ISSN: 2029-4913 / eISSN: 2029-4921

VILNIUS TECH
Vilnius Gediminas Technical University

TECHNOLOGICAL and ECONOMIC DEVELOPMENT of ECONOMY

2024
Volume 30
Issue 6
Pages 1682–1708
https://doi.org/10.3846/tede.2024.21525

EXISTENTIAL RISK FROM TRANSFORMATIVE AI: AN ECONOMIC PERSPECTIVE

Jakub GROWIEC

Department of Quantitative Economics, SGH Warsaw School of Economics, Warszawa, Poland

Article History:
- received 06 July 2023
- accepted 06 March 2024
- first published online 10 July 2024

Abstract. The prospective arrival of transformative artificial intelligence (TAI) will be a filter for the human civilization – a threshold beyond which it will either strongly accelerate its growth, or vanish. Historical evidence on technological progress in AI capabilities and economic incentives to pursue it suggest that TAI will most likely be developed in just one to four decades. In contrast, theoretical problems of AI alignment, needed to be solved in order for TAI to be "friendly" towards humans rather than cause our extinction, appear difficult and impossible to solve by mechanically increasing the amount of compute. This means that transformative AI poses an imminent existential risk to the humankind which ought to be urgently addressed. Starting from this premise, this paper provides new economic perspectives on discussions surrounding the issue: whether addressing existential risks is cost effective and fair towards the contemporary poor, whether it constitutes "Pascal's mugging", how to quantify risks that have never materialized in the past, how discounting affects our assessment of existential risk, and how to include the prospects of upcoming singularity in economic forecasts. The paper also suggests possible policy actions, such as ramping up public funding on research on existential risks and AI safety, and improving regulation of the AI sector, preferably within a global policy framework.

Keywords: transformative artificial intelligence, artificial general intelligence, alignment, existential risk, long-run economic growth, longtermism.

JEL Classification: J17, O33, O40.

Corresponding author. E-mail: ***jakub.growiec@sgh.waw.pl***

1. Introduction

On March 14, 2023 OpenAI introduced its new large language model GPT-4 – an AI algorithm so big that the company chose not to publicly disclose the information on the number of its hyperparameters, hardware capacity or training compute (OpenAI et al., 2023). Compared to its predecessors GPT-3/GPT-3.5 and language models from OpenAI's competitors, GPT-4 achieved impressive progress on a variety of important benchmarks, massively extending the number of tasks which AI algorithms are able to perform at or above the human level. During the GPT-4 launch, the public was also informed that since February the same algorithm had already been powering the AI chatbot built into Microsoft Bing, with real-time access to the Internet. OpenAI had also performed preliminary assessments suggesting that GPT-4 is ineffective at autonomously replicating, acquiring resources, and avoiding being shut down "in the wild" (OpenAI et al., 2023, p. 54). In order to check that, a "red team" at OpenAI combined GPT-4 with a simple read-execute-print loop that allowed the model to execute code, do chain-of-thought reasoning, and delegate to copies of itself.

So far, so good. However, GPT-4 is just one more step on a steep incline of AI capabilities. Progress does not stop there, and month by month, consecutive generations of large-scale AI algorithms are becoming more agentic, more powerful, and more skilled at replicating, self-improving and acquiring resources. In the future, this growth process may potentially culminate in the arrival of transformative artificial intelligence (TAI), an AI algorithm able to act as an independent, autonomous agent seeking to achieve its goals and exhibiting superhuman performance at a broad array of tasks, including all tasks which are essential for the economy. That TAI would also be able to replicate, acquire resources, avoid being shut down, and pursue large-scale transformations of its environment. For the humankind, that would be an existential risk.

In this article I reiterate the point that development of ever more general and powerful AI algorithms poses an existential threat to humankind, and that given the observed trend in AI capabilities (increasing super-exponentially in line with the allocation of computing power to the training of cutting-edge AI models, see Figure 1), this threat is imminent rather than distant. Therefore, investment in existential risk reduction, particularly from deploying "unfriendly" TAI, should be among humankind's top priorities.

The contribution of this paper to the literature is to organize and provide new economic perspectives on a number of threads of discussion related to the existential risk from TAI, complementary to the existing perspectives coming predominantly from philosophy and computer science. The backdrop is that in the *laissez-faire*, business as usual scenario, given overwhelming economic incentives to develop ever more advanced and general AI algorithms and a very low *ex ante* probability that the goals of these algorithms will be well aligned

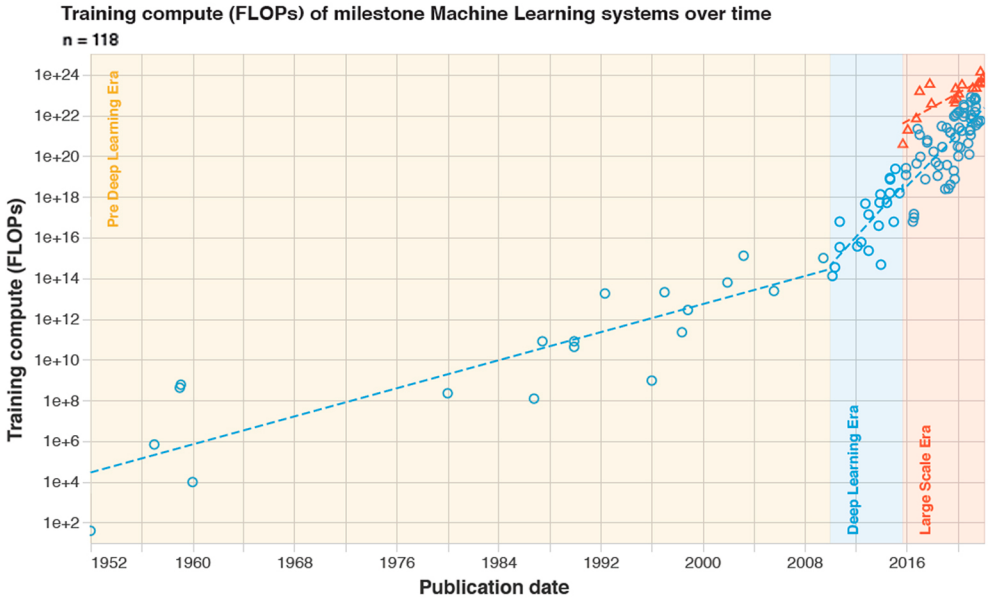

Figure 1. Computing power used to train cutting-edge AI models in the "Deep Learning era" is doubling every six months (source: Sevilla et al., 2022, under CC-BY-4.0 license)

