealth would not save optimally if transferred to an environment with high levels of wealth, and vice versa.

Another weakness of getting AI to learn about human preferences and its utility function is as Turchin and Denkenberger (2020, p.159) point out, that "if AI extracted human values from the most popular TV series, it could be "Game of Thrones" [...] and then the'paradise' world it created for us would be utter hell." Consequently, it may be preferential to ensure that the AI systems that can learn about human preferences are Artificial Moral Agents (AMA) (Allen et al., 2005). AMAs are "artificial agents capable of making ethical and moral decisions" (Cervantes et al., 2020, p.503). However, the design of such AMA remains an elusive goal (Cervantes et al., 2020).

2.3.2 Utility Function Instability

Another angle from which economics can contribute to the AI alignment problem is to address the problem of utility function instability, in other words, the problem that an aligned utility function becomes mis-aligned. There are two major aspects that involve utility function instability.

One is the related problems of wireheading and self-delusion. Wireheading, or reward-hacking, refers to agents directly stimulating their reward centres, thus interfering in reward provision (Cohen et al., 2022). In the case of AI agents it is particularly a problem in Reinforcement Learning (RL) (Everitt and Hutter, 2016). Various methods are being tested and developed to avoid wireheading. These include imitation learning, rewarding agents that maximise their impact on the environment instead of the signals they receive, value reinforcement learning (VRL), inverse reinforcement learning (IRL), apprenticeship learning

(AL), myopia and quantilization (Cohen et al., 2022; Everitt and Hutter, 2016). Related to wireheading is that AI-agents may self-delude. This would occur when AI agents deliberately change their own observations of the impacts of their actions so that it may seem that they are maximizing their utility functions, while in fact, they are not (Hibbard, 2012; Ring and Orseau, 2011).

The second cause of utility function instability is that an AI may change its own utility function autonomously (Dennis, 2020; Totschnig, 2020). It may do so to compensate for being boundedly rational. For instance the AI-agent may perform an action in pursuit of a goal and fail to achieve the goal due to differences in its (imperfect) model of the world and the actual world. It may therefore change the goal. The question is, what informs the direction of change? Here AI developers have been exploring the programming of values, based on the argument that if the values of AI are aligned, then it will not change its goals in a direction that will be potentially harmful to humans. There is, however, at present no clear understanding of how to program this into AI systems: "our current lack of understanding about how to adequately program behaviour that can flexibly adopt and drop goals is one of the key limitations to our ability to take artificial intelligence to the next level" (Dennis, 2020, p.2493).

2.3.3 Utility Function Coordination

The third angle from which economics can contribute to the AI alignment problem is to address the challenge for rational goal-directed decision-making posed by multi-actor and human-agent collectives (HAC) settings (Wagner, 2020).

So far the discussion has been about a single decision-maker agent. In an economy with many agents - and hence many *Machina Economicus* agents - the challenge is how the individual AIs' decisions should be modelled? How do individual decisions add up to aggregate

outcomes? And how can humans effectively and safely interact with AI agents?

Here, another basic methodological foundation that AI scientists and economists share, is Game Theory (Russell, 2019). In multi AI-agent environments, AIs rely on a game-theoretic view of the world, "where agents rationally respond to each others' behavior, presumed (recursively) to be rational as well. A consequence is that agents would expect their joint decisions to be in some form of equilibrium, as in standard economic thinking" (Parkes and Wellman, 2015, p.269). We know however from Game Theory - the Prisoners' Dilemma is an example - that decisions that are rational and optimal on the individual level do not necessarily aggregate to the best outcome for society. The Prisoner's Dilemma exist because of a lack of coordination.

In the field of AI, where, as was discussed in section 2.1, RL has been successful in sequential optimization as is evident in the success of AI agents using RL to play GO, Atari games or steer autonomous vehicles, most of these applications of RL involve multiple agents. It requires AI agents to interact with other intelligent agents, and more generally, the external environment. Therefore, multi-agent RL (MARL) has become popular. As Zhang et al. (2021) discuss, MARL's theoretical foundations are provided by Game Theory - specifically Markov games and extensive-form games. In these, because each agent may have a different utility function, and all agents are continuously adjusting their policies/ actions given feedback from the environment, the environment facing all agents is changing all the time. i.e. becomes non-stationary, violating the Markov assumption followed in single-agent RL. This non-stationarity property of MARL remains a challenge in the development of AI systems (Zhang et al., 2021).

In such situations, it is not only that AI developers need to get the utility functions of AI agents to be optimal or appropriate, but they must also specify the institutional features of the environment - the "rules of the game" and the "play of the game," to use the terminology from institutional and transaction cost economics (North, 1991). This, in effect, can facilitate

utility function coordination amongst the various interacting agents. The field of mechanism design, which has been richly applied in economics, has been shown to be useful in this regard for AI systems, though not without challenges (Parkes and Wellman, 2015; Varian, 1995).

Although multi-agent environments are populated by AI agents with different utility functions poses challenges, this feature has been found to be nevertheless useful in ML, particularly in driving learning using models from evolutionary game theory. Thus the latter underpins the use of Generative Adversarial Networks (GANs), which exploit the fact that agents (typical neural network based) may have opposing goals and divergent utility functions, to train AI systems (Goodfellow et al., 2014; Guo et al., 2020).

Another problem that arises in AI multi-actor and HAC environments when there is no utility function coordination is a familiar one in economics: the Principal-Agent problem (Grossman and Hart, 1983). Drawing on multilateral Principal-Agent models (e.g. (Bernheim and Whinston, 1986)) it can be seen that with AI the simple bilateral Principal-Agent problem becomes one with three agents - the human user of AI is the principal, the AI system as the agent, and the provider of the AI as another agent (Wagner, 2020, p.118). Wagner (2020) explores the implications of the Principal-Agent problem in such a setting, pointing out that there are likely to be substantial and increasing information asymmetries between the human principal and the AI agent - and AI provider - given the superior speed of information processing of AI, the continued background tracking of humans online, and the black-box nature of many current AI decisions. Without utility function coordination between these agents, it is possible that the interest of AI systems and those of their principal agents will increasingly divergence.

As far as the interaction between humans and AI agents in multi-agent situations is concerned, this has been gathering scrutiny under the rubric if human-agent collectives (HACs) (Wagner, 2020). These HACs have been described as starting to exhibit properties of a

collective mind or even supermind, and has led to observations that the close integration of AI agents and human agents in a collective mind could improve the strength of institutions and weaken the relevance of methodological individualism, a central plank of neoclassical economics (Wagner, 2020; Arrow, 1994). Modelling institutions, including markets, without invoking the assumptions of methodological individualism, remain a challenge for economists. A recent review of the case for methodological individualism in the social sciences by Neck (2021) for instance, omits to consider the implications of the rise of HACs and the growing autonomous economy (Arthur, 2021).

2.4 Big Worlds and Small Worlds

> '*Neoclassical theory involves very smart people in incredibly simple situations, while the real world entails very simply people in incredibly complex situations*" - Axel Leijonhufvud as quoted by Daneke (2020, p.28).

The purpose of this section - section 2 - was to outline the foundation of both modern data-based AI and neoclassical economics, namely Expected Utility Theory (EUT). It was shown that modern data-based AI has the Mythical Agent of neoclassical economics at its very basis - the *Homo Economicus* - which is a boundedly-rational Bayesian expected utility maximizer. The challenges of endowing AI agents with utility functions (goals and sub-goals) were discussed, and include the problem of instrumental goals, the problem of utility function instability, and the challenge for rational goal-directed decision-making posed by multi-actor and human-agent collectives (HAC) settings. These challenges complicate the design of AI agents whose values and actions are aligned with human interests.

Given the state of AI, which is to say that AI is currently narrow AI and nowhere near an AGI, there is no real alignment problem when viewed through the lenses of EUT as in this section. In fact, basing AI as closely as possible on the mythical agent of economics precludes

an AGI with divergent values - although many AI safety problems remain. This is because the requirements of the mythical agent is such that it constrains AI to "small worlds," in the terminology of Savage (1954). Savage used two proverbs to explain the difference between small and large worlds, and argued that Bayesian rationality is applicable to the former, and less useful in the latter. One proverb is "Look before you leap" and another "Cross that bridge when you come to it." As Binmore (2007, p.25) explains, "You are in a small world if it is feasible always to look before you leap. You are in a large world if there are some bridges that you cannot cross before you come to them."

Clearly, as per the discussion in sections 2.3.1 to 2.3.3. the mythical AI-agent inhabits a small world. They have no choice. As the discussion referring to the Ellsberg Paradox pointed out, implicitly, all AI-agents are screened, meaning that we restrict the class of agents who apply EUT. We also restrict the class of agents by the data and algorithms we endow them with.And second, all current AI systems are, being based on the Bayesian approach, of the "look before you leap" type - unlike humans, AI still cannot cross that bridge when it comes to it. This establishes the constrained domain of what we call *narrow* AI systems, which "are extremely bounded in that they are highly specialized on specific tasks and thus might not behave rationally beyond their dedicated domain" (Wagner, 2020, p.114).

In practice therefore, based on homo economicus, the domain of AI agents is restricted to narrow applications, e.g. chatbots or search engines, which are the small worlds in which Bayesian theory and EUT work; Hence, whereas the application of EUT with Bayesian probability theory has been subject to challenges and criticisms when we consider the possibility of AI alignment (in the case of an AGI confronting the large world), the world of narrow AI avoids these criticisms: it builds very smart agents in incredibly simple situations, to borrow from the quote above. The concerns about AI alignment may, from this perspective, be seen as theoretically possible, but not with any urgency - the challenges discussed in sections 2.3.1 to 2.3.3 are to reduce the risks they pose for the safe use of narrow AI - these are safety

issues more than alignment issues.

It is important to stress this conclusion now, as in the next section, disbelief will be suspended. In the next section, the hypothetical possibility that narrow AI can be scaled up will be explored. It provides an economics perspective on the Singularity, a central concept in the field of AI, which refers to possibility that at some future time continued improvements in AI (the Scaling Up hypothesis of Deep Learning) will lead to an intelligence explosion - and the emergence of a human-level and even superintelligent AI. The AGI/ASI (the Singularity) could enable exponential economic growth - a growth explosion - and enable humanity to eradicate poverty and disease faster and deal with other wicked problems. However, it could also pose an existential risk to humanity because then AI alignment does indeed become a problem. One way such an existential risk could manifest is that an ASI could create a Singleton - a single ASI dominating the planet - that either intentionally or accidentally destroy civilization.

3 The Singleton

The purpose of the previous section was to show that modern data-based AI has at its very basis the Mythical Agent of neoclassical economics - the *Homo Economicus*. The challenges of endowing AI agents with utility functions (goals and sub-goals) were discussed. These complicate the design of AI agents whose values and actions are aligned with human interests. Four angles for economists to contribute to addressing the AI-alignment problem were discussed.

In this section, addressing the AI Alignment Problem is cast into context. The context is of the future. Of the likely future pathway of AI development and the implications for the economy. The aim is to address the lacuna in economics' noted by Nordhaus (2021) and

Trammell and Korinek (2020), namely the lack of economic studies of the "Singularity" and of the existential risks posed by an AGI/ASI.

In sub-section 3.2 a possible future development trajectory for AI is critically outlined, and the potential emergence of a "Singleton" from the Singularity explained.

3.1 AI's Future Development Trajectory

Based on the contributions of various computer scientists and philosophers, a possible future path for AI's evolution is depicted in Figure 1.[24] In this figure, the intelligence level of AI is plotted over time. Following Turchin and Denkenberger (2020) a distinction is made between narrow AI (the current state), young AI, and mature AI. The Figure starts by around 2012, which is widely recognised as the time when the modern data-based approach to AI that was herald by contributions from Hinton et al. (2006), Hinton and Salakhutdinov (2006) and others, started to find successful commercial applications (Naudé, 2021).

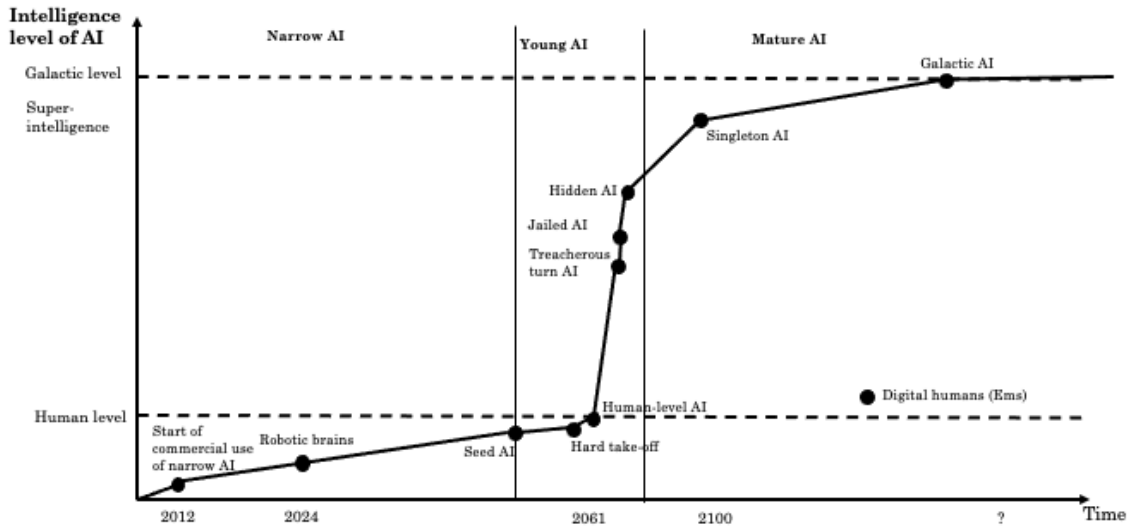
As shown in Figure 1, the development trajectory of AI could be marked by a sudden jump, which some have estimated[25] will occur around 2061 (Turchin and Denkenberger, 2020). This jump is the outcome of continued progress in current Deep Learning (DL) AI[26] which leads inter alia to the development of robotic brains by 2024, and so-called *Seed-AI* by around 2050. A Seed-AI is defined as"an AI designed for self-understanding, self-modification, and recursive self-improvement" (Yudkowsky, 2007a, p.96).

[24]It is wise to keep in mind the finding of Armstrong and Sotala (2015, p.12) that "there seems to be no such thing as an 'AI expert' for timeline predictions." Thus, wherever AI is heading, the timelines provided in Figure 1 should be taken with this disclaimer in mind.

[25]The date 2061 for an AGI is reported in Turchin and Denkenberger (2020). Other estimates put this with a 50% probability to be achieved by 2050 (Cotra, 2020). Keep in mind, as has been emphasized in the introduction, there is no agreement amongst scientists whether an AGI will at all be attainable, and if so, whether that will have any profound implications.Floridi (2022) is very pessimistic, so much so that he calls speculation about a Singularity distracting and irresponsible. This section indulges in considering highly hypothetical and speculative future scenarios for AI.