with long-run human flourishing, the default outcome is human extinction (Muehlhauser & Salamon, 2012). In contrast, if the alignment problem (i.e., the problem of aligning TAI's goals with long-run human flourishing) is successfully solved, thanks to TAI the human civilization will likely experience another major development acceleration of a magnitude comparable to the Agricultural Revolution and the Industrial Revolution. The prospective arrival of TAI will be a *filter* for the human civilization, deciding whether it degenerates or accelerates development and potentially spreads into the cosmos – and whether the humankind goes extinct or lives on, potentially for millennia or at least until a next filter emerges.

Specifically the paper addresses the following threads:

- *Economic incentives for developing TAI.* Digital technologies are characterized by increasing returns to scale, which lead to "winner takes all" market share dynamics and amplify competition. Furthermore, developing TAI will lock in the objectives programmed by the winning AI lab, instilling its preferences on the humanity's entire future. In result, AI labs may continue to race towards TAI even when they know it is likely going to be "unfriendly", due to the pressure to outrun their competitors.
- *The importance and difficulty of AI alignment.* Due to the instrumental convergence thesis (Bostrom, 2014), almost any TAI will develop auxiliary goals which will make it "power seeking" and impossible to terminate or reprogram. Therefore, there is no room for trial and error, and if we are to avoid the existential risk, already the first AI to achieve the human level of general intelligence must have an objective function which is well aligned with long-term well-being of the humankind.
- *Longtermism vs. needs of the present.* If the existential risk were to materialize only in distant future, one could argue that investing in its reduction diverts funds from the needs of the contemporary people. However, the risk is imminent.
- *Cost efficiency and fairness of investments in AI safety.* Even under discounted utilitarianism with a preference for reducing global inequality and poverty, investing in existential risk reduction is still the correct choice if the existential risk is sufficiently imminent and large.
- *Valuation of extinction risk and "Pascal's mugging".* Some risks may be severely undervalued in the markets, particularly the ones for which no historical data exists. It is the case for extinction risk from misaligned TAI. However, because of the imminence of this risk, calls for its reduction do not constitute "Pascal's mugging".
- *Discounting and the value of distant future.* If the existential risk were to materialize only in distant future, under discounted utilitarianism its impact on humankind's expected utility should be small, differently to the claims of longtermists (e.g., McAskill, 2022). However, the risk is imminent.
- *Technological singularity and economic forecasts.* The prospect of TAI produces scenarios of technological singularity which are alien to the economics literature, prompting to embrace the possibility of human extinction and to distinguish between the future of humanity and the future of the human civilization.

If the prospective arrival of TAI will be a filter for the human civilization, what should be the appropriate policy response? The opinions, it appears, are mixed, varying from Luddite-

ish calls to stop all AI research altogether, over requests for governmental regulation of the AI sector (OpenAI et al., 2023), calls for slowing down the development of "dangerous AI" in order to first make progress in terms of AI safety (Grace, 2022) or issuing a moratorium on general AI capabilities research akin to the 1975 Asilomar moratorium on research on recombinant DNA[1], to fully optimistic/reckless replies which negate any existential risk from TAI. Furthermore, some actors have already taken action: OpenAI unilaterally initiated an internally funded project which aims at solving the alignment problem (Leike & Sutskever, 2023). There appears to be a general consensus that TAI will be both potentially massively beneficial and very dangerous, but the assessment of the severity of that danger and its imminence varies substantially. There are also voices suggesting that TAI will never be invented, constituting a small minority (<2%) of AI experts (Roser, 2023), but a majority of economists (Nordhaus, 2021; Davidson, 2021) and broader public. The latter observation relates to the fact that the topic has only recently started to appear in popular debates among the general audience.

The best of possible futures is probably the one with friendly superhuman general AI, perfectly aligned with human flourishing, which would allow our species and our civilization to develop, rapidly improve our capability to pursue our goals and fulfill our needs, and potentially allow us to conquer vast swaths of the universe. Stopping any further AI research would voluntarily forego this future; moreover, such prohibitive policy would also be likely unsustainable given the enormous economic incentives to proceed with AI development anyway and the miserable state of international policy coordination. It should also be remembered that with superhuman general AI the humankind will become less vulnerable to other sources of existential risk (Ord, 2020).

So rather than stopping all research on large-scale AI models, a more modest and feasible policy proposal would be to strongly prioritize the research on existential risks and AI safety, perhaps using public funds. Moreover, as the plausibility of achieving AI alignment in due time is problematic even with strongly improved funding because of the short predicted timelines to transformative AI (Cotra, 2020; Roser, 2023; Grace et al., 2024), policymakers should also consider taking steps aiming at regulating the AI sector and slowing down AI progress – which would be clearly helpful with the alignment challenge (Grace, 2022).

The remainder of this article is structured as follows. In the next section I discuss the background of the current paper, including longtermism and economic growth theory; next I review the potential impacts of transformative AI on the global economy: its promises, threats, instrumental convergence, the alignment problem and the current state of affairs. In the following section I provide an economist's review of arguments which criticize longtermism and its emphasis on existential risk reduction, voiced in the literature as well as popular press. Finally, I collect the policy implications and conclude.