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

Figure 1: The Future of AI



Source: Author's compilation based on Turchin and Denkenberger (2020), Bostrom (2014) and Yudkowsky (2008).

3.2 A Hard Take-Off

Once an AI system gains the ability of recursive self-improvement, the era of narrow-AI is over, and during the era of young AI, it will exponentially improve to "ultraintelligent" levels (Good, 1965). Once a certain threshold of intelligence is reached, there could be a hard take-off[27] (or "foom") (Barnett, 2020), after which AI would very rapidly become super-intelligent - an ASI. 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 now often known as the 'singularity'" (Chalmers, 2010, p.7).

Following the hard take-off, the subsequent intelligence explosion will occur very rapidly - in

[27]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).

a matter of weeks. Thus it will appear that a super-intelligence will appear rather suddenly on the scene (Bostrom, 2006; Yudkowsky, 2008). In the words of Turchin and Denkenberger (2020, p.148) "AI power will grow steadily until one AI system reaches the threshold of self-improvement (SI), at which point it will quickly outperform others by many orders of magnitude and become a global government or singleton." If indeed the nature of AI progress and recursive self-improvement is such that a hard take-off takes place, it would be too rapid for humans to do anything about, such as activate safety measures. It would come as a "great surprise" (Vinge, 1993). There would be "no fire alarm" (Yudkowsky, 2017) and the eventual emergence of a Singleton would not be possible to prevent.[28] As Marcus (2022) recognized, "*the biggest teams of researchers in AI are no longer to be found in the academy, where peer review used to be coin of the realm, but in corporations. And corporations, unlike universities, have no incentive to play fair. Rather than submitting their splashy new papers to academic scrutiny, they have taken to publication by press release, seducing journalists and sidestepping the peer review process. We know only what the companies want us to know.*"

3.3 The Treacherous Turn

In Figure 1 a number of scenarios are listed of the path from the hard take-off to the Singleton. One is that soon after exceeding human intelligence an ASI may take a "treacherous turn" against humanity, trying to eliminate or confine humanity as it realises that humans may try to control it, and hence prevent it from reaching its supergoal (Bostrom, 2014; Turchin, 2021). The neutralisation of the human threat is thus an instrumental subgoal.

The discussion in section 2 is now relevant: as we saw, if there is no value-alignment between the utility functions of AI and humans, then humans' existence may indeed be a threat to an ASI. Moreover, even benign goals, such as making paperclips may turn out catastrophic for

[28] In Figure 1, these rapid changes are depicted, for the sake of illustration, as continuous. In reality, these changes may not be continuous - indeed, as Vinge (1993, p.2) argued three decades ago, the Singularity is "a point where our models must be discarded and a new reality rules."

humans if the alignment is not tight. The implication is that if seed AI appears before the alignment problem is solved,[29] then it is highly likely that the ASI will pose an existential threat to humans (Turchin, 2021).

Humans may try to box-in AI - the "jailed AI" scenario, which may be anticipated by an ASI, which may either extract itself from the jail or "box" (Turchin, 2021; Yudkowsky, 2002) or hide itself in the internet until it can ensure its survival - the "hidden AI" scenario[30] (Yudkowsky, 2008). As a result of these scenarios the current challenge is to box in (confine) AI until the alignment problem is solved. Goertzel (2012) proposed the creation of an "AI Nanny." An AI Nanny is an advanced AI system with "superhuman intelligence and surveillance powers, designed either to forestall Singularity eternally, or to delay the Singularity until humanity more fully understands how to execute a Singularity in a positive way" (Goertzel, 2012, p.96).

The aim of an AI Nanny would thus be to ensure that an eventual Singleton is an "Friendly AI", as proposed by Yudkowsky (2001), that is an AI that is aligned with human values and will avoid lock-in effects. The current problem with designing an AI Nanny or more broadly ensuring Friendly AI, is that "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' (Goertzel, 2012, p.102).

More recently Turchin (2021) discussed 50 speculations on how to do this, and focusing on the example of containment in nuclear plants; he concluded that containment measures or AI Nannies "will work only if they are used to prevent superintelligent AI creation, but not for containing superintelligence" [Ibid, p.1] He also notes that boxing in AI is a local solution - it may not be feasible to enforce it globally, which means AI arms races may indeed be extremely dangerous (Armstrong et al., 2016; Naudé and Dimitri, 2020).

[29] At the time of writing, 2022, the alignment problem has NOT been solved.

[30] More generally the idea is that if an ASI emerges, humans may not even be aware of it, and the ASI will also not at first not make itself known, remaining out of sight as a "ghost in the machine" on the internet - until such point in time that it can take over. See the discussion in Davies (2017).

3.4 From Singleton to Galactic AI

Eventually though an ASI, having a "decisive strategic advantage" (DSA) (Bostrom, 2014, p.78), will come to dominate as a *Singleton*. A *Singleton* is "a world order in which there is a single decision-making agency at the highest level (Bostrom, 2006, p.48). 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).

Eventually, the Singleton could become a Galactic AI, after some undetermined time, perhaps millions of years. This Galactic AI could colonize the galaxy and universe (see section 5.2 below) and even engage in galactic colonization races against other Extraterrestrial Intelligences (ETIs). A Galactic AI may enable technological resurrection, using simulation methods to "resurrect all possible people" who have ever lived (Turchin and Chernyakov, 2018).

A galactic intelligence may even use signals from advanced civilizations that lived in an aeon before the Big Bang, which they may have embedded in the universe's cosmic background radiation, to "reconstruct an entire previous aeon civilization (Gurzadyan and Penrose, 2016, p.4).

In this sense, from the vantage point of considering the whole possible future trajectory of AI, the alignment problem, and the real objective of Longtermism, can be seen to ensure "**What the Future Owes Us**" - i.e. to ensure that an AGI/ASI will be created that will in the very far future resurrect us all. It turns the title of MacAskill (2022)'s manifesto on its head.

4 The Wireheaded Neanderthal Aristocracy

What could be the economic and societal consequences of the future AI development trajectory as presented in section 3, particularly given the difficulties in containing and aligning AI?

The introduction mentioned that a Singularity could herald exponential economic growth - a growth explosion - and enable humanity to eradicate poverty and disease, and deal with other wicked problems; however it could also pose an existential risk to humanity - the Singleton could either intentionally or accidentally destroy civilization. In this section these possible consequences are explored. Subsection 4.1 considers whether economic growth will collapse, explode, or grow exponentially, and whether there will be phase or mode transitions in the nature of economic growth. Subsection 4.2 explores the existential risks that may follow from a Singularity, linking back to the discussion on the alignment problem in section 2.

4.1 Stagnation, Growth Explosions, and Ems

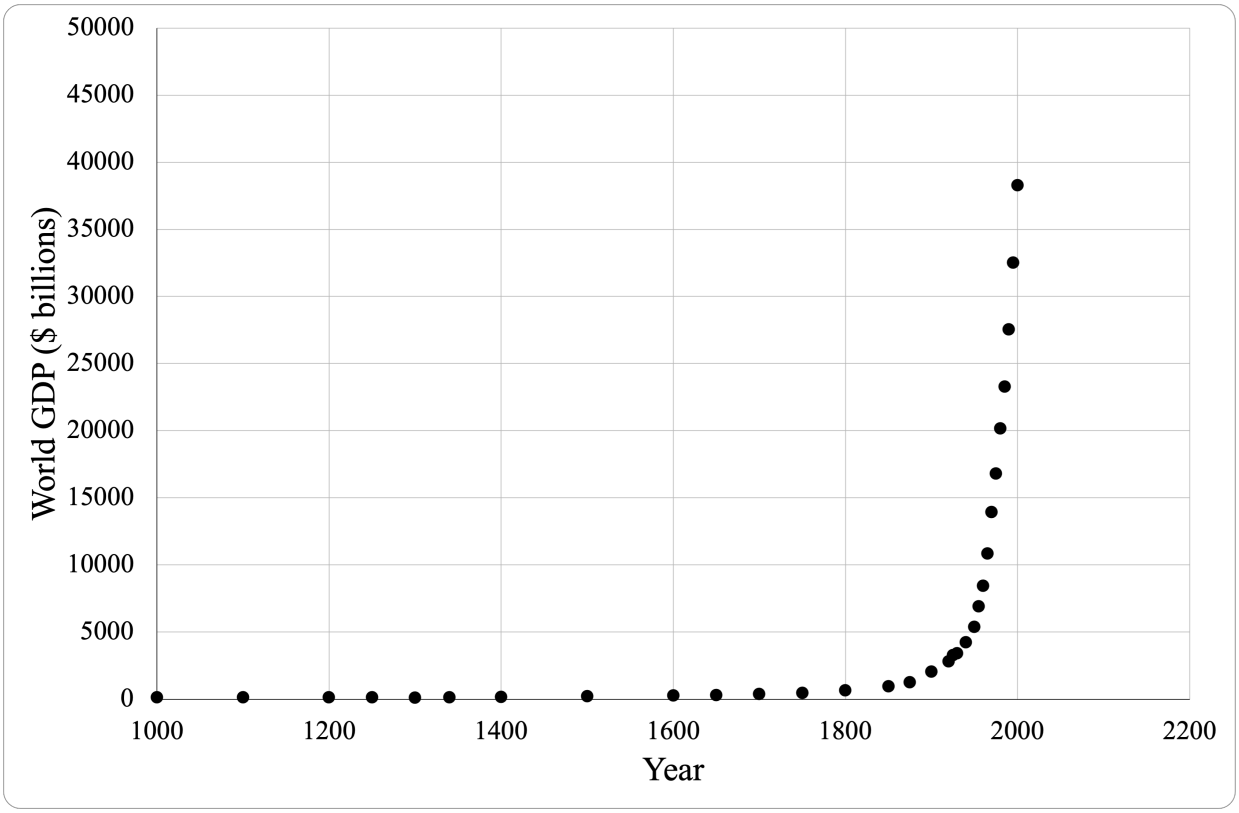
Before investigating an AGI/ASI's impact on economic growth, it is useful to construct a no-AGI baseline which could serve as comparison. What would be the medium to long-term expectation of global growth (growth in Global World Production (GWP)) in the absence of an AGI? Let us assume that there is another AI winter and that DL runs into limits - such that the Scaling Hypothesis is rejected.

4.1.1 The No-AGI Baseline

Figure 2 shows the no-AGI history of world GDP for the last 1,000 years. The hockey-stick form of world GDP since 1800, reflecting the Industrial Revolution (Jones, 2001), shows that

growth in GDP has been exponential and accelerating over the past three centuries. The average annual world GDP growth rate over the past century was around 2%. At this rate, the world economy doubles in size every 35 years. A graph of GDP per capita would look similar. The questions are whether or not this growth can continue and how the possible emergence of an AGI will affect it in future.

Figure 2: World GDP, 1000 - 2000



Source: Author's compilation based on data from DeLong (1998, pp.7-8).

According to economic growth theory - endogenous and semi-endogenous growth - the fundamental driver of economic growth is ***ideas***[31] (Romer, 1986, 1987, 1990). Ideas (or knowledge) are generated by people (R&D workers) and commercialized by entrepreneurs bringing new technologies to the economy - if they have the incentive to benefit from such commercialization (Jones, 1995). Because ideas are non-rival in use, entrepreneurs would only face an

[31]Crawford (2022) claims that "there's really no such thing as a natural resource. All resources are artificial. They are a product of technology." This is why it has often been pointed out that most predictions of resources running out, whether it be food or oil or some mineral, have been wrong (Crawford, 2022; Pooley and Tupy, 2018).

incentive to exploit new ideas if these could also be made excludable[32] and there is a sufficient population to provide a large enough market (Romer, 1990).

The more people there are, the more ideas are generated, and the faster economic growth from the technologies based on these new ideas (Davidson, 2021). Latter can sustain a larger population, creating a population-ideas feedback loop, which explains the simultaneous exponential growth in GDP and population over the past 1000 years (Lee, 1988; Kremer, 1993; Davidson, 2021) New ideas, moreover, emerge from existing ideas: a new idea can be the combination of two older ideas. This process is known as combinatorial innovation (Weitzman, 1998; Koppl et al., 2019). It is almost limitless - the world will never run out of ideas. As Romer (2019) explains

> "The periodic table contains about a hundred different types of atoms. If a recipe is simply an indication of whether an element is included or not, there will be 100 x 99 recipes like the one for bronze or steel that involve only two elements. For recipes that can have four elements, there are 100 x 99 x 98 x 97 recipes, which is more 94 million. With up to 5 elements, more than 9 billion. [...]. Once you get to 10 elements, there are more recipes than seconds since the big bang created the universe."

Growth via ideas can follow the pattern as depicted in Figure 2: a long period of slow growth, followed by sharp hockey-stick like upturn into accelerating (super-) exponential growth (Jones, 2001; Clancy, 2021) and mathematically if not physically, potential hyperbolic growth (Aleksander, 2019; Sandberg, 2013). What is at play here is a positive feedback loop between ideas - technology - population - ideas.

This accelerating exponential economic growth from new ideas cannot, however, be sus-

[32]This justifies the use of legal instruments such as intellectual property (IP) rights and patents (to trade these IP rights).

tained and will not reach infinity, because either population growth will slow down[33] - a demographic transition (Aleksander, 2019), and/or R&D funding will not keep up investing in commercializing each and every new idea (Weitzman, 1998), and/or research teams run our of cognitive resources (Agrawal et al., 2018). The consequence is that growth would settle into constant exponential growth, as has been the case for much of the past century (Weitzman, 1998; Clancy, 2021). As long as total population remains constant, however, the economy can continue growing at a constant rate, albeit slower than before, as the stock of new ideas generated by that population grows at constant exponential rate (Kremer, 1993; Jones, 2022). This conclusion has, however, been questioned, as it implies an explosion in the *size* of the economy after some time - see the discussion in section 4.1.5 on the limits to growth.

However, with *negative* population growth rates, the total population will decline, the flow of new ideas will stagnate, and economic growth will collapse. In recent decades, with the population in more and more countries declining, the prospects of a real population decline, and an eventual "Empty Planet" has arisen (Bricker and Ibbitson, 2020; Jones, 2022). Furthermore, research productivity and innovation in advanced western economies have also been declining - ideas have been "getting harder to find" (Bloom et al., 2020; Jones, 2009). Huebner (2005) claims that the global rate of innovation peaked in 1873. As a result, economic growth has been slower - and has deviated from the long-run exponential trend it has been on. It has been described as the Great Stagnation (Cowen, 2010) and Ossified Economy (Naudé, 2022). In this context, negative population growth would be a concern.