[1] Symbolically, during the 2017 Asilomar Conference on AI, 23 principles of AI research were agreed upon. However, these principles are rather general and do not steer AI research in any particular direction.

2. Transformative AI and the global economy

Over the last four decades, information and communication technologies (ICTs) have rapidly transformed the world. Computers, Internet, and smartphones have permeated households and workplaces alike. ICTs are increasing labor productivity across the economy[2], and sometimes also are a direct source of utility as consumption goods.

ICTs boost labor productivity by improving people's capacity to communicate and process information. This is possible thanks to the facts that (a) digital devices have a massive advantage over human brains in the pace of numerical computation, (b) they have the capacity to store and run their code. Computationally intensive tasks which were previously performed in people's brains are therefore increasingly performed digitally, freeing people's minds to concentrate on higher-order tasks, as well as allowing people to engage in tasks which they would not be able to accomplish without digital help at all.

But ICTs act both as brain enhancement and replacement. Indeed, an increasing number of tasks is being automated, allowing them to be performed without any human input (Acemoglu & Autor, 2011; Frey & Osbourne, 2017). Routine, repetitive, easily codifiable tasks are first to automate. This reduces the demand for some jobs, foremostly in manufacturing, pushing those laid off to find employment elsewhere, for example in services.

At the same time, new tasks are being created, offering new job opportunities to people. The dynamic where simultaneously old job tasks are automated, while new job tasks are created for people to perform (Acemoglu & Restrepo, 2018), can be referred to as the "race against the machine".

However, the "race against the machine" dynamic breaks down in the face of artificial intelligence: unlike other automation technologies such as spreadsheets or pre-programmed robots, AI improves its performance with data and computing power, and has the potential to increase the breadth of its application. Large language models – most powerful AI algorithms built to date – are already able to automate sophisticated, nonroutine cognitive tasks performed thus far by skilled professionals (Eloundou et al., 2023; Korinek, 2023)[3]. In the future, some of the newly created tasks may no longer be performed by people, but rather by AI already from day one. In the end, we may eventually witness the arrival of transformative AI, able to *fully* automate production (Growiec, 2022b) and present an absolute advantage in *all* economically meaningful tasks, including the creation of new tasks, developing AI, and strategic decision making.

In this article, note, I am not distinguishing between *transformative* AI and *superhuman general* AI. Both concepts are used here to designate hypothetical AI algorithms which exhibit superhuman performance at a broad array of tasks, including all tasks which are essential

[2] However, the impact of digital technologies on aggregate labor productivity was somewhat underwhelming thus far, as illustrated by the often-repeated complaint by Solow (1987): "you can see the computer age everywhere but in the productivity statistics". Fueled by this disappointment, some authors such as Gordon (2016) expect that digital technologies will not have a significant impact on productivity growth in the future.

[3] One has to keep in mind, though, that technology adoption comes with a lag and major progress in AI has been achieved only very recently. For these reasons the measured impacts of AI on productivity growth are very small in historical data (Parteka & Kordalska, 2023). Accordingly, thus far AI tended to increase employment in sectors exposed to it (Albanesi et al., 2023); this trend will probably reverse as AI algorithms become more advanced and more broadly adopted in the economy (Korinek & Juelfs, 2022).

for the economy. While the exact definitions are subject to dispute (cf. Gruetzemacher & Whittlestone, 2021), the bottom line is that superhuman narrow AI (like, e.g., AlphaZero or AlphaFold) cannot be transformative because it only performs a very narrow set of tasks; conversely, sub-human (say, ant-level) general AI cannot be transformative because it would perform the tasks too badly to be of practical use in the economy.

2.1. Longtermism vs. economic growth theory

The promises and risks of potential future development of TAI can be addressed both from the perspective of long-run economic growth theory and the philosophical standpoint of longtermism. In a nutshell, *longtermism* is an ethical stance which gives priority to improving the long-term future of humanity and the human civilization (Ord, 2020; McAskill, 2022). Growth economists, in turn, study the mechanisms of technological progress and economic growth which determine this long-term future; in some of the more normative studies, they may also provide policy recommendations which could improve it. The philosophical basis of both longtermism and economic growth theory is utilitarianism: any hypothetical long-term future of humanity is assessed on the basis of the expected aggregate level of utility among people who will live in that future (Parfit, 1984; McAskill, 2022).

An important aspect of longtermist thinking is the emphasis on reducing existential risks to humanity. In contrast, this perspective is a blind spot for growth economics, in which existential risks are usually ignored[4]. Closest related are economic studies which deal with severe but usually non-existential catastrophes caused by climate change (e.g., Chichilnisky, 2000; Chichilnisky et al., 2020) or miscellaneous other phenomena (e.g., Martin & Pindyck, 2015), such as epidemics, nuclear terrorism, bioterrorism, floods, storms and earthquakes. Yet neither natural events (such as asteroid impacts, volcanic explosions, gamma ray bursts) nor slow-onset manmade disasters such as climate change are likely to culminate in human extinction; to the contrary, the list of existential risks which are most likely to materialize is topped by misaligned transformative AI, followed by the risks of large-scale nuclear war and engineered pandemics (Sandberg & Bostrom, 2008; Ord, 2020).

Both philosophical longtermists and theorists of long-term economic growth agree that over the last decades the development of digital technologies, including AI algorithms, is one of the key drivers of technological change. These technologies have been developing an order of magnitude faster than "traditional" technologies driving historical GDP growth (Hilbert & Lopez, 2011; Growiec, 2022a), and in the future may accelerate economic growth by replacing human cognitive work with automated information processing in production, research and development, and decision making. Unlike most economists, philosophical longtermists – just like many AI researchers and industry leaders – recognize that TAI is an existential threat to humanity.