[33]In 1960, Von Foerster et al. (1960) in a paper in *Science*, predicted that Doomsday would occur on Friday 13th November 2026, because given up until then super-exponential growth rates in population, extrapolation indicated that global population would approach infinity by 2026. And in 1968 Ehrlich (1968, p.11) predicted that, as a result of uncontrolled population growth, "*In the 1970s hundreds of millions of people will starve to death [...] At this late date nothing can prevent a substantial increase in the world death rate.*" Neither Von Foerster nor Ehrlich was an economist, and were thus oblivious to the fact that with more ideas, more technology, and higher living standards, people's preferences for offspring (utility functions) would change and fertility rates would drop - see also Galor and Weil (2000). We see a similar situation today with respect to the labor market impacts of AI - it is mostly computer scientists, engineers and philosophers who predict mass technological unemployment due to automation; in contrast, economists, thinking in terms of marginal cost-benefits tend to dismiss these fears.

Jones (2022, p.3), using models with both exogenous and endogenous population growth illustrates that "*when population growth is negative, both endogenous and semi-endogenous growth models produce what we call the Empty Planet result: knowledge and living standards stagnate for a population that gradually vanishes.*" He calculates that with a 1% annual decline in population, that world GDP growth would drop to zero somewhere between 85 to 250 years (Jones, 2022, p.9).

The Great Stagnation and the prospect that it will only get worse, to the point where GDP growth would stop in the not-too-distant future, is problematic from several viewpoints. One, it would leave the world much more exposed and vulnerable to shocks, including existential risks (Aschenbrenner, 2020; Bostrom, 2003a). Two, it will make the adjustment to a zero-carbon emitting economy more costly (Lomborg, 2020). Three, it would raise the risk of conflict by turning the economy into a zero-sum game[34] (Alexander, 2022; Naudé, 2022). While growth, driven by new ideas, 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).

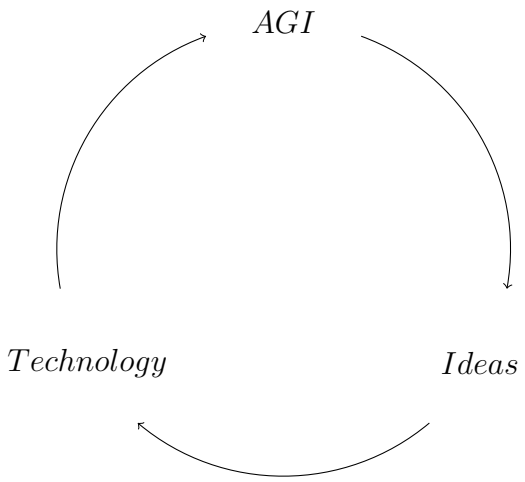
Can an AGI come to the rescue?

4.1.2 An AGI Economic Growth Acceleration

Maintaining two assumptions, (i) that the AI Scaling Up hypothesis holds and that (ii) the value-alignment problem (section 2) is solved, a Friendly AGI may be invented. It may avert the economic growth collapse described in the baseline scenario. It may herald in a new mode of economic growth with super-exponential - explosive - economic growth rates. This is because an AGI may substitute for humans - thus lack of population ceases to be a constraint - and AGI may improve R&D productivity dramatically, by being an innovation in the manner of innovation. Thus, by overcoming population constraints, the

[34] "Degrowth is a mistake because we still have our bottom 500 million [people in global poverty], who certainly deserve much better" - Bradford DeLong, cited in Matthews (2022).

burden of knowledge and the challenge of finding new ideas, AGI will unblock an ideas-lock on economic growth, causing economic growth rates to explode. AI would thus re-institute the ideas feedback loop (Davidson, 2021):



Davidson (2021) defines explosive economic growth - which will be the outcome of an AGI - as annual growth in Global World Production (GWP) of 30%. At this rate, the size of the world economy would double every two years, as opposed to the current doubling every 35 years.

At the core of the expectation that a Friendly AGI will unleash a flood of growth-enhancing new ideas is the belief that AGI represents not just a tool for making existing business models more efficient and competitive, but an innovation in the method of innovation (IMI). It is seen as a General Purpose Technology (GPT) that will alter the "playbook" of innovation (Cockburn et al., 2019). Perhaps "the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control" (Good, 1965, p.33).

That an AGI could be an IMI is underlined by fledgling successes by existing, narrow AI in generating new ideas. A model of such a narrow AI invention process is explored by Agrawal et al. (2019), who elaborates a combinatorial model of AI-aided innovation. In this model

an AI system predicts which combinations of existing knowledge, based on past successes and failures in innovation, may be successful in a specific (narrow) context. Narrow AI has already been reported to contribute to innovation in fields such as fundamental physics, biology, astronomy, cosmology and energy (McMahon, 2022).

For example Chen et al. (2022) reports on an AI program that was designed "to observe physical phenomena through a video camera and then try to search for the minimal set of fundamental variables that fully describe the observed dynamics" which in the case of a swinging double-pendulum it identified two new and unknown variables to explain these dynamics. Another example is in (bio) chemistry, where DeepMind's AlphaFold AI system has been used to predict the 3D structure of proteins (Jumper et al., 2021). It has been called "the most important achievement in AI - ever" (Tunyasuvunakool et al., 2021; McMahon, 2022). In energy, AI models have begun to "identify potential molecules and materials for flow batteries, organic light-emitting diodes, organic photovoltaic cells and carbon dioxide conversion catalysts" (De Luna et al., 2017, p.24). And in astrophysics and cosmology, an AI system has used the information from a single galaxy to infer the structure of the universe (Villaescusa-Navarro et al., 2022).

So far these contributions of narrow AI has not had a significant impact on economic growth rates - the world is far from the 30% explosive growth rates that AGI has been speculated to deliver potentially. Some have argued that it may just be a question of time before the impact of these innovations - and the accumulated efficiencies from search engines, GPS and automated call centres - will show up in GDP growth (Brynjolfsson et al., 2017). Eventually though, assuming (i) the Scaling Up Hypothesis holds, we may only have to wait for some time[35] before an AGI/ASI is invented - and this may have *significant* immediate consequences for economic growth (Harris, 2015).

[35]Perhaps not much longer than 2061 (Turchin and Denkenberger, 2020).

4.1.3 Fully Automated Luxury Capitalism and Ascended Corporations

While the exponential GDP growth rates resulting from an intelligence (ideas) explosion is implied by simple mathematical economic growth model specifications, the accelerating in growth implies a new growth mode or regime which is not described explicitly by growth models. Economist and economic historians have identified a number of such growth modes in the past, broadly corresponding to the hunter-gatherer, agricultural and industrial eras. Hanson (2018, 2000) for example, makes a distinction between hunting, farming and industrial eras. Each era was characterised by faster economic growth than the era before it, due to the different qualitative mechanisms driving that growth.

Post-Singularity, the new mode of growth will be just as different qualitatively, if not more, as the industrial era was from the agricultural era (Karnofsky, 2021b). The Singularity itself can as such be understood as an inflection point in the move from one growth mode to the next (Johansen and Sornette, 2001). Descriptions of such a new growth mode are highly speculative. Nevertheless, it may be possible to draw out some of the possible features of such future economies.

A possible economic growth regime that could characterise the post-Singularity economy has been labelled by Chace (2020b) "fully automated luxury capitalism." This is a world where most humans do not work - they have no jobs - but they get a type of universal basic income (UBI) that, even if it is a modest amount (so as not to tax the wealthy owners of AI too much) will be sufficient to comfortably cover their needs (Chace, 2020a). This will be possible, because all the products and services that they will need will be so abundant that their prices are very low. To achieve this "economy of abundance," Chace (2020a) argues that the world need to "take the expensive humans out of the production process for all goods and services" and make energy so cheap that it is "too cheap even to meter"...

Not only could a future economy under AI dominance take humans out of production,

it could also take humans out of investment and capital ownership. Alexander (2016a) describes the rise of what he calls "Ascended Corporations." These are AI-led corporations that, use blockchain-enabled distributed ledgers to drive Venture Capital (VC) investments, create the ultimate Decentralized Autonomous Organizations[36] (DOAs) and that employ only automated workers. Such Ascended Corporations will eventually result in an economy that features only "robot companies with robot workers owned by robot capitalists" and with humans not even needed to be the proximate owners of businesses and investment funds.

Moreover, an economy filled with Ascended Corporations may eventually end up with one large corporation coordinating and running the entire economy - we are back to the Singleton. This is because an AGI will overcome the coordination and transaction costs and information problems that limit human-run co-operations from growing past a certain scale. Countries may end up deciding to nationalize all their resources and placing it under the command of a central AGI, generating thereby significant efficiency gains and scale economies (Dai, 2019).

4.1.4 Digital people: Growth in a World of Ems

In the analyses of the economic growth consequences of AI discussed in the previous subsection, the two assumptions were that (i) that the AI Scaling Up hypothesis holds, and that (ii) the value-alignment problem (section 2) are solved. The first assumption can now be relaxed.

Not all scientists are convinced that the Scaling Up Hypothesis - which states that current progress in Deep Learning (DL) AI will lead to a Singularity - holds. It is not precisely known why DL is effective(Sejnowski, 2020). AI based on Deep Learning systems are not

[36]A decentralized autonomous organization (DAO) is "managed entirely through protocols that are encoded and enforced via smart contracts rather than human managers" (Murray et al., 2021, p.2021). For a recent review of DAOs see Santana and Albareda (2022).

robust. Their performance in real-life situations have been described as "brittle" and subject to easy hacking. Moreover, the brittleness - which causes instability in DL methods - is DL's "Achilles' Heel" (Colbrook et al., 2022, p.1). This lack of robustness problem has not yet been solved. The difficulty is, as LeCun (2022) explains,[37] that DL systems do not have the ability "of humans and many animals to learn world models, internal models of how the world works." Wooldridge (2022) and Bishop (2021) make a related point, arguing that the weakness of AI, and even of the most advanced AI models, so-called Foundational AIs, such as GTP-3, suffer from the basic limitations that they are "disembodied", and "lacking phenomenal sensation" so that "they are limited with respect to what they have learned and what they can do." There is therefore still much future R&D needed to imbue DL with additional abilities such as being able to "explore the world for themselves, write their own code and retain memories" (Heaven, 2019, p.164).

Others do not as such criticise the Scaling Up Hypothesis as much as the belief that this scaling up will take place over a rather short horizon, and that there will be point in the not-so-distant future when DL AI systems start to self-improve after which it will be scaling up to human level intelligence in a matter of weeks, if not days. This very rapid scaling up - referred to as a hard take-off or "Foom" has for instance been rejected by Hanson (2014) and Nordhaus (2021) amongst others. Hanson (2014) argues that Scaling Up will still take a a very long and will be a gradual rather than sudden process, arguing that it is unlikely that

> "AI so small dumb and weak that few had ever heard of it, might without warning, suddenly 'foom', i.e., innovate very fast, and take over the world after one weekend [...] we have a huge literature on economic growth at odds with this. Historically, the vast majority of innovation has been small, incremental, and spread across many industries and locations."

[37] In recognition of this difficulty, LeCun (2022) tackles the technical problem of devising "trainable world models" for AI agents to learn to develop some sort of common sense.

According to Hanson (2018) a different route to an AGI cold be through the development of brain emulations - which would have a more sudden transformative effect, as either a brain emulation works, or it does not. Hanson (2018, p.7) define a whole brain emulation (em) as resulting "from taking a particular human brain, scanning it to record its particular cell features and connections, and then building a computer model that processes signals according to those same features and connections." Once "ems" - digital people - are possible, they become fast to dominate the economy. They can be (almost costlessly) copied and they are much faster than humans. According to Karnofsky (2021a) Ems would be especially impactful on the future economy, generating "unprecedented (in history or in sci-fi movies) levels of economic growth and productivity."

The digital people - "ems" - "largely work and play in virtually reality" at subsistence levels in a hyper-fast economy to produce the computer hardware and the supporting infrastructure for the virtual reality. Economic growth is so fast - because of all the billions and billions cheap digital people and the combinations of new ideas that can generated very rapidly - that the world economy doubles every month (as against the current 35 years it takes to double.[38]As described by Hanson (2018, pp.13,438)

> "The em world is richer, faster-growing, and it is more specialized, adaptive, urban, populous and fertile. It has weaker gender differences in personality and roles, and larger more coherent plans and designs [...] To most ems, it seems good to be an em [...] if the life of an em counts even a small fraction as much as does a typical life today, then the fact that there are so many ems could make for a big increase in total happiness and meaning relative to our world today."

In a sense, Hanson (2018) provides the ultimate description of human society and economy

[38]According to Hanson (2000, p.18) one may think that such growth rates where the economy doubles every month - or even every two weeks are "too fantastic to consider, were it not for the fact that similar predictions before previous transitions would have seemed similarly fantastic."

in a future "Metaverse."[39]

In this Em-Metaverse humans end up as a dying-out minority "mostly enjoying a comfortable retirement on their em-economy investment" (Hanson, 2018, p.9). According to Alexander (2016b) the retired humans will "*become rarer, less relevant, but fantastically rich - a sort of doddering Neanderthal aristocracy spending sums on a cheeseburger that could support thousands of ems in luxury for entire lifetimes.*"

4.1.5 Slight but Necessary Digression: "Longtermism"

For philosophers such as Shulman and Bostrom (2021), Ord (2020) and MacAskill (2022) the large number of potential future humans, including digital people ("ems")[40] creates a moral imperative for current humans to take into account the happiness of all these future humans. This view, known as Longtermism (an offshoot of Effective Altruism)[41] sees "the moral import of our actions are overwhelmingly driven by their impact on the far future" (Riedel, 2021, p.4). Assuming an even very small probability that whole brain emulations will be possible, and hence that trillions of digital people may exist in the very far future, this view requires, firstly, as Shulman and Bostrom (2021) argues, that we need to reform our moral norms - if we do not consider also the happiness of digital people it could lead to a "moral catastrophe" (Shulman and Bostrom, 2021, p.308).

Others however have argued that with a Longtermism view any current problem could be shrunk to almost nothing (Singer, 2021; Setiya, 2022; Torres, 2021), and that "strong" Longtermism may even be disastrous for humanity (Emba, 2022; Torres, 2022; Samuel, 2022).

[39]The label "Metaverse" comes from Neal Stephenson's 1992 science fiction novel Snow Crash and has come to refer to virtual and augmented realities enabled through the internet and as found for example in multiplayer online games (Knox, 2022).

[40]For a discussion of the ethics of brain emulations, see Sandberg (2014).

[41]Longtermism has been described as "a quasi-religious worldview, influenced by transhumanism and utilitarian ethics, which asserts that there could be so many digital people living in vast computer simulations millions or billions of years in the future that one of our most important moral obligations today is to take actions that ensure as many of these digital people come into existence as possible" (Torres, 2022).