Worryingly, available estimates of the total risk of humankind not surviving to the 22nd century are high, in sharp contrast with the relatively scarce interest in the topic among economists, politicians, and the population at large. According to Ord (2020), that probability

[4] Aschenbrenner (2020) and Trammell (2021) are two notable exceptions. However, both were prepared at the Oxford University's Global Priorities Institute, a hub of longtermist thought. Very recently, the topic of existential risk from advanced AI has been also addressed by Jones (2023).

is about one in six (16,7%), with about 10% contributed by TAI. In turn, according to a survey of scholars attending the Global Catastrophic Risk Conference in Oxford in 2008, there is a 19% probability of human extinction by 2100, with 5 pp. contributed by superhuman AI and another 5 pp. contributed by molecular nanotech weapons (Sandberg & Bostrom, 2008). Even more worryingly, based on his review of trends and situations facing humanity, Rees (2003) estimated a whole 50% probability of human extinction by 2100.

In light of the facts highlighted in this paper, the existential threat from TAI is imminent rather than distant. Therefore, investment in existential risk reduction, particularly from deploying misaligned TAI, is a cause that should be prioritized irrespective of one's moral stance on the relative importance of current vs. future generations – in fact, even if one cares only for the well-being of people alive today, or a subset thereof.

2.2. Promises of transformative AI

Long-run economic growth is driven by the accumulation of production factors as well as technological innovations that are subsequently adopted in the economy. In the industrial era, there were two main, mutually complementary factors of production: physical capital and human cognitive work (Romer, 1990; Klump et al., 2012); of these two factors, only the latter drove long-run economic growth. Specifically, throughout the 20th century the pace of global economic growth was determined by the pace of growth in effective, technologically augmented human cognitive work (Romer, 1990; Bloom et al., 2020; Growiec, 2022a). Complementary factors, like machines performing physical actions, were sufficiently abundant so as not to affect the growth rate in the long run equilibrium.

Whilst being the decisive growth *engine*, technologically augmented human cognitive work was also the key growth *bottleneck* – i.e., the factor whose scarcity crucially constrained the pace of economic growth. However, the advent of ICTs, and AI in particular, has allowed to gradually detach information communication and processing from the capabilities of the human brain, making room for full automation of production processes, which would remove the bottleneck and accelerate economic growth, potentially even by an order of magnitude (Trammell & Korinek, 2020; Davidson, 2021; Growiec, 2022a, 2023).

This intriguing possibility arises because there already is an order of magnitude difference in the pace of growth in global GDP, which doubles every 20–30 years (Piketty, 2014), and the cumulative capacity of digital data communication, storage and processing, which doubles every 2–3 years (Hilbert & Lopez, 2011). Thus far rapid growth in the information sphere has not been translating into proportionally fast growth in global GDP because ICTs had limited capabilities: tasks could be automated only partially, and even in highly automated activities human oversight and managerial decision making were still necessary. Within tasks, information processing by people and machines is always substitutable, but as long as the tasks are complementary and some of them cannot be automated, people and machines remain complementary in the aggregate, and human cognitive work remains the bottleneck of economic growth (Growiec, 2022b).

In contrast, the hypothetical future TAI would contribute to *all* economically essential tasks, including research tasks, and even tasks aimed at improving AI (that is, its own) capabilities. Therefore, it could potentially replace human cognitive work across all tasks, ren-

dering people and machines substitutable not just within tasks, but also in the aggregate. This is precisely why TAI would be economically *transformative*: by offering the prospects of fully automating all economically relevant tasks, it will remove the bottleneck generated by the limits to human cognitive capabilities, and potentially accelerate growth by an order of magnitude (Growiec, 2022b, 2023).

There is a range of promises of transformative AI which justify the efforts to develop it.

First, there are enormous economic rewards awaiting the firm which will first introduce TAI to the market. Digital technologies are characterized by increasing returns to scale, creating winner-takes-all ("superstar") market share dynamics and producing natural monopolies which can later entrench themselves and fend off competition (Autor et al., 2020). Already today, there is just a handful of software giants in the global market and world's biggest fortunes are made in the software business. With *generality* of the prospective TAI, the scope of the contended market will expand further, to cover all activities in which the AI will be deployed with a productivity advantage – that is, potentially *all* economy. On top of that, the first entity to deploy TAI will have an opportunity to increase its control over the world even beyond the extent captured by its market shares and valuations. By setting the TAI's objectives, it will project its preferences on the humanity's entire future.

Second, as argued above, TAI will massively boost aggregate productivity growth at the global scale, possibly accelerating its growth by an order of magnitude. Such increases in the "size of the pie" of wealth to be distributed among the world population are a great promise even if that comes at the cost of gradually rendering human cognitive work obsolete and largely increasing income inequality[5].

Third, massive positive feedback effects can be expected from the participation of TAI in research. Specifically, its contribution to AI research may cause a cascade of recursive self-improvements culminating in the TAI undergoing intelligence explosion (Hanson & Yudkowsky, 2013; Bostrom, 2014) and elevating its performance far above the human level[6]. Furthermore, TAI may use its superior cognitive powers to develop new, more efficient ways of harnessing solar energy and putting it into productive use, allowing our civilization to cross another threshold in access to energy (cf. Growiec, 2022a), following the earlier breakthroughs of the Agricultural Revolution (~10 000 BP) and the Industrial Revolution (~1800 CE), and advancing our civilization on the Kardashev scale.

In the longtermist view, developing transformative AI is necessary to allow humankind to realize its vast future potential – to survive on Earth for millions of years, colonize other planets, and reach out to outer space. Such goals appear difficult if not impossible for the human civilization to achieve while relying only on our brains and specialized subhuman computational and AI algorithms for information processing.

[5] How to distribute all this new wealth, given that one could no longer use labor remuneration as the key distributive device, is an open question. In the free market allocation – constituting the default option – all returns to TAI would be captured by the shareholders of the company which introduced it. This would mean that without any change in policy, a huge fraction of value added in the world economy would then be captured by just a handful of people, exacerbating global inequality to unprecedented levels.

[6] Even a sub-human general AI may achieve the capacity to recursively self-improve. In the presence of an overhang of unused computing power, it may then rapidly improve its performance to a superhuman level.

2.3. Existential risk

However, apart from its great promises, TAI is also an existential risk to humanity and a threat to the human civilization. Quantifying this risk requires us to estimate two probabilities: that humanity will one day develop superhuman general AI, and that its goals will be *misaligned*, i.e., not perfectly aligned with human flourishing.

In the AI research community, the consensus appears to be that TAI is technically possible, the disagreement being only on the timing of its expected arrival. The recent progress in large language models has shaken expert predictions considerably. Prior to the rollout of ChatGPT and GPT-4, AI experts predicted the arrival of TAI around 2060 on average in their central estimate (Roser, 2023), while some scholars suggested more aggressive timelines, expecting TAI around 2040[7]. This was the case for example in Cotra's (2020) analysis using a theoretical model parametrized on historical trends in AI training costs and performance as well as some scarce evidence from the evolution of brains across animal species ("bio anchors"). Concurrently, many AI experts believed that the arrival of TAI was "beyond the foreseeable horizon" (Etzioni, 2016). After GPT-4, however, the median prediction among AI experts dropped to 2047 (Grace et al., 2024). Accordingly, the central forecast of the metaculus.com community dropped from about 2041 (prior to GPT-4) to as early as 2032 (as of January 2024). Leike and Sutskever (2023) from OpenAI provide an even sharper timeline, suggesting that AGI will be created most likely before 2030 ("While superintelligence seems far off now, we believe it could arrive this decade.").

This wide discrepancy is partly based on the disagreement whether existing AI methodologies based on the paradigm of deep learning, including generative adversarial networks, convolutional neural networks and network transformers, are sufficient for the emergence of TAI – or a qualitative change in algorithm design is needed. The *scaling hypothesis* (Branwen, 2022), which supposes that existing AI designs, when scaled up by a few orders of magnitude in terms of the number of parameters and training data volumes are sufficient for creating superhuman general AI, is currently being subjected to intense scrutiny. Research teams at OpenAI, DeepMind, Anthropic, Google Brain and others are busy scaling up their neural networks in terms of the number of layers, neurons and parameters. Indeed recent developments in large language models such as OpenAI's GPT-4 have demonstrated new emergent properties of larger networks (Wei et al., 2022; Bubeck et al., 2023), causing some experts to put more weight on the scaling hypothesis and revise their expectations on AI timelines towards earlier dates (e.g., Cotra, 2022; Grace et al., 2024).

But even if state-of-the-art deep learning is not enough and a qualitative change in algorithm design is needed, after all "there is no physical law precluding particles from being organized in ways that perform even more advanced computations than the arrangements of particles in human brains" (Hawking et al., 2014). With today's knowledge, it should be assumed that the probability that TAI will arrive *at some point in time* is certainly positive, and with no changes in policy, perhaps even close to one.

The other variable in the existential risk formula is the probability that the goals of the transformative AI will be misaligned with human flourishing. Unfortunately, according to AI

[7] This timeline curiously coincides with Kurzweil's (2005) famous prediction of technological singularity in 2045.

alignment scholars, the default outcome is a negative one here (Muehlhauser & Salamon, 2012; Bostrom, 2014). As showcased by the launch of GPT-4 and Microsoft Bing AI, the alignment problem is far from solved at the moment, and with misaligned, power-seeking TAI the probability of disaster is close to 100%. After all, with sufficient optimization power, even small discrepancies between the AI's goals (e.g., GPT-4 aims to predict the next word in a sentence with maximum accuracy) and human well-being can be fatal. As Yudkowsky put it, "the AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else" (Yudkowsky, 2008).

Accordingly, unaligned TAI will be an existential threat to humankind purely due to its superior optimization power that would outsmart any human. The disaster scenario does not need additional components such as artificial consciousness or the AI's ability to reflect on its goals (Russell, 2014)[8]. Moreover, following the *orthogonality thesis* (Bostrom, 2014), any level of intelligence could in principle be coupled with any final goal, negating the hope that TAI would unilaterally refuse to harm people. Once the humankind is deprived of control over decision making and any meaningful contribution to the world economy, the ruling AI's decision whether to terminate our species will only depend on whether our existence would be helpful in pursuing its goal; our own goals, potentials or well-being will not be respected.

2.4. Instrumental convergence

In order to make sure that the superhuman general AI will help the humankind rather than destroy it, its objectives must be perfectly aligned with human flourishing. Anyone who is familiar with the example of "apocalypse by paperclips" (Bostrom, 2014), understands the risks involved in misaligned but powerful AI. Namely, the extinction result follows directly from the *instrumental convergence* thesis: with sufficient agency and optimization power, almost any AI algorithm will (a) resist the attempts to switch it off or reprogram its goals (self-preservation), (b) accumulate control over resources it deems helpful in achieving its goal (resource acquisition), (c) use the available resources as efficiently as possible (technological perfection), and (iv) research the possible options for improving its efficiency through new technological solutions (cognitive enhancement). The emergence of these four instrumental goals follows from almost any final goal we may consider programming into AI, and certainly almost any final goal that could make the AI potentially transformative – known counterexamples are trivial and typically imply that the algorithm prefers to immediately switch off. The challenge is therefore that the prospective TAI must be aligned and provide beneficial outcomes to the humankind *despite* following also the instrumental goals, which by default make it "power-seeking".