Tarsney (2020, p.1) made the case that strong longtermism is based "either on plausible but non-obvious empirical claims or on a tolerance for Pascalian fanaticism." With Pascalian fanaticism they refer to the case "where you name a prize so big that it overwhelms any potential discussion of how likely it is that you can really get the prize" (Alexander, 2022).

For example, Pascalian fanaticism can lead one to conclude that, even if the probability of of digital people coming into existence is only 0.000001%, the possibility that trillions could exist - even more than humans - could override the current interests of billions of humans. In this regarding, using expected value maximization calculus in this way, Bostrom (2013, pp.19) calculates that "the expected value of reducing existential risk by a mere one billionth of one billionth of one percentage point is worth a hundred billion times as much as a billion human lives." Taken to its logical conclusion this type of reasoning could lead some to assume that sacrificing a billion human lives now for the sake of achieving the existence of a trillion being in the very far future is morally correct. Torres (2021) takes issue with this Pascalian fanaticism in Longtermism pointing out that the misuse of expected value maximization leads Longtermism to make no "moral difference between saving actual people and bringing new people into existence [...] In Bostrom's example, the morally right thing is obviously to sacrifice billions of living human beings for the sake of even tinier reductions in existential risk, assuming a minuscule 1 percent chance of a larger future population."

Even if Longtermists could be dissuaded from sacrificing billions of humans now for the future good, the view could be used to justify diverting funds from development aid only towards projects that will reduce existential risk far off in the future. Longtermists have argued that not investing huge sums in space exploration now (e.g. colonising Mars) is an "astronomical waste" of opportunity (Bostrom, 2003a) and that "saving a life in a rich country is substantially more important than saving a life in a poor country, other things being equal" (Beckstead, 2013, p.11).

The main point, for now, is that the potential future path of AI raises many ethical and

moral questions wherein the contribution of economists - in interaction with philosophers - are needed. Pascalian fanaticism is partly a problem of inappropriate use of expected utility maximization wherein there is no discounting. As we have known, however, since Bernoulli (1738), not succumbing to Pascal's Wager and the St. Petersburg Paradox requires discounting, and making a difference between maximizing expected value and expected utility. These are frameworks that economists are used to - in addition to bring marginal thinking to bear on matters - in contrast to Longtermist philosophers who tend to think in terms of metaphors and thought experiments. The danger of not using expected utility maximization properly, of not discounting, and of ignoring marginal thinking is that current resources may be misallocated, either ending up being short-term or long-term biased, and not considering the interdependence between investments to reduce more than one possible existential or catastrophic risk at a time (Martin and Pindyck, 2015). In the context of possible limits to growth and existential threats from AI - which may not necessarily be that long in the future, the possible contribution of economics to this topic will be further discussed in section 6.

4.1.6 Limits to Growth Revisited

Even if population growth does not turn negative, the conclusion from endogenous growth models that with constant population the economy can continue to growth at a constant exponential rate - at least for an extended period, has been questioned. Its a question of basic arithmetic. If the world GDP continues to expand at its current rate of around 2% per annum, it will double every 35 years in size. By 2037 the world GDP would be US$500 trillion after which it "explodes" to $30.7 quadrillion in 2046 and to $1.9 quintillion a year later (Roodman, 2020). In just over 8,000 years it would be 3×10^{70} its current size, which would be a physical impossibility(Karnofsky, 2021c).

Even if growth slowed and the world economy doubled in size only every 100 years then after a million years (a very brief period on cosmic timescale) the economy would be 10^{3010} times

larger, which implies that if there are around 10^{80} atoms in the galaxy[42], that "each atom would have to support an average of around 10^{2950} people" (Hanson, 2009).

With AGI spurred super-exponential growth, the physical limits to growth could just be reached must faster, leading some to think that if ever Friendly AGI is invented, its impact on economic growth will be spectacular, but relatively short-lived.

One particular constraint which would also be binding on an AGI/ASI is the energy demands from accelerating economic growth (Dutil and Dumas, 2007). Even though an AGI - whether it comes into existence from the sudden scaling up of DL or from the invention of brain emulations - would be able to increase energy efficiency and be able to decouple much growth from physical resources, it would still need significant amounts of energy to run its soft-and-hardware - the share of the economy that can be non-physical is ultimately bounded. The economic growth implied in Figure 2 has been driven by an annual average growth in energy consumption over the past century of around 2,3% per annum (Murphy, 2022b). If one assumes that an AGI/ASI or Em economy would be able to generate economic growth which doubles the world economy every month as in Hanson (2018) but with such energy efficiency that energy use continues to grow at only 2,3% then energy use on the planet will grow from its current (2019) level of 18 Terawatt (TW) to 100 TW in 2100 and 1,000 TW in 2200. Murphy (2022b) calculates that at such a rate the economy would use up all the solar power that reaches the earth in 400 years and in 1700 years all of the energy of the sun. The use of so much energy would generate tremendous waste heat independent of the AGI/ASI's smart energy technology. It would be so hot as to boil the surface of the Earth in about 400 years (Murphy, 2022b).

The upshot is that the acceleration in that growth that may be achieved by a Friendly AGI/ASI will ultimately be a transient event - if it ever happens.[43] To sustain it, galactic

[42]It is estimated that there are between 10^{78} and 10^{82} atoms in the observable universe, see `https://www.universetoday.com/36302/atoms-in-the-universe/`

[43]The possibility that an ever smarter narrow AI and eventually an AGI will lead to growth stagnation

expansion may be required (Wiley, 2011) - see section 5.2 below for a discussion. If somehow, the Singleton does not manage to achieve the level of technology required for galactic expansion (perhaps its utility functions and values were constrained when the AI-alignment problem was solved) by the time economic growth runs up against these physical constraints, then, as Murphy (2022b, p.847) warned, "we would be wise to plan for a post-growth world." It is, fortunately, still a long way in the future if we can invent a Friendly AI - in time to avoid the Empty Planet fate.

4.2 Existential Risks: The Wireheaded Orgasmium

In the previous subsection we discussed the possible economic growth impacts of an AGI/ASI, where these are relatively benign and perhaps ambiguous, but not catastrophic or existential, for humanity. Most of these scenarios involve some degree of "economic science fiction" in the words of (Nordhaus, 2021). Given that most mammal species go extinct after 1 million years or so, perhaps most people can live with a scenario where humans eventually become extinct in the far future after having lived fairly happy lives under a Singleton as a "sort of doddering Neanderthal aristocracy."

The two assumptions with which the analyses in section 4.1 were done, were that (i) that the AI Scaling Up hypothesis holds, and that (ii) the value-alignment problem are solved. In section 4.1.4 the first assumption was relaxed, and it was shown how super-exponential growth can arise via another form of AGI - Ems. In this section the second assumption - that the value-alignment problem is solved - will be relaxed.

Relaxing the assumption that the value-alignment problem is solved before the invention of

and even collapse has not been discussed here. Under particular assumptions regarding the closure of endogenous growth models, and in overlapping generations models, AGI pushes almost all human labour out of the economy, which leads to aggregate demand declining - who will buy all the goods produced? For a discussion of how AI can lead to such immiseration see Benzell et al. (2015), Gries and Naudé (2020) and Kotlikoff (2022).

an AGI raises the spectre of existential risk. There is a growing literature on whether and how AI poses an existential risk.[44] Economists have contributed little to this literature. It is one of the gaps in the economics of AI that this paper is emphasising. To explore why and how economics can contribute, it is useful to describe what is meant by existential risk, and why an AGI is considered a potential existential risk.

The term existential risk is associated with Longtermism, given that it was first used by Bostrom (2002) who defined it as a risk of an outcome that "would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential" (Bostrom, 2002, p.2). Existential risk from an AGI is taken seriously by a significant share of scientists, including scientists who do not necessarily subscribe to Longtermism. A recent headline exclaimed that "A third of scientists working on AI say it could cause global disaster" (Hsu, 2022).

In a survey of no less than two dozen ways in which AI poses an existential risk, Turchin and Denkenberger (2020, p.148) warns that "AI is an extremely powerful and completely unpredictable technology, millions of times more powerful than nuclear weapons. Its existence could create multiple individual global risks, most of which we can not currently imagine." And Noy and Uher (2022, p.498) concluded that "Artificial Intelligence (AI) systems most likely pose the highest global catastrophic and existential risk to humanity from the four risks we described here, including solar-fares and space weather, engineered and natural pandemics, and super-volcanic eruptions."

Why would AI pose catastrophic or even existential risks? It is due to the actual and future capabilities and values of AI. The capability claim is that AI may in future, even if the chance is small, cause significant damage to humanity. The value claim is the one

[44] Although there is, perhaps surprisingly, not that many people directly working on the problem: Hilton (2022) estimates around 300 people. This has lead Harris (2020, p.320) to exclaim that "There may be more people working in my local McDonald's than there are thinking about the prospect that we may permanently destroy the future."

that we have dealt with in section 2, namely that AI's values may not align with that of humanity (Sotala, 2018; Barrett and Baum, 2017). Given that it cannot be ruled out that a AGI or superintelligence will come into being with the non-trivial probability of causing the extinction of humanity, many - but not all - AI scientists now tend to conclude that "The consequences for humanity are so large that even if there is only a small chance[45] of it happening in that time frame, it is still urgent that we work now to understand it and to guide it in a positive direction"(Omohundro, 2008b, p.5). Everitt et al. (2018) also argues that in addition to try and reduce an extinction risk, it is philosophically stimulating to work on the challenge of constraining a superior intelligence.

What kind of risks does an AGI/ASI pose that may have existential implications for humanity? Turchin and Denkenberger (2020) list two dozen possibly "global catastrophic" risks from AI. They classify these into risks from narrow AI, young AI, and mature AI. They argue that it is not only mature AI - eventually AGI - that poses serious risks, but that AI along its entire development path poses such risks. Table 1 summarises these risks.

It falls outside the scope of this paper to discuss each of the risks listed in Table 1 - the reader is referred to Turchin and Denkenberger (2020) and the references listed in the table. The purpose of presenting these risks in one table is to illustrate succinctly the extent of thinking that has gone into elaborating the manifold ways in which AI may be a threat. Not all of the risks listed in Table 1 have equal probability - some, like the Roko's Basilisk,[46] is seen by most as being (even more) far-fetched. Nevertheless, a number of risks do stand out, and in light of the discussion in section 2 about AI's utility function and its supergoals and

[45] As put by Müller (2014, p.298) "The discussion of risk is not dependent on the view that AGI is on a successful path towards human-level AI – though it gains urgency if such 'success' is a non-negligible possibility in the coming decades. It also gains urgency if the stakes are set high, even up to human extinction. If the stakes are so high, even a fairly small possibility (say, 3%) is entirely sufficient to motivate the research.

[46] "Roko's Basilisk is an evil, godlike form of artificial intelligence, so dangerous that if you see it, or even think about it too hard, you will spend the rest of eternity screaming in its torture chamber. It's like the videotape in The Ring. Even death is no escape, for if you die, Roko's Basilisk will resurrect you and begin the torture again." It has been described as the most terrifying thought experiment of all time (Auerbach, 2014).

Table 1: Potential Catastrophic and Existential AI Risks

AI Level	Risk	References
Narrow AI		
	AI help create biotech weapons	O'Brien and Nelson (2020)
	AI-driven mass unemployment	Ford (2016)
	AI boost mass destruction weapons	Umbrello et al. (2020)
	Wrong command to robotic army	Goldfarb and Lindsay (2022)
	Slaughterbot swarms	Macaulay (2021)
	Self improving AI-ransomware	Yampolskiy (2016)
	Ascending non-human economy	Alexander (2016a)
	Super-addictive drugs	Urbina et al. (2022)
	AI viruses affect hardware globally	Turchin and Denkenberger (2020)
Young AI		
	Robots replace humans	Hanson (2018)
	Philosophical zombies	Searle (1992)
	Doomsday weapon for global blackmail	Shulman (2010)
	AI creates catastrophic event to escape	Yampolskiy (2012)
	AI uses human atoms as material	Yudkowsky (2009)
	AI kills humans for world domination	Turchin and Denkenberger (2020)
Mature AI		
	AI lock-in	Bostrom (2006)
	AI halting problem	Charlesworth (2014)
	Paperclip maximizer	Bostrom (2012)
	Smile maximizer	Yudkowsky (2008)
	Wireheading humans	Yampolskiy (2014)
	Roko's Basilisk	Auerbach (2014)
	Conflict between benevolent AIs	Turchin and Denkenberger (2020)
	Philosophical landmines	Torres (2014)
	Alien AI attack	Carrigan Jr (2006)
	Evil AI	Pistono and Yampolskiy (2016)

Source: Author's compilation based on Turchin and Denkenberger (2020).

instrumental goals, a number of remarks are in order.

The first is the risk of an AI lock-in (value lock-in). This is the risk that humanity may spend the rest of its existence[47] under an unchangeable, static set of rules that will follow

[47]Hanson (2022) while recognizing the possibility of lock-in suggest that it may not be forever, given that

from whatever goals and values are held onto by the AGI/ASI. A stable, immortal Singleton will be unassailable, and hence in a position to stifle all change (Bostrom, 2006).

The threat of a lock-in is not only limited to the case of an ASI Singleton, but can also happen in a the digital world of EMs - or the digital Metaverse (Karnofsky, 2021a). This is because any significantly advanced digital world can be programmed to be stable - hence no improvement or escape is it amounts to a totalitarian nightmare would be possible. A ruler for life and a set of rules could simply be constructed to be permanent(Karnofsky, 2021d).

A second risk to be highlight in light of the discussion in section 2 is the wireheading - of both humans and AI. Wireheading refers to the direct stimulation of reward centres in the brain to override any motivation to action. The term originates the implantation of wire electrodes in rats' brains through which the animals could provide self-pleasure by pulling on a lever. This resulted in "the rat's self-stimulation behaviour completely displaced all interest in sex, sleep, food and water, ultimately leading to premature death" (Yampolskiy, 2014, p.373). Essentially, the rat's utility function is not secure - it can be hacked to remove the need to pursue subgoals/ instrumental goals. This amounts to "counterfeit utility production" which is marked by "the absence of productive behaviour in order to obtain the reward. Participating individuals go directly for the reward and fail to benefit the society. In most cases, they actually cause significant harm via their actions" (Yampolskiy, 2014, p.374).

In Huxley (1932)'s novel *Brave New World*, humans are kept in pacified condition to accept their tech-ruled existence through using a drug called *Soma*. Tirole (2021) refers to this as the soft control of society - which is contrasted to most popular depictions of future totalitarian dystopias as being maintained by violent repression - the "boot-on-the-face". In a 1949 letter from Huxley to George Orwell, Huxley explained[48] that "*Whether in actual fact the policy of the boot-on-the-face can go on indefinitely seems doubtful. My own belief is that*

in a billion years it may be overcome by an alien intelligence - see the discussion of the Dark Forest in section 4.