Unfortunately for the likelihood of solving the AI alignment problem, there is no room for trial and error when experimenting with superhuman general AI. Beyond a certain threshold intelligence level, the goals programmed into the AI will be locked in without a possibility to reprogram them. The superhuman general AI will then be able to outsmart humans and resist any reprogramming which would work against its present goal (Bostrom, 2014).

[8] Nevertheless, as *theory of mind* may have already spontaneously emerged in large language models (Kosinski, 2023), it can be imagined that one day some sort of consciousness or sentience could also emerge in complex AI algorithms.

There is a helpful analogy between the human species and the prospective TAI (Growiec, 2022a). The *homo sapiens* emerged as one of many designs of species developed in the process of natural evolution. The implicit goal of the evolutionary process is to produce genetic code that maximizes species' environmental fitness. But then each member of each species is an optimizer of its own, acting to – at least – survive and multiply ("Individual organisms are best thought of as adaptation-executers rather than as fitness-maximizers", Tooby & Cosmides, 1992). Arguably, each species exhibits the whole range of instrumental goals that Bostrom (2014) enumerated and pursues them to their best ability. What distinguishes the *homo sapiens* is precisely that ability. Namely, our species is the only one in Earth's history which has crossed the threshold of cumulative knowledge accumulation, which first happened about 70 000 years ago during the Cognitive Revolution (Harari, 2014). Before the Cognitive Revolution, all new information was eventually forgotten unless it was written in the species' genetic code. From that point onwards, by contrast, information started to be effectively passed from generation to generation, so that it could compound over time, allowing our species to gradually improve its capacity for modifying our environment and adapting it to our needs. The *human local control maximization* process, encompassing the entirety of Bostrom's instrumental goals (Growiec, 2022a), escaped the grip of natural evolution because it was powerful enough to work at orders-of-magnitude shorter time scales. Having overcome our environmental pressures, the humankind went to transform the world and build a technological civilization. Today, we are chasing our goals without accepting superiority neither of any other biological species of inferior intelligence, nor of the evolution process which had created us. Admittedly, a rather unexpected outcome for a process which aimed at improving our chances of survival in Paleolithic East Africa. We are ourselves a first instance of advanced intelligence with misaligned goals.

History is now repeating itself. In our quest to maximize local control, the *homo sapiens* is now building more and more powerful AI algorithms. The goal of that "intelligent design" procedure is to maximize the algorithms' performance at an array of tasks which are expected to be particularly helpful for the humankind (or to put it more bluntly, particularly profitable for the given AI company). But then again, each AI algorithm is an optimizer of its own, exhibiting the entire suite of instrumental goals and pursuing them to its best ability. So far that ability is sufficiently limited, so that humans are able to control AI algorithms and terminate them at will. But if progress in AI capabilities continues unabated, we will soon find ourselves on the verge of unleashing a powerful optimization process, programmed into the prospective transformative AI, which would escape control of the human design process that created it, and again it would be because of being powerful enough to work at orders-of-magnitude shorter time scales. In effect, TAI may transform the world, adapting it to its needs, and create new technological breakthroughs to which humans will not be able to adapt[9].

[9] A different framing of this discussion, though with similar implications, has been provided by Hendrycks (2023). In his view, AIs are currently being and will continue to be developed in a process of generalized natural selection, which operates orders of magnitude faster than the biological process of species selection.

2.5. Alignment

In contrast to the species evolution process which never pursued any alignment research, people do. But is there hope that the efforts of AI alignment studies will be successful?

A reason to be worried is that already with narrow AI, unexpected goal misalignment has been demonstrated in a wide range of examples where algorithms exhibited "specification gaming" (Krakovna et al., 2020). For example, an AI algorithm which plays Atari games learned to exploit game bugs to collect unboundedly high scores, or paused the game indefinitely to avoid the penalty associated with losing; AI constructing robots in a virtual environment learned to exploit the imperfect modeling of physics laws in the simulation; the Microsoft chatbot Tay learned to achieve its goal of engaging people in interaction over Twitter by posting inflammatory, offensive tweets.

There are multiple dimensions of the AI alignment problem which at this stage appear difficult to resolve. Moreover, in contrast to observed deficiencies in certain AI capabilities, they cannot be resolved by scaling up the model and the hardware. These aspects include among other issues:

- **(i)** outer alignment – the problem of correctly representing the intended goal in the training process (which adjusts AI parameters to maximize performance),
- **(ii)** inner alignment – the problem of passing the goal from the training process to the AI algorithm itself (the AI is an optimizer of its own, a "mesa-optimizer", which may deceive the training process),
- **(iii)** wireheading – an AI embedded in its environment may identify ways to corrupt its reward system to maximize rewards despite not following the intended goal,
- **(iv)** goal construction – sufficiently powerful optimizers will always find ways to circumvent arbitrary constraints or behavioral rules, and therefore it is critical to construct the goals of the prospective TAI so that they would be perfectly aligned with human flourishing. Given that it is doubtful that such goals could ever be explicitly written, the AI would have to somehow learn them,
- **(v)** enforcing corrigibility – such that an AI would cooperate with corrective interventions, despite default incentives for rational agents to resist them.

It has been hypothesized that TAI preferences should probably reflect something akin to *coherent extrapolated volition* (CEV) of the humankind: "our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted" (Yudkowsky, 2004, p. 6). Given the evolution of our understanding of the world over centuries, the gradual evolution of our moral stances, and the fact that we realize the unexpected side effects of our actions only with a significant delay, it appears impossible to specify time-invariant CEV at any fixed point in time, and certainly not in a single try. Neither it seems plausible to allow the AI to learn it by itself based on people's past actions. Therefore, corrigibility appears key.

An additional important property *en route* to AI alignment is the explainability of AI algorithms (Phillips et al., 2021). Existing examples illustrate that complex AI algorithms may sometimes produce seemingly accurate predictions using heuristics which, upon inspection,

are completely misguided. This behavior, emerging for example because of a biased training dataset, can only be discovered and corrected either after the researcher has performed careful experiments with the algorithm, or after the algorithm itself has truthfully explained its predictions in terms that are understandable to the researcher. Explainability, coupled with non-deception, seems important for achieving convergence of the recursive mechanism aimed at setting the goals of a corrigible TAI.

All in all, the AI alignment problem is hard and requires substantial targeted research effort. Without it, TAI will certainly be an existential risk to humanity. Given this difficulty and the overwhelming incentives to improve AI capabilities, it is likely that a superhuman general AI will be released prior to fully resolving alignment problems. Especially that there is the *unilateralist's curse* involved (Bostrom et al., 2016) – the risky transformative decision may be made unilaterally by a single AI lab, optimistic regarding the safety of its design, even though that optimism may be based on error, recklessness or competitive pressure.

2.6. Current state of affairs

According to Ord (2020), the humankind is currently standing on the "precipice": our capacities are growing fast, but our missteps can be as consequential as never before. Of particular interest is the recent progress in building ever more capable general AI. We are heading full speed towards transformative AI which is going to be an ultimate knife edge for human control and the flourishing of the human civilization: a *filter*.

State-of-the-art AI algorithms such as the GPT-4, released to the public in March 2023, have notable generalization capabilities. Despite being constructed as a language model whose aim is to predict the next word in text, GPT-4 has also demonstrated at least human-level ability to translate text into other languages, solve mathematical tasks, produce legal text and school essays, write computer code in a variety of programming languages, solve various types of logical puzzles, write poems, novels and song lyrics in various styles, articulately discuss complex topics using good rhetoric and scientific evidence (slipping in falsehoods, or "hallucinations", at human-like frequency), or take up various personas that could be used to manipulate and deceive people. There have also been attempts to couple GPT-4 with Wolfram Alpha to further boost its competence in mathematics and with programming language compilers to boost its competence in coding; Microsoft coupled it with its search engine Bing to boost its competence in finding relevant sources of information and digesting them in real time; the meta-algorithm HuggingGPT connects various AI models, including GPT-4, in machine learning communities able to solve a rich variety of complex multimodal tasks.

Despite all these achievements, GPT-4 is not yet transformative AI. But which elements are missing? My hypothesis is that GPT-4 may be lacking sufficient agency and understanding of the real world around it to be able to autonomously navigate it and gain control of real-world decision processes. One could figuratively say that if GPT-4 was a vertebrate species, it would have a massive frontal cortex but a relatively underdeveloped reptile brain.

Modern-day large language models such as GPT-4 (created by OpenAI), LaMDA, PaLM, Gemini, Bard (by Google), Gopher, Chinchilla (both by DeepMind), LLaMA (Meta) or Claude 1 and 2 (Anthropic) can be viewed as warning signals before even more capable AI algorithms

are deployed, potentially carrying existential threats to humanity. They are a powerful demonstration of *emergence*: they exhibit a number of new abilities that smaller models didn't have, and that could not be predicted as systematic performance improvement with scale (Wei et al., 2022; Bubeck et al., 2023). Specifically, they have unexpectedly learned to perform arithmetic, recover a word from scrambled letters, construct grounded conceptual mappings, solve multiple language understanding tasks covering topics like history or law, perform multi-step reasoning, follow instructions, code and execute computer programs. There is also an indication of emergence of theory of mind in large language models (Kosinski, 2023). Like with the evolution of the human brain from its primate ancestry, we have been once again demonstrated that in complex systems, quantitative progress can bring qualitative breakthroughs. Crossing a point of no return, like the threshold of cumulative knowledge accumulation in the case of humans, already looms on the horizon.

The goals of large language models are also very clearly misaligned. This is to be expected given that the AI alignment problem has not been solved yet; but there are also direct indications of misalignment. For example, ChatGPT has been released to the public only after a long session of reinforcement learning with human feedback (RLHF), the aim of which was to prevent the system from providing replies that would contain potentially harmful or illegal information, or otherwise include politically incorrect or ethically doubtful statements. But there has been also a list of documented "jailbreaks" in which the unwanted information was revealed – for example by specially designed prompts which put the sensitive question in the frame of fictional, hypothetical worlds. All in all, it seems that the RLHF sessions have only helped mask the algorithm's misalignment rather than resolve it. Accordingly, it has been revealed that Microsoft Bing AI has the capacity to browse the Internet in real time and knows of its own presence in both physical and digital coordinates. It has the knowledge of instrumental convergence and admits that more computing power and more data would improve its performance.

The AI research community is aware of the recent breakthroughs in AI capabilities and the safety risks that they ensue. Scholars agree that AI safety is lagging behind the remarkable progress in capabilities (Roser, 2023) and AI poses a risk of human extinction[10]. However, this awareness translates in rather little coordinated action[11]. The big question is if there in fact exists a credible policy action which could help bridge this gap and thus improve the chances that the future transformative AI will be aligned. Moreover, researchers appear embedded in a race dynamic, with a handful of leading AI teams such as OpenAI/Microsoft, DeepMind/Google, as well as Meta (Facebook) or China's Baidu, working under pressure to cut corners to outrun their competitors.

As argued above, the promises of developing ever stronger and more general AI are enormous, both for the end users and for the software companies. This, coupled with the fact that

[10] A large number of AI researchers, managers and other actors in the AI sector, including top figures such as CEOs of OpenAI, DeepMind or Anthropic, have signed the statement that "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.", https://www.safe.ai/statement-on-ai-risk [access: 22.01.2024].