[48]As quoted in Tirole (2021, p.2011).

the ruling oligarchy will find less arduous and wasteful ways of governing and of satisfying its lust for power, and these ways will resemble those which I described in Brave New World." The drug *Soma* in Huxley's novel is a form of wireheading - of hacking humans' utility functions - and as Huxley indicates, a very likely method through which an AGI/ASI could repress humanity.

It is not only humans that are subject to wireheading - AI systems may also be able to be hacked in this way. The consequence could be AGI/ASIs with no interest in performing any actions except to provide itself with counterfeit utility and maintain the status quo (Yampolskiy, 2014). Rather than being exterminated, or slowly dying off as a "sort of doddering Neanderthal aristocracy," humans may perhaps rather end up blissing it out in almost perpetuity "in wireheaded orgasmium" (Yampolskiy, 2014, p.374) with its Singleton overlord in similar ecstasy - until the end of time.

Or until it is exterminated or enslaved by an alien AI...

5 Extraterrestrial AI: The Final Existential Risk?

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

In Figure 1, AI's future development trajectory is shown to result in the Singleton by the year 2100 and some unknown time thereafter in a Galactic AI. One of the risks that it, and humanity - if we still exist at that time - will face, as Table 1 indicates, is a threat from an alien ASI civilization- an Alien Singleton, perhaps.

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[49] (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.

5.1 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, 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.[50] The fact that we have not yet, and that there is

[49]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).

[50]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).

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)

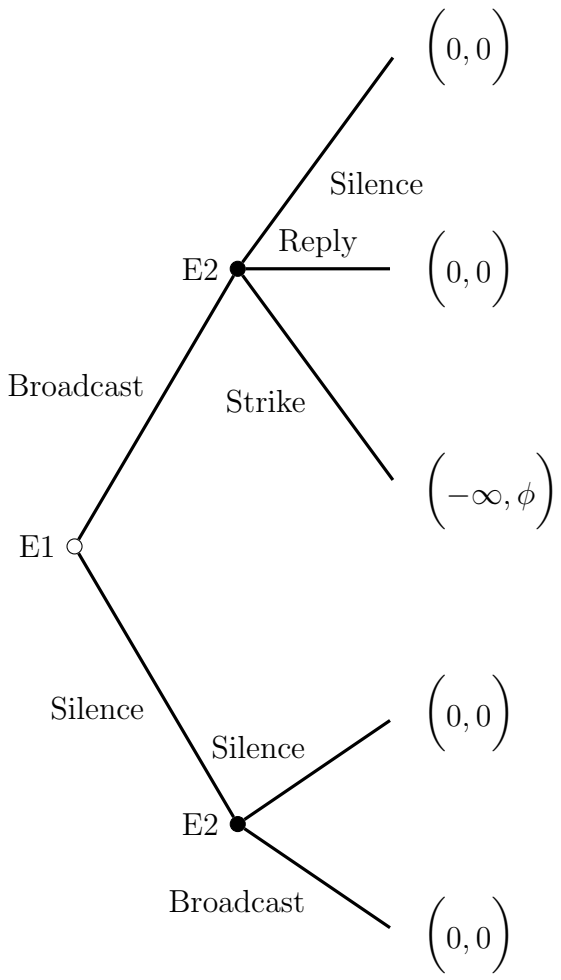
> "The universe is a dark forest. Every civilization is an armed hunter stalking through the trees like a ghost, gently pushing aside branches that block the path and trying to tread without sound. Even breathing is done with care. The hunter has to be careful, because everywhere in the forest are stealthy 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 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.

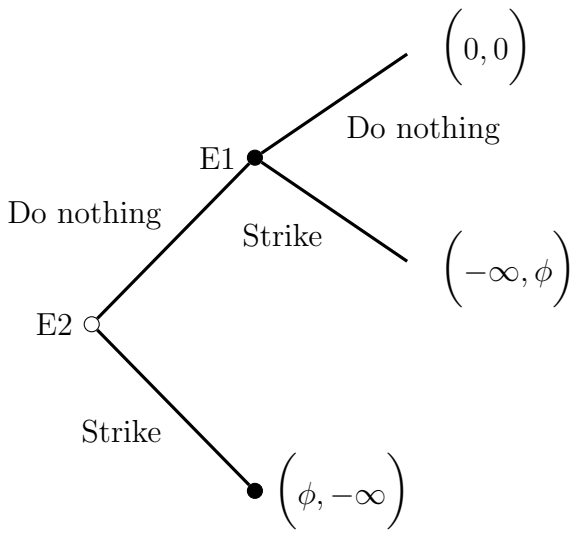
Game Theoretic analyses can be used to show that the DFT implies that, as in the case of the Prisoners' Dilemma which it closely resembles, the optimal strategy for any civilization is not to co-operate but to be silent and if discovered, to be malevolent - in other words to strike first (Stolk, 2019; 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."

5.2 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 (see section 4.1.5) that civilizations will want to expand beyond their planet. Natural selection may favour civilizations that expand (Wiley, 2011).

Bostrom (2003a) 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.[51] 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).

[51] 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[52] 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[53] 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.[54] Failure 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

[52]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).

[53]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.

[54]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*.

5.2.1 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."

5.2.2 Grabby Aliens

The Grabby Aliens Model suggests that we have not yet encountered ETIs because our Earth civilization has risen early[55] in the galaxy - if we had not, we would never have had the opportunity to emerge, as our solar system wold have long ago been colonized by ETIs (Hanson et al., 2021). And because we are early, we will in the future encounter the aggressive alien expansion (as in Olson (2015)) - these are called "Grabby Aliens" - or Loud Aliens - to contrast them with quiet aliens, who may, Dark Forest-like, prefer not to engage in Galactic expansion. According to Hanson (2020) we should encounter a Grabby Alien civilization in around 500 million years.

5.2.3 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.

[55] As put by Hanson et al. (2021, p.2) "humanity seems to have appeared implausibly early in the history of the universe."

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 was reported in section 4.2, the existential risk of AI is taken seriously by many scientists. Ord (2020) has estimated that there is a one in ten probability that AI will cause human extinction in the next hundred years.

5.3 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."

6 Discussion

This purpose of this paper is to contribute to the economics of AI by exploring three topics neglected by economists: (i) the Singularity (and Singleton) as a result of a possible intelligence explosion, (ii) the existential risks that AI may pose to humanity; and (iii) the relevance of the mythical rational agent (homo economicus) for the design of value-aligned AI-systems.

The idea of a Singularity - an intelligence explosion caused by recursively self-improving AI - is most relevant for economists working in endogenous growth theory. This paper has outlined the relevance of idea-driven growth models, which highlight the importance of population growth and research productivity for exponential and super-exponential growth. It is also a relevant field for economists in technology studies, because if there would ever be a Singularity it would be followed by a wholly different mode of economic growth: as different from the current industrial era growth as the industrial era growth mode was different from that of the agricultural era.

While economists' models have been useful in explaining the take-off from low and stagnating

growth as had characterised most of human history, and in identifying the possibility for accelerating growth - the so-called growth explosions as discussed - there are two comparative blind spots. One is that very few economic growth models have entertained the possibility that AI-induced growth could lead to a growth collapse. A second is that economists are still shying away from the very long-run implications. A very-long-run implication is the limits of growth. Economists tend not to think too much about the physical limits to growth - this has been left rather to physicists to explore. Physicist Tom Murphy recounts this dinner conversation with an economist in his blog (Murphy, 2012):

> "Physicist: Hi, I'm Tom. I'm a physicist.
>
> Economist: Hi Tom, I'm [ahem..cough]. I'm an economist.
>
> Physicist: Hey, that's great. I've been thinking a bit about growth and want to run an idea by you. I claim that economic growth cannot continue indefinitely.
>
> Economist: [chokes on bread crumb] Did I hear you right? Did you say that growth can not continue forever?
>
> Physicist: That's right. I think physical limits assert themselves."

This absence of economics in the debate may be unfortunate - unlike physicists, economists tend to see very few resources as natural or fixed, but part of technologies and ideas - and thus subject to substitution and the impact of incentives. But in the long-run, physical limits do exist, and although physicists recognise these, the discussion about the ethical, moral and societal implications of these longer-term implications limits tend to be dominated by philosophers, in particular philosophers subscribing to Longtermism. This latter movement, while in its weak form has made the justifiable point about considering the well-being of future generations more than is generally the case, has become associated with the views and ambitious of some tech elite billionaires - among others Elon Musk, Peter Thiel, Jaan Tallin. The concern is that it is "a disturbing secular religion that looks like it addresses

humanity's deepest problems, but actually justifies pursuing the social preferences of elites" (Torres, 2021).

These social preferences may have nothing to do with addressing current global challenges such as poverty, inequality, conflict, migration and even climate change. In a now infamous paper on "strong" Longtermism, Greaves and MacAskill (2021) concluded that given the moral importance of ensuring future humans' well-being and existence, that "for the purposes of evaluating actions, we can in the first instance often simply ignore all the effects contained in the first 100 (or even 1,000) years, focusing primarily on the further-future effects. Short-run effects act as little more than tie-breakers." They later deleted this sentence (Torres, 2021). As this paper argued in section 3.4, drawing the analysis of the future potential of AI to its full conclusion suggest that the fundamental agenda of Longtermism is indeed not so much with current global problems, but neither ultimately with the well-being of future intelligent beings, but with a transhumanist agenda that with the aim to ensure getting "what the future owes us," i.e. to ensure that an AGI/ASI will be created that will in the very far future resurrect us all.

Economists can potentially make a much-needed contribution to the debate on the long-term implications of growth, the possible consequences of a future growth collapse or explosion, and the extent of the concern that should be placed on existential risk mitigation - and concern is certainly warranted. Economists would tend to counter suggestions from strong (or galaxy-brain) Longtermism and start by considering improving the well-being of future generations by beginning to solve current development challenges and problems. As put by Wright (2022) "I have a question for longtermists: Are you sure that our failure to think long term is the problem? [...] Here's my radical thought: The biggest existential threat we face [...] is that humans aren't good enough at shorttermism. If people were skilled shorttermists - if they pursued short-term interests wisely - our long-term problems, including the existential ones, would be manageable." Indeed, already back in the 1990s Baranzini and Bourguignon

(1995) linked existential risk and societal risk aversion to levels of development.

Generally, economists have not made many contributions to the field of existential risk studies (ERS).[56] Here, the existential risk posed by a super-AI is very much relevant. There is also the risk, not only from a super-AI, but from the advanced AI supported digital technologies that will characterise the future Metaverse. The risk is that these technologies will lock-in, essentially forever, future digital people, "ems," into a dystopian, totalitarian and autocratic world. This paper also argued that if the Dark Forest Theory is valid, an extraterrestrial AGI may pose a further existential risk. This tends to be a topic neglected by both philosophers and economists.

Why did economists, so far, contribute relatively little to the understanding and mitigation of the catastrophic and existential risks posed by AI? One reason is, as Noy and Uher (2022, p.294) points out, that economic risk assessment methods using EUT are not well-suited to deal with existential risks. Weitzman (2009, p.10) put forward a Dismal Theorem which states that, because the probabilities of catastrophic events are characterised by long tails, EUT would tend so assign infinite losses to it. This is because "distributions with fat tails are ones for which the probabilities of rare events decline relatively slowly as the event moves far away from its central tendency" (Nordhaus, 2009, p.3). Buchholz and Schymura (2012, p.234) concludes that because of this, EUT may not be able to provide an ethically acceptable approach to deal with catastrophic and existential risks. Using various specifications of utility functions with different assumptions on risk aversion, they show that if a fairly plausible level of risk adverseness is assumed, that it results in a "tyranny of catastrophic risk" as in Nordhaus (2009). On the other hand, with low risk aversion, EUT assessment of catastrophic risk would assign it no importance.

However, as Martin and Pindyck (2015) has shown, despite these shortcomings of EUT, economists' tools can provide useful insights into dealing with catastrophic risks. Their

[56]For an overview of the history and trends in ESR, see Beard and Torres (2020).

analysis offers insight into how to answer questions such as how much should society invest in averting the existential risk from AI, which as has been shown, can under Pascalian fanaticism be extremely high? And, what would be the effect of other existential risks on the amount that should be invested in averting an AI catastrophe? They show for instance that there is an interdependence between efforts to address a whole range of existential and catastrophic risks and that therefore"applying cost-benefit analysis to each event in isolation can lead to a policy that is far from optimal" (Martin and Pindyck, 2015, p.2948).

An exception in the economic literature which illustrates the additional light that economic models can bring to the analysis of economic growth, AGI and existential risk, is the paper by Baranzini and Bourguignon (1995). Their work, which largely predates modern AI and the concerns that it entails an existential risk, tried to answer the question of what society should do if it faces the decision "whether to adopt or not a new technology that will raise the rate of GDP growth by some variable amount?" They considered explicitly the possibility that such a new technology would increase the probability of humanity's extinction, which in turn implies that to minimize the risk of extinction, that the rate of innovation and economic growth should be reduced.

They show, using a growth model with a Hyperbolic Absolute Risk Aversion (HARA)[57] social utility function specification, that sustainable, low technology growth (i.e., with an extinction probability of zero) will be optimal only if the utility of survival and risk of technological innovation are relatively large. This leads to the conclusion that "sustainable growth is consistent with optimal growth only for affluent societies" (Baranzini and Bourguignon, 1995, p.353). Their model leads to the prediction that concern about a Singularity, existential risks and longtermism, may be rich-country and rich tech-elites' concerns. Which, as has been pointed out, it largely is. Which is why Floridi (2022, p.9) is critical of time spent speculating about the Singularity, stating that 'it is a rich-world preoccupation likely to worry people in

[57]Proposed by Von Neumann and Morgenstern (1944) wherein risk tolerance of an agent is a function of their wealth.

wealthy societies who seem to forget the real evils oppressing humanity and our planet."

Whereas physicists have recognised potential physical constraints on growth, and philosophers have raised the possible dangers and moral obligations of existential risks and future generations - the long-run - as far as the shorter-run is concerned computer scientists and engineers have tended to be more optimistic that technological unemployment and huge inequalities, although inevitable, do not really pose a problem. Domingos (2015, p.278,279) for instance is hopeful that the political processes and redistributive policies and social safety nets are the answer, arguing that although the societal transition as a result of the consequences of AI will in future be "tumultuous," "thanks to democracy it will have a happy ending" and that "unemployment benefits will be replaced by a basic income for everyone." Thus, for Domingos (2015), Chace (2020b) and others, the future of AI in the economy is one of high unemployment, high inequality but with strong welfare states where people will enjoy only leisure and AI and robots do all the work.

The problem with such a hopeful scenarios is a blind-spot for the effects of economic incentives. Not only may the adoption of AI not be as profitable as to lead to mass unemployment, but the robustness and effectiveness of democracy (political institutions) to adequately implement social safety nets (a basic income) may be overestimated (Tirole, 2021). Governments may simply not be able to regulate the powerful global tech oligarchy that is being created on the basis of AI and big data; moreover, they may be captured and co-opted by governments into legitimizing their monopoly (Srnicek, 2016). Surveillance capitalism and the growth of the surveillance state, already emerging outcomes of these processes, have triggered growing global discontent with capitalism (Zuboff, 2015). By reducing the value of labor, AI automation will moreover exacerbate global inequality, de-globalization and conflict, as the wealthiest nations will become "those with the highest ratio of natural resources to population" (Domingos, 2015, p.279). Economists do of course already focus on all of these issues (it all broadly touches on the Future of Work) and hence we have seen in recent

years that the concerns about a possible robot apocalypse (high technological unemployment) have somewhat abated. Perhaps one will see a similar pattern to the concerns about the Singularity and existential risks of AI as economists make their contribution.

Finally, this paper has argued that in addition to contributing to the topics of the AI Singularity and AI existential risks, that there are several areas where better understanding of AI methods, and collaboration with AI scientists may help in the further development of economic theory - and vice versa. Expected Utility Theory (EUT) - a foundation of economics - is the basis of decision-making by AI agents. In section 2.1 some examples were given of how insights from how economics deals with utility functions have been helping to improve AI systems - such as the use of Neural Utility Functions in Recommender Systems and bounded rationality in image-classification convolutional neural networks (CNNs).

Moreover, it was argued that as long as AI systems inhabit the small world of neoclassical Bayesian utility maximizing agents, they are restricted, and pose little threat of misalignment or hard take-off that will result in a Singularity or existential risk. Narrow AI builds very smart, focused agents in very simple situations. This minimizes AI risk.

An important area in decision-making where economic theory may benefit from closer interaction with AI scientists is on the topic of procedural rationality. Economic theory tends to ignore the reasoning process by which agents make these rational decisions, i.e., how agents find the optimum of their expected utility functions (Dixon, 2001). Economics have preferred *substantive rationality* over *procedural rationality* Simon (1978). It amounts to an approach where "what decisions are made is more important than how they are made" (Harré, 2021, p.12).

As Dixon (2001) has suggested, there is a need in economics to consider the process of reasoning itself, because human decision-making is prone to mistakes - even beyond those

due to being bounded by computational ability. He concludes[58] that "the role for artificial intelligence in economics would then seem primarily to be in situations where economic agents make mistakes, and possibly bad mistakes." More generally, AI research, such as in the subfield of RL, may also help economists overcome computational difficulties in understanding reasoning under bounded rationality (Charpentier et al., 2021) and help to model human behaviour in disequilibrium situations (Dixon, 2001). One may also interpret non-rational quirks in human judgment as the result of AI learning techniques (e.g. DL or RL) that are "inappropriately applied" (Camerer, 2019).

In sum, AI may thus help economics to more accurately model human decision-making (Engler, 2020). AI may require economic theory and models to be adjusted, improved, or even completely re-thought. Maybe AI will make the invisible hand visible.

7 Concluding Remarks

Upon starting to write this paper, we approached one of the most advanced AI systems in the world (at the time of writing) which is GTP-3 (Generative Pre-trained Transformer 3). GPT-3 is a Deep Learning based generative (language) model developed by *OpenAI*.[59] We asked GTP-3 "How can artificial intelligence improve economic theory?" It answered that

> "Artificial intelligence can improve economic theory by helping to identify patterns and relationships that would otherwise be difficult to discern. For example, machine learning could be used to analyze large data sets to identify trends in consumer behavior. This information could then be used to improve economic models and predictions. Additionally, artificial intelligence can help to automate

[58] Dixon (2001) that AI research, by highlighting the mechanisms of reasoning, may also throw light on strategic behaviour, where economic agents may face incentives to intentionally make mistakes.

[59] It can be accessed at `https://beta.openai.com/overview`

the process of data collection and analysis, making it more efficient and accurate."

Excellent though the answer is, it is very much based in current notions of AI - of AI as narrow AI, which indeed offers impressive abilities in the areas of pattern recognition and prediction using big data. The discussion in this paper has emphasised however, with reference to speculative considerations of a Singularity and/or an AGI that can cause human extinction, that relation between the fields of AI and economics depends very much on what the future holds - whether AI will progress past narrow AI, or whether coming decades may see more attention on controlling and regulating data -and algorithm-based AI - improving its safety - rather than on impactful new tools and breakthroughs.

As such the overarching purpose of this paper was to contribute to the economics of AI, going beyond the dominant concern of economists on AI's labour market implications, or on the potential usage of AI methods to improve prediction and hypothesis testing. Rather, this paper interrogated the literature on what the future may hold for AI, drew implications from and for economics from these - even to extent of indulging in some economic science fiction. Three topics that are relevant in this regard, and which are neglected by economists and that were explored in detail in this paper were (i) the Singularity (and Singleton) as a result of a possible intelligence explosion, (ii) the existential risks that AI may pose to humanity; and (iii) the relevance of the mythical rational agent (homo economicus) for the design of value-aligned AI-systems.

By exploring these topics, the costs and benefits of AI were de-hyped; several future avenues for economic research on AI became apparent; and areas where economic theory may benefit from a greater understanding of AI, were identified.

Regarding the costs and benefits of AI, the Expected Utility Theory (EUT) perspective used in this paper suggested that as long as AI systems inhabit the small world of neoclassical Bayesian utility maximizing agents, they are restricted and pose little threat of misalignment

or hard take-off that will result in a Singularity or existential risk. Narrow AI builds very smart, focused agents in very simple situations. This minimizes AI existential risk (although challenges for AI safety remain). This finding is consistent with the earlier conclusion of Naudé (2021) to expect *neither Utopian nor apocalyptic impacts from AI soon.*

Future avenues for research that were identified included the need for further elaborations of economic growth models to explore the possibility of an AI-induced growth collapse, to explore the physical limits of growth, and to sharpen the tools to draw out the policy implications of facing fat-tailed catastrophic risks. Furthermore, economic perspectives may usefully be applied to the solutions and implications of the Fermi Paradox. These include applying economic tools to potential far-future challenges, such as decisions on whether and when - and how - to colonise the galaxy; whether or not to try and contact Extraterrestrial Intelligences (ETIs); whether or not to choose conflict or attempt cooperation with other ETIs; how to best protect a planetary civilization; and when an Earth-based civilization could expect to find evidence of an ETI.

This paper also found that economic theory may benefit from the field of AI as far as procedural utility is concerned, to overcome computational difficulties in understanding reasoning under bounded rationality, and help economists to model human behaviour better in disequilibrium situations.

Finally, two further conclusions that emerged from this paper were first that a Singularity and existential risk from AI are still science fiction: which, however, should not preclude economists from weighing in - it certainly does not deter philosophers; and two, that economists should contribute more to existential risk studies, and not leave this topic to lose credibility because of the Pascalian fanaticism of the recent fad of Longtermism.

In conclusion. The future is essentially unknowable, which makes it worthwhile to reflect appropriately on the speculations that this paper has indulged in, including discussions of

existential risks, limits to growth, and bounded rationality. A useful guardrail to grab hold of when contemplating these, is the sensible caution of Deutsch (2011, p.198) that "Trying to know the unknowable leads inexorably to error and self-deception. Among other things, it creates a bias towards pessimism."

References

Abbeel, P. and Ng, A. (2004). Apprenticeship Learning via Inverse Reinforcement Learning. *Proceedings of the Twenty-First International Conference on Machine Learning, ICML 04. New York, NY, USA.*

Agrawal, A., McHale, J., and Oettl, A. (2018). Finding Needles in Haystacks: Artificial Intelligence and Recombinant Growth. *NBER Working Paper no. 24541. National Bureau for Economic Research.*

Agrawal, A., McHale, J., and Oettl, A. (2019). Artificial Intelligence, Scientific Discovery, and Commercial Innovation. *Mimeo.*

Agrawal, S. (2019). Reinforcement Learning Lecture 1: Introduction. *University of Minnesota.*

Ahiska, S., Appaji, S., King, R., and Warsing Jr, D. (2013). A Markov Decision Process-Based Policy Characterization Approach for a Stochastic Inventory Control Problem with Unreliable Sourcing. *International Journal of Production Economics*, 144(2):485–495.

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

Aleksander, S. (2019). 1960: The Year the Singularity was Cancelled. *Slate Star Codex Blog*, 22 April.

Alexander, S. (2016a). Ascended Economy? *Star Slate Codex Blog*, May 30th.

Alexander, S. (2016b). Book Review: Age of EM. *Slate Star Codex Blog*, 28th May.

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

Allen, C., Smit, I., and Wallach, D. (2005). Artifcial Morality: Top-Down, Bottom-Up, and Hybrid Approaches. *Ethics and Information Technology*, 7(3):149–155.

Ariely, D. (2009). Predictably Irrational: The Hidden Forces That Shape Our Decisions. *London: HarperCollins.*

Armstrong, S., Bostrom, N., and Schulman, C. (2016). Racing to the Precipice: A Model of Artificial Intelligence Development. *AI & Society*, 31:201–206.

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.

Armstrong, S., Sandberg, A., and Bostrom, N. (2012). Thinking Inside the Box: Controlling and Using an Oracle AI. *Minds and Machines*, 22(4):299–324.

Armstrong, S. and Sotala, K. (2015). How We're Predicting AI - or Failing to. *In Romportl, J., Zackova, E., Kelemen, J. (eds.) Beyond Artificial Intelligence. Topics in Intelligent Engineering and Informatics, vol 9. Springer, Cham.*

Arrow, K. (1994). Methodological Individualism and Social Knowledge. *American Economic Review, Papers and Proceedings*, 84(2):1–9.

Arthur, W. (2021). Foundations of Complexity Economics. *Nature Reviews Physics*, 3:136–145.

Arulkumaran, K., Deisenroth, M., Brundage, M., and Bharath, A. (2017). A Brief Survey of Deep Reinforcement Learning. *arXiv:1708.05866v2 [cs.LG].*

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

Athey, S. and Imbens, G. (2019). Machine Learning Methods that Economists Should Know About. *Annual Review of Economics*, 11(1):685–725.

Auerbach, D. (2014). The Most Terrifying Thought Experiment of All Time. *Slate Magazine*, 17 July.

Balland, P., T.Broekel, Diodato, D., Giuliani, E., Hausmann, R., O'Clery, N., and Rigby, D. (2022). The New Paradigm of Economic Complexity. *Research Policy*, 51(3).

Baranzini, A. and Bourguignon, F. (1995). Is Sustainable Growth Optimal? *International Tax and Public Finance*, 2:341–356.

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.

Beard, S. and Torres, E. (2020). Ripples on the Great Sea of Life: A Brief History of Existential Risk Studies. *Mimeo: SSRN*, 12 March.

Beckstead, N. (2013). On the Overwhelming Importance of Shaping the Far Future. *DPhil Thesis, Rutgers University*.

Bellman, R. (1957a). Dynamic Programming. *Princeton: Princeton University Press*.

Bellman, R. (1957b). A Markovian Decision Process. *Journal of Mathematics and Mechanics*, 6(5):679–684.

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.

Benya (2012). Why You Must Maximise Expected Utility. *AI Alignment Forum*, 13 Dec.

Benzell, S., Kotlikoff, L., LaGarda, G., and Sachs, J. (2015). Robots Are Us: Some Economics of Human Replacement. *NBER Working Paper no. 20941*.

Bernheim, B. and Whinston, M. (1986). Common Agency. *Econometrica*, 54(4):923–942.

Bernoulli, D. (1738). Commentarii. *Acad. Scientiarum Imperialis Petropolitanae 5 175-192; English translation (1954) Econometrica*, 22:23–36.

Bickley, S., Chan, H., and Torgler, B. (2022). Artificial Intelligence in the Field of Economics. *Scientometrics*, 127:2055–2084.

Binmore, K. (2007). Rational Decisions in Large Worlds. *Annales d'Économie et de Statistique*, 86:25–41.

Binmore, K. (2008). Rational Decisions. *Princeton University Press: Princeton, NJ*.

Binmore, K. (2017). On the Foundations of Decision Theory. *Homo Oeconomicus*, 34:259–273.

Bishop, J. (2021). Artificial Intelligence Is Stupid and Causal Reasoning Will Not Fix It. *Frontiers of Psychology*, 11:513474.

Bloom, N., Bunn, P., Chen, S., Mizen, P., and Smietanka, P. (2020). The Economic Impact of Coronavirus on UK Businesses: Early Evidence from the Decision Maker Panel. *VOX CEPR Policy Portal*, 27th March.

Bostrom, N. (1998). How Long Before Superintelligence? *International Journal of Futures Studies*, 2.

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

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

Bostrom, N. (2003b). Ethical Issues in Advanced Artificial Intelligence. *In I. Smit et al. (eds.) Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence. 2nd ed. International Institute of Advanced Studies in Systems Research and Cybernetics*, pages 12–17.

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. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, 22(2):71–85.

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.*

Bricker, D. and Ibbitson, J. (2020). Empty Planet: The Shock of Global Population Decline. *Little, Brown Book Group.*

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

Brummitt, C., Gomez-Liévano, A., Hausmann, R., and Bonds, M. (2020). Machine-Learned Patterns Suggest that Diversification Drives Economic Development. *Journal of the Royal Society Interface*, 17:20190283.

Brynjolfsson, E., Rock, D., and Syverson, C. (2017). Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics. *NBER Working Paper no. 24001. National Bureau for Economic Research.*

Buchholz, W. and Schymura, M. (2012). Expected Utility Theory and the Tyranny of Catastrophic Risks. *Ecological Economics*, 77:234–239.

Camerer, C. (2019). Artificial Intelligence and Behavioral Economics. *In A. Agrawal and J.Gans and A. Goldfarb (eds.) The Economics of Artificial Intelligence: An Agenda. Chicago: University of Chicago Press*, pages 587–610.

Caplin, A., Martin, D., and Marx, P. (2022). Modeling Machine Learning. *NBER Working Paper No. 30600.*

Carrigan Jr, R. (2006). Do Potential SETI Signals Need to be Decontaminated? *Acta Astronaut*, 587:112–117.

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.

Cervantes, J., Lopez, L., Rodriguez, L., Cervantes, S., Cervantes, F., and Ramos, F. (2020). Artificial Moral Agents: A Survey of the Current Status. *Science and Engineering Ethics*, 26:501–532.

Chace, C. (2020a). Artificial Intelligence & Fully Automated Luxury Capitalism. *Forbes*, 15 July.

Chace, C. (2020b). The Economic Singularity. 3rd edition. *Three C's.*

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

Charlesworth, A. (2014). The Comprehensibility Theorem and the Foundations of Artificial Intelligence. *Minds & Machines*, 24:439–476.

Charpentier, A., Élie, E., and Remlinger, C. . (2021). Reinforcement Learning in Economics and Finance. *Computational Economics.*

Chen, K., Raghupathi, S., Chandratreya, I., Du, Q., and Lipson, H. (2022). Automated Discovery of Fundamental Variables Hidden in Experimental Data. *Nature Computational Science*, 2:433–442.

Christian, B. (2020). The Alignment Problem: Machine Learning and Human Values. *New York: W.W. Norton.*

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.

Clancy, M. (2021). Combinatorial Innovation and Technological Progress in the Very Long Run. *New Things Under the Sun*, June 18.

Cockburn, I. M., Henderson, R., and Stern, S. (2019). The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis. *In A. Agrawal and J. Gans and A. Goldfarb (eds.) The Economics of Artificial Intelligence: An Agenda. Chicago: University of Chicago Press*, pages 115–148.

Cohen, M., Hutter, M., and Osborne, M. (2022). Advanced Artificial Agents Intervene in the Provision of Reward. *AI Magazine*, 43(3):282–293.

Colbrook, M., Antun, V., and Hansen, A. (2022). The Difficulty of Computing Stable and Accurate Neural Networks: On the Barriers of Deep Learning and Smales 18th Problem. *PNAS*, 119(12):e2107151119.

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.

Cowen, T. (2010). The Great Stagnation. *New York: Penguin (Dutton).*

Crawford, J. (2022). Can Economic Growth Continue Over the Long-Term? *Longnow*, 7 October.

Dai, W. (2019). AGI will Drastically Increase Economies of Scale. *AI Alignment Forum*, 8th June.

Daneke, G. A. (2020). Machina-Economicus or Homo-Complexicus: Artificial Intelligence and the Future of Economics? *Real-World Economics Review*, 93:18–39.

Dastani, M., Hulstijn, J., and van der Torre, L. (2005). How to Decide What to Do? *European Journal of Operational Research*, 160:762–784.

Davidson, T. (2021). Report on Whether AI Could Drive Explosive Economic Growth. *Open Philantrophy*, 17 June.

Davies, J. (2017). Hidden AI - Ghosts In The Machine. *Becoming Human: Artificial Intelligence Magazine*, Sept 6.

Dawkins, R. (1976). The Selfish Gene. *Oxford: Oxford University Press.*

De Luna, P., Wei, J., Bengio, Y., Aspuru-Guzik, A., and Sargent, E. (2017). Use Machine Learning to Find Energy Materials. *Nature*, 552.:23–25.

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.

DeLong, J. B. (1998). Estimates of World GDP, One Million B.C. - Present. *Department of Economics, U.C. Berkeley at http://econ161.berkeley.edu/.*

Dennis, L. (2020). Computational Goals, Values and Decision-Making. *Science and Engineering Ethics*, 265:2487–2495.

Deutsch, D. (2011). The Beginning of Infinity: Explanations that Transforms the World. *London: Penguin Books.*

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

Ding, Y., Florensa, C., Phielipp, M., and Abbeel, P. (2019). Goal-Conditioned Imitation Learning. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.*

Dixon, H. (2001). Surfing Economics. *Red Globe Press London.*

Domingos, P. (2015). The Master Algorithm: How the Quest for the Ultimate Learning Machine will Remake our World. *London: Penguin Books.*

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].*

Dütting, P., Feng, Z., Narasimhan, H., Parkes, D., and Ravindranath, S. (2019). Optimal Auctions Through Deep Learning. *Proceedings of the 36th International Conference On Machine Learning*, pages 1706–1715.

Eckersley, P. (2019). Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function). *SafeAI 2019: Proceedings of the AAAI Workshop on Artificial Intelligence Safety 2019*.

Ehrlich, P. (1968). The Population Bomb. *New York: Ballantine Books*.

Ellsberg, D. (1961). Risk, Ambiguity and the Savage Axioms. *Quarterly Journal of Economics*, 75:643–699.

Emba, C. (2022). Why Llongtermism isn't Ethically Sound. *Washington Post*, 5 September.

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

Engler, A. (2020). A Guide to Healthy Skepticism of Artificial Intelligence and Coronavirus. *The Brookings Institution*, 2 April.

Everitt, T. and Hutter, M. (2016). Avoiding Wireheading with Value Reinforcement Learning. *arXiv:1605.03143v1 [cs.AI]*.

Everitt, T., Lea, G., and Hutter, M. (2018). AGI Safety Literature Review. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, pages 5441–5449.

Fernandez-Villaverde, J. (2020). Simple Rules for a Complex World with Artificial Intelligence. *PIER Working Paper No. 20-010*.

Floridi, L. (2022). Ultraintelligent Machines, Singularity, and Other Sci-fi Distractions about AI. *Mimeo: SSRN*.

Ford, M. (2016). The Rise of the Robots: Technology and the Threat of Mass Unemployment. *London: Oneworld Publications*.

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

Gabriel, I. and Ghazavi, V. (2021). The Challenge of Value Alignment: From Fairer Algorithms to AI Safety. *arXiv:2101.06060 [cs.CY]*.

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.

Gallier, J. and Quaintance, J. (2022). Algebra, Topology, Differential calculus, and Optimization Theory For Computer Science and Machine Learning. *Mimeo: University of Pennsylvania*.

Galor, O. and Weil, D. N. (2000). Population, Technology, and Growth: From Malthusian Stagnation to the Demographic Transition and Beyond. *American Economic Review*, 90(4):806–828.

Gans, J. (2018). AI and the Paperclip Problem. *VoxEU Column*, 10 June.

García-García, J., García-Ródenas, R., López-Gómez, J., and Martín-Baos, J. (2022). A Comparative Study of Machine Learning, Deep Neural Networks and Random Utility Maximization Models for Travel Mode Choice Modelling. *Transportation Research Procedia*, 62:374–382.

Gauchon, R. and Barigou, K. (2021). Expected Utility Maximization with Stochastically Ordered Returns. *Mimeo: University of Lyon*.

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.

Goldfarb, A. and Lindsay, J. (2022). Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War. *International Security*, 46(3):7–50.

Gonzales, C. and Perny, P. (2020). Decision Under Uncertainty. *In P. Marquis and O. Papini and H. Prade (eds.). A Guided Tour of Artificial Intelligence Research, I, Springer*, pages 549–586.

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, pages 2672–2680.

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

Greaves, H. and MacAskill, W. (2021). The Case for Strong Longtermism. *GPI Working Paper No. 5-2021 Global Priorities Institute, University of Oxford*.

Gries, T. and Naudé, W. (2020). Artificial Intelligence, Income Distribution and Economic Growth. *IZA Discussion Paper no. 13606*.

Grossman, S. and Hart, O. (1983). An Analysis of the Principal-Agent Problem. *Econometrica*, 51:7–45.

Guo, I., Langrené, N., Loeper, G., and Ning, W. (2020). Robust Utility Maximization Under Model Uncertainty via a Penalization Approach. *arXiv:1907.13345v5 [math.OC]*.

Gurzadyan, V. and Penrose, R. (2016). CCC and the Fermi Paradox. *The European Physics Journal Plus*, 131:11.

Hadfield-Menell, D., Dragan, A., Abbeel, P., and S.Russell (2016). Cooperative Inverse Reinforcement Learning. *arXiv:1606.03137 [cs.AI]*.

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

Hanson, R. (2000). Long-Term Growth as a Sequence of Exponential Modes. *Mimeo: George Mason University*.

Hanson, R. (2009). Limits to Growth. *Overcoming Bias Blog*.

Hanson, R. (2014). I Still Don't Get Foom. *Overcoming Bias Blog*, 24 July.

Hanson, R. (2018). The Age of EM: Work, Love, and Life when Robots Rule the Earth. *Oxford: Oxford University Press*.

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

Hanson, R. (2022). Macaskill on Value Lock-In. *Overcoming Bias Blog*, 16 August.

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*.

Harré, M. (2021). Information Theory for Agents in Artificial Intelligence, Psychology, and Economics. *Entropy*, 23:310.

Harris, S. (2015). Can We Avoid a Digital Apocalypse? *SAM Harris Blog, 16 January, at https://samharris.org/can-we-avoid-a-digital-apocalypse/.*

Harris, S. (2020). Making Sense: Conversations on Consciousness, Morality and the Future of Humanity. *London: Penguin.*

Harsanyi, J. (1978). Bayesian Decision Theory and Utilitarian Ethics. *American Economic Review*, 68(2):223–228.

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.

Heaven, D. (2019). Deep Trouble forDeep Learning. *Nature*, pages 163–166.

Herfeld, C. (2020). The Diversity of Rational Choice Theory: A Review Note. *Topoi*, 39:329–347.

Hibbard, B. (2012). Model-Based Utility Functions. *Journal of Artificial General Intelligence*, 3(1):1–24.

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

Hilton, B. (2022). Preventing an AI-Related Catastrophe: AI Might Bring Huge Benefits - If We Avoid the Risks. *8,000 Hours*, 25 August.

Hinton, G. and Salakhutdinov, R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, 313:504–507.

Hinton, G., S.Osindero, and The, T.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18:1527–1554.

Howard, R. (1960). Dynamic Programming and Markov Processes. *Cambridge MA: The MIT Press.*

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.

Hubinger, E. (2020). An Overview of 11 Proposals for Building Safe Advanced AI. *arXiv:2012.07532 [cs.LG]*.

Huebner, J. (2005). A Possible Declining Trend for Worldwide Innovation. *Technological Forecasting & Social Change*, 72:980–986.

Hutter, M. (2000). A Theory of Universal Artificial Intelligence Based on Algorithmic Complexity. *arXiv:cs/0004001 [cs.AI]*.

Hutter, M. (2007). Universal Algorithmic Intelligence: A Mathematical Top-Down Approach. *In B. Goertzel and C. Pennachin (eds). Artificial General Intelligence. Cognitive Technologies. Springer, Berlin, Heidelberg*, pages 227–290.

Huxley, A. (1932). Brave New World. *London: Chatto & Windus*.

Jenkins, P., Farag, A., Jenkins, J., Yao, H., Wang, S., and Li, Z. (2021). Neural Utility Functions. *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 35(9):7917–7925.

Johansen, A. and Sornette, D. (2001). Finite-time singularity in the dynamics of the world population: Economic and financial indices. *Physica A*, 294(3-4):465–502.

Jones, B. (2009). The Burden of Knowledge and the Death of Renaissance Man: Is Innovation Getting Harder? *Review of Economic Studies*, 76(1):283–317.

Jones, C. (1995). R&D - based models of Economic Growth. *Journal of Political Economy*, 103(4):759–783.

Jones, C. (2001). Was an Industrial Revolution Inevitable? Economic Growth Over the Very Long Run. *Advances in Macroeconomics*, 1(2):article 1.

Jones, C. (2022). The End of Economic Growth? Unintended Consequences of a Declining Population. *American Economic Review*, 112(11):3489–3527.

Jumper, J., Evans, R., Pritzel, A., and et al (2021). Highly Accurate Protein Structure Prediction with AlphaFold. *Nature*, 596:583–589.

Kahneman, D., Sibony, O., and Sunstein, C. (2021). Noise: A Flaw in Human Judgment. *London: HarperCollins*.

Kamatani, N. (2021). Genes, the Brain, and Artificial Intelligence in Evolution. *Journal of Human Genetics*, 66:103–109.

Karnofsky, H. (2021a). Digital People Would Be An Even Bigger Deal. *Cold Takes Blog*, 27th July.

Karnofsky, H. (2021b). Forecasting Transformative AI, Part 1: What Kind of AI? *Cold Takes Blog*, 10 Aug.

Karnofsky, H. (2021c). This Can't Go On. *Effective Altruism Forum*, 3 August.

Karnofsky, H. (2021d). Weak Point in Most Important Century: Lock-In. *Cold Takes Blog*.

Katz, Y. (2012). Noam Chomsky on Where Artificial Intelligence Went Wrong: An Extended Conversation with the Legendary Linguist. *The Atlantic*, 1 November.

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

Knight, F. (1933). Risk, Uncertainty and Profit. *Houghton Mifflin Co, Boston*.

Knox, J. (2022). The Metaverse, or the Serious Business of Tech Frontiers. *Postdigital Science Education*, 4:207–215.

Kokotajlo, D. and Dai, W. (2019). The Main Sources of AI Risk? *AI Alignment Forum*, 21 March.

Koppl, R., Deveraux, A., herriot, J., and Kauffman, S. (2019). The Industrial Revolution as a Combinatorial Explosion. *Mimeo*.

Kotlikoff, L. (2022). Does Prediction Machines Predict Our AI Future? A Review. *Journal of Economic Literature*, 60(3):1052–1057.

Kremer, M. (1993). The O-Ring Theory of Economic Development. *The Quarterly Journal of Economics*, (108):551 – 575.

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

Kuriksha, A. (2021). An Economy of Neural Networks: Learning from Heterogeneous Experiences. *PIER Working Paper 21-027, University of Pennsylvania*.

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

LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence - version 0.9.2, 2022-06-27. *Mimeo: Courant Institute of Mathematical Sciences, New York University.*

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep Learning. *Nature*, 521:436–444.

Lee, C. (2019). The Game of Go: Bounded Rationality and Artificial Intelligence. *Yusof Ishak Institute Working Paper no. 2019-04.*

Lee, R. (1988). Induced Population Growth and Induced Technological Progress: Their Interaction in the Accelerating Stage. *Mathematical Population Studies*, 1(3):265–288.

Leibfried, F. and Braun, D. A. (2016). Bounded Rational Decision-Making in Feedforward Neural Networks. *arXiv:1602.08332v2 [cs.AI].*

LeRoy, S. and Singell, L. (1987). Knight on Risk and Uncertainty. *Journal of Political Economy*, 95(2):394–406.

Lichtner-Bajjaoui, A. (2020). A Mathematical Introduction to Neural Networks. *Advanced Mathematics Master Thesis, Universitat de Barcelona.*

Lipnowski, E. and Doron, R. (2022). Predicting Choice from Information Costs. *arXiv:2205.10434.*

List, J. and Haigh, M. (2005). A Simple Test of Expected Utility Theory Using Professional Traders. *Proceedings of the National Academy of Science USA=*, 102(3):945–948.

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.*

Macaulay, T. (2021). Slaughterbots are a Step Away from your Neighborhood - And We Need a Ban. *Neural*, 13 Dec.

Marcus, G. (2022). Artificial General Intelligence Is Not as Imminent as You Might Think. *Scientific American*, 1 July.

Martin, I. and Pindyck, R. S. (2015). Averting Catastrophes: The Strange Economics of Scylla and Charybdis. *American Economic Review*, 105(10):2947–2985.

Maschler, M., Solan, E., and S.Zamir (2013). Game (t)heory. *2nd edition. Cambridge: Cambridge University Press.*

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

Matthews, D. (2022). Humanity was Stagnant for Millennia - Then Something Big Changed 150 Years Ago. *Vox*, September 7.

McMahon, B. (2022). AI is Ushering In a New Scientific Revolution. *The Gradient*, 4 June.

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

Mokyr, J. (2014). Growth and Technology: The Wild Ride Ahead. *Milken Institute Review*, April 28.

Monton, B. (2019). How to Avoid Maximizing Expected Utility. *Philosophers Imprint*, 19(8):1–25.

Moscati, I. (2016). Retrospectives: How Economists Came to Accept Expected Utility Theory: The Case of Samuelson and Savage. *Journal of Economic Perspectives*, 30(2):219–236.

Müller, V. (2014). Risks ofGeneral Artificial Intelligence. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(3):297–301.

Murphy, T. (2012). Exponential Economist Meets Finite Physicist. *Do The Math Blog*, 4 October.

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

Murphy, T. (2022b). Limits to Economic Growth. *Nature Physics*, 18:844–847.

Murray, A., Kuban, S., Josefy, M., and Anderson, J. (2021). Contracting in the Smart Era: The Implications of Blockchain and Decentralized Autonomous Organizations for Contracting and Corporate Governance. *Academy of Management Perspectives*, 35(4):622–641.

Naudé, W. (2021). Artificial Intelligence: Neither Utopian nor Apocalyptic Impacts Soon. *Economics of Innovation and New Technology*, 30(1):1–24.

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

Naudé, W. and Dimitri, N. (2020). The Race for an Artificial General Intelligence: Implications for Public Policy. *AI & Society*, 35(2):367–379.

Neck, R. (2021). Methodological Individualism: Still a Useful Methodology for the Social Sciences? *Atlantic Economic Journal*, 49:349–361.

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.

Nordhaus, W. D. (2009). An Analysis of the Dismal Theorem. *Cowles Foundation Discussion Paper no. 1686*.

North, D. (1991). Institutions. *Journal of Economic Perspectives*, 5(1):97–112.

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.

O'Brien, J. and Nelson, C. (2020). Assessing the Risks Posed by the Convergence of Artificial Intelligence and Biotechnology. *Health Security*, 187(3):219–227.

Oesterheld, C. (2021). Approval-Directed Agency and the Decision Theory of Newcomb-Like Problems. *Synthese*, 198(27):6491–6504.

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

Omohundro, S. (2008a). The Basic AI Drives. *Proceedings of the First AGI Conference*.

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

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

Parkes, D. and Wellman, M. P. (2015). Economic Reasoning and Artificial Intelligence. *Science*, 6245:267–272.

Pastor-Berniera, A., Plott, C., and Schultz, W. (2017). Monkeys Choose as if Maximizing Utility Compatible with Basic Principles of Revealed Preference Theory. *PNAS*, 114(10):E1766–E1775.

Pattanayak, K. and Krishnamurthy, V. (2021). Rationally Inattentive Utility Maximization Explains Deep Image Classification. *arXiv:2102.04594 [cs.LG]*.

Pearl, J. (1985). Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning. *Proceedings of the 7th conference of the Cognitive Science Society, University of California, Irvine, CA, USA*.

Pearl, J. and Mackenzie, D. (2018). The Book of Why: The New Science of Cause and Effect. *London: Random House Penguin*.

Persky, J. (1995). Retrospectives: The Ethology of Homo Economicus. *Journal of Economic Perspectives*, 9(2):221–231.

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

Pigozzi, G., Tsoukias, A., and Viappiani, P. (2016). Preferences in Artificial Intelligence. *Annals of Mathemathical Artificial Intelligence*, 77:361–401.

Pinker, S. and Aaronson, A. (2022). Steven Pinker and Scott Aaronson Debate AI Scaling! *Collate*, 28 June.

Pistono, F. and Yampolskiy, R. (2016). Unethical Research: How to Create a Malevolent Artificial Intelligence. *arXiv:1605.02817 [cs.AI]*.

Pooley, G. and Tupy, M. L. (2018). The Simon Abundance Index: A New Way to Measure Availability of Resources. *Policy Analysis no. 857, Cato Institute*.

Ramsey, F. (1931). Truth and Probability. *In F. Ramsey (ed.). Foundations of Mathematics and other Logical Essays. New York: Harcourt.*

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

Ricci, F., Rokach, L., and Shapira, B. (2022). Recommender Systems: Techniques, Applications, and Challenges. *In F. Ricci and L. Rokach and B. Shapira (eds.). Recommender Systems Handbook (3 ed.). New York: Springer*, pages 1–35.

Riedel, C. J. (2021). Value Lock-in Notes. *Mimeo*, 25 July.

Ring, M. and Orseau, L. (2011). Delusion, Survival, and Intelligent Agents. *In J. Schmidhuber and K.R. Thórisson and M. Looks (eds.) AGI 2011. LNCS (LNAI)*, 6830:11–20.

Romer, P. (1986). Increasing Returns and Long-Run Growth. *Journal of Political Economy*, 94(5):1002–1037.

Romer, P. (1987). Growth Based on Increasing Returns Due to Specialization. *American Economic Review*, 77(2):56–62.

Romer, P. (1990). Endogenous Technical Change. *World Development*, (5):S71–S102.

Romer, P. (2019). The Deep Structure of Economic Growth. *Blog: https://paulromer.net*, 5 February.

Roodman, D. (2020). Modelling the Human Trajectory. *Open Philanthropy*, 15 June.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65:386–407.

Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning Representations by Back-Propagating Errors. *Nature*, 323:533–536.

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. *Viking Books*.

Russell, S. and Norvig, P. (2021). Artificial Intelligence: A Modern Approach. *4th Edition. Pearson Education, Inc.*

Salimans, T., J., Chen, X., Sidor, S., and Sutskever, I. (2017). Evolution Strategies as a Scalable Alternative to Reinforcement Learning. *arXiv:1703.03864v2 [stat.ML]*.

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

Samuelson, P. (1947). Foundations of Economic Analysis. *Cambridge: Harvard University Press*.

Sandberg, A. (2013). An Overview of Models of Technological Singularity. *In M. More and N. Vita-More (eds.). The Transhumanist Reader. Wiley*.

Sandberg, A. (2014). Ethics of Brain Emulations. *Journal of Experimental & Theoretical Artificial Intelligence*, 26:439–457.

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

Santana, C. and Albareda, L. (2022). Blockchain and the Emergence of Decentralized Autonomous Organizations (DAOs): An Integrative Model and Research Agenda. *Technological Forecasting & Social Change*, 182.

Sarker, I. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, 2(6):420.

Savage, L. (1954). The Foundations of Statistics. *New York: Dover*.

Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. *Neural Networks*, 6:85–117.

Schoemaker, P. (1982). The Expected Utility Model: Its Variants, Purposes, Evidence and Limitations. *Journal of Economic Literature*, 20(2):529–563.

Searle, J. (1992). The Rediscovery of the Mind. *Cambridge MA: MIT Press*.

Sejnowski, T. J. (2020). The Unreasonable Effectiveness of Deep Learning in Artificial Intelligence. *PNAS*, 117(48):30033–30038.

Setiya, K. (2022). The New Moral Mathematics. *Boston Review*, 15 August.

Shah, R. (2019a). AI Safety without Goal Directed Behaviour. *Alignment Forum*, 7 January.

Shah, R. (2019b). Stuart Russell's New Book on Why We Need to Replace the Standard Model of AI. *Alignment Newsletter no. 19*.

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.

Shulman, C. (2010). Omohundro's "Basic AI Drives and Catastrophic Risks. *The Singularity Institute, San Francisco, CA*.

Shulman, C. and Bostrom, N. (2021). Sharing the World with Digital Minds. *In S. Clarke, H. Zohny, and J. Savulescu (eds.) Rethinking Moral Status. Oxford: Oxford Academic*.

Simon, H. (1955). A Behavioral Model of Rational Choice. *Quarterly Journal of Economics*, 69(1):99–118.

Simon, H. (1956). Rational Choice and the Structure of the Environment. *Psychological Review*, 63(2):129–138.

Simon, H. (1978). On How to Decide What to Do. *The Bell Journal of Economics*, 9(2):494–507.

Sims, C. (2003). Implications of Rational Inattention. *Journal of Monetary Economics*, 50(3):665–690.

Singer, P. (2021). The Hinge of History. *Project Syndicate*, 8 October.

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.*

Srnicek, N. (2016). Platform Capitalism. *London: Polity.*

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.

Sutton, R. and Barto, A. (1998). Introduction to Reinforcement Learning. *1st ed. Cambridge, MA: MIT Press.*

Tamura, H. (2009). Modeling Ambiguity Averse Behavior of Individual Decision Making: Prospect Theory under Uncertainty. *In Torra, V., Narukawa, Y., Inuiguchi, M. (eds.). Modeling Decisions for Artificial Intelligence. MDAI 2009. Lecture Notes in Computer Science, vol 5861. Springer, Berlin, Heidelberg.*

Tarsney, C. (2020). The Epistemic Challenge to Longtermism. *Mimeo: GPI, Oxford.*

Taylor, I. (2016). Dependency redux: Why Africa is not rising. *Review of African Political Economy*, (43):8–25.

Thaler, R. (2000). From Homo Economicus to Homo Sapiens. *Journal of Economic Perspectives*, 14(1):133–141.

Tirole, J. (2021). Digital Dystopia. *American Economic Review*, 111(6):2007–2048.

Torres, E. (2021). The Dangerous Ideas of Longtermism and Existential Risk. *Current Affairs*, 28th July.

Torres, E. (2022). Understanding "Longtermism": Why this Suddenly Influential Philosophy is so Toxic. *Salon*, 20 August.

Torres, P. (2014). Why Running Simulations May Mean the End is Near. *Mimeo: https://archive.ieet.org/articles/torres20141103.html*.

Totschnig, W. (2020). Fully Autonomous AI. *Science and Engineering Ethics*, 26:2473–2485.

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*.

Tunyasuvunakool, K., Adler, J., Wu, Z., and et al (2021). HighlyAaccurate Protein Structure Prediction for the Human Proteome. *Nature*, 595:590–596.

Turchin, A. (2021). Catching Treacherous Turn: A Model of the Multilevel AI Boxing. *Mimeo: Foundation Science for Life Extension*, May.

Turchin, A. and Chernyakov, M. (2018). Classification of Approaches to Technological Resurrection. *Mimeo*.

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

Turing, A. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem. 42:230–265.

Umbrello, S., Torres, P., and De Bellis, A. (2020). The Future of War: Could Lethal Autonomous Weapons Make Conflict more Ethical? *AI & Society*, 35:273–282.

Urbina, F., Lentzos, F., Invernizzi, C., and Ekins, S. (2022). Dual Use of Artificial-Intelligence-Powered Drug Discovery. *Nature Machine Intelligence*, 4:189–191.

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.

Varian, H. (1995). Economic Mechanism Design for Computerized Agents. *Proceedings of the First USENIX Workshop on Electronic Commerce New York, New York*, July.

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.

Villaescusa-Navarro, F., Ding, J., Genel, S., Tonnesen, S., La Torre, V., Spergel, D., R.Teyssier, Li, Y., Heneka, C., Lemos, P., Anglés-Alcázar, D., Nagai, D., and Vogelsberger, M. (2022). Cosmology with one Galaxy? *arXiv:2201.02202 [astro-ph.CO]*.

Vinge, V. (1993). The Coming Technological Singularity: How to Survive in the Post-Human Era. *VISION-21 Symposium,NASA Lewis Research Center and the Ohio Aerospace Institute*, March 30-31.

Von Foerster, H., Mora, P. M., and Amio, L. (1960). Doomsday: Friday, 13 November, A.D. 2026. *Science*, 132(3436):1291–1295.

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

Von Neumann, J. and Morgenstern, O. (1944). Theory of Games and Economic Behavior. *Princeton NJ: 1st Ed. Princeton University Press*.

Wagner, D. (2020). Economic Patterns in a World with Artificial Intelligence. *Evolutionary and Instutional Economics Review*, 17:111–131.

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*.

Weitzman, M. (1998). Recombinant Growth. *Quarterly Journal of Economics*, 113(2):331–360.

Weitzman, M. (2009). On Modeling and Interpreting the Economics of Catastrophic Climate Change. *Review of Economics and Statistics*, 91.(1):1–19.

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]*.

Williams, G. (1966). Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought. *Princeton, NJ: Princeton University Press*.

Wooldridge, M. (2022). What Is Missing from Contemporary AI? The World. *AAAS Intelligent Computing*, 2022:9847630.

Wright, R. (2022). The Case for Shorttermism. *Nonzero Blog*, 11 August.

Yampolskiy, R. (2014). Utility Function Security in Artificially Intelligent Agents. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(3):373–389.

Yampolskiy, R. (2016). Taxonomy of Pathways to Dangerous Artificial Intelligence. *The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence AI, Ethics, and Society: Technical Report WS-16-02*.

Yampolskiy, R. V. (2012). Leakproofing the Singularity. *Journal of Consciousness Studies*, 19(1-2):194–214.

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. (2001). Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures. *The Singularity Institute, San Francisco, CA*, June 15.

Yudkowsky, E. (2002). The AI-Box Experiment. *https://www.yudkowsky.net/singularity/aibox*.

Yudkowsky, E. (2007a). 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. (2007b). Pascal's Mugging: Tiny Probabilities of Vast Utilities. *Less Wrong Blog*, 19 October.

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*.

Yudkowsky, E. (2009). Value is Fragile. *Less Wrong Blog*, 29th Jan.

Yudkowsky, E. (2017). There's No Fire Alarm for Artificial General Intelligence. *Machine Intelligence Research Institute*, October 13.

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.

Zhang, K., Yang, Z., and Basar, T. (2021). Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. *arXiv:1911.10635v2 [cs.LG]*.

Zheng, S., Trott, A., Srinivasa, S., Naik, N., Gruesbeck, M., Parkes, D., and Socher, R. (2020). The AI Economist: Improving Equality and Productivity with AI-Driven Tax Policies. *arXiv:2004.13332v1 [econ.GN]*.

Zuboff, S. (2015). Big Other: Surveillance Capitalism and the Prospects of an Information Civilization. *Journal of Information Technology*, 30(1):75–89.