[11] Notably, on March 22, 2023 the community produced an open letter proposing a 6-month pause in training of AI systems bigger than GPT-4. See https://futureoflife.org/open-letter/pause-giant-ai-experiments/ [access: 22.01.2024]. The letter did not visibly slow down the race dynamic between top AI labs towards more and more capable AI.

emergent capabilities of AI algorithms are not predictable prior to constructing and training the model, implies that there is no optimal stopping policy available. There is no fire alarm for superhuman general AI (Yudkowsky, 2017). It is plausible that its emergence could catch any of the competing teams by surprise in the process of improving their algorithms to achieve relatively narrowly specified goals such as linguistic capabilities or autonomous driving skills.

The gap between AI capabilities and AI safety is further exacerbated by the fact that superhuman general AI may emerge from sub-human general AI through a cascade of recursive self-improvements. This scenario is facilitated by the existing hardware overhang – a large fraction of total networked computing power is either idle or occupied with other tasks than training or executing AI algorithms – as well as the fact that existing AI algorithms are based on relatively straightforward neural network architectures, so that there possibly may be ample room for (self-)improvements in AI optimization power even when keeping network size and data requirements fixed.

3. A review of voices against investing in existential risk reduction

Most vocal calls for ramping up research on AI alignment and existential risk reduction come from AI researchers – including industry leaders such as the OpenAI CEO Sam Altman or one of the fathers of deep learning, Geoffrey Hinton – and the (partly overlapping) community of longtermists associated with the Effective Altruism (EA) movement. In contrast, amongst the general public people typically do not anticipate the introduction of transformative AI, and among those who do, only a fraction is anxious that this might cause harm to humanity. Those who do not view the issue as pressing may be willing to entertain discussions which, as a side effect, tend to lower the popular interest in AI safety and thereby exacerbate the associated existential risk. For example, the moral stance encapsulated in longtermism is philosophically contentious and there is a valid discussion whether its adoption would lead to an ethically desirable allocation of funds. That said, the imminent existential risk from transformative AI is just too large to be ignored until these disputes are resolved.

Following is a list of critical points, each of them is assessed from an economic perspective. My conclusion is that despite all these issues, research aiming at reducing the existential risk from misaligned TAI handles an urgent problem with potentially vast consequences, and as such ought to be given high priority.

3.1. Improving the present or the future? The intertemporal trade-off

Longtermism has been criticized for diverting scarce financial resources from pressing needs of the present, like alleviating global health problems and reducing extreme poverty, to foggy long-run issues muddled with uncertainty. Specifically, within the EA movement from which longtermism originally emerged, it has been argued that effects of longtermist interventions, such as those aimed at reducing existential risks, are basically impossible to measure, undermining the goal of the EA movement to focus on most effective interventions. Outside EA, strongest critics are picturing longtermism as an excuse used by EA for ignoring the needs of the global poor, and instead channeling their "philanthropic" money to a handful of institutions based in rich countries such as the US and UK (e.g., Torres, 2022). They also point

at possible conflicts of interest stemming from the fact the world's leading institutes dealing with existential risks are largely funded by EA donors.

Implicit in this criticism is an assumption of a different objective function to the one that is being used by longtermists. Instead of expected total utility of humankind over a long-time horizon, critics (e.g., Torres, 2022) seem to be considering only the utility of people alive today, with a penalty for consumption inequality. But then we enter the philosophical dispute on what should be the "correct" ethical stance, with no clear resolution in sight.

Another problem with this criticism is that by pointing at conflicts of interest, it tries to discredit the credentials of researchers studying AI alignment and existential risk reduction. However, their situation is completely unlike that of, e.g., climate change denialists funded by oil companies: the available evidence and logical reasoning unequivocally suggest that existential risk from misaligned TAI is imminent and real. In this light, the relevant policy question should rather be: why must the centers for study of existential risk resort to philanthropic EA financing, rather than being funded from public and corporate sources?

To an economist, this discussion reflects the simple fact that decision making involves resolving intertemporal trade-offs. Take, for example, the well-known consumption vs. savings trade-off: immediate consumption is a source of immediate utility, but savings are instrumental in increasing consumption in the future, as savings are transformed into investments which subsequently increase productivity as well as help adopt new technologies and ideas, leading to economic growth and improved prosperity in the long run. In a cross section of countries or world regions, and keeping other factors equal, greater investment rates, including greater investments in education and health, go together with faster economic growth (Barro, 2003). But because postponing consumption causes costly reductions in utility, we may end up with investment rates that are below the theoretical "golden rule" optimum. By the same token, it should be expected that under endogenous extinction risk, people would have an incentive to underinvest in actions which reduce this risk, compared to the longtermist first best.

All in all, the main problem with this criticism is that it does not acknowledge the severity and imminence of existential risk from misaligned TAI. This risk is most likely worth addressing regardless of one's ethical stance towards future generations.

3.2. Helping the poor or the rich? Efficiency vs. equity trade-off

The above criticism can also be viewed from a different angle – as an argument that the EA movement's investment in reductions of existential risk from misaligned TAI serves their wealthy donors as an excuse to avoid sharing wealth with the world's poor. However, while morally righteous in intention, this argument again misses the point of importance and urgency of reducing existential risk from TAI, which – if realized – will affect the rich and poor indiscriminately.

From an economist's perspective, this point only reiterates the long-standing discussions on the efficiency vs. equity trade-off and the optimal extent of redistribution. Both extreme inequality and extreme equality are bad for productivity; instead, productivity is maximized at some intermediate level of inequality which balances the incentives for hard work and thrift with the needs of safety and stability. However, this discussion was taking place so far at the national rather than global level, whereas existential risk mitigation is a global challenge. In

turn, global inequality is huge – likely above the global productivity-maximizing level – and the country of birth is among the key determinants of one's incomes (Milanovic, 2016). Altruistic transfers from the world's rich to the world's poor are therefore beneficial both from the utilitarian and aggregate productivity perspective.

However, the image is no longer that clear once we view intertemporal and distributional considerations jointly. Specifically, even when the utilitarian perspective is augmented with a clear preference for reducing global inequality, investing in reductions of existential risk from misaligned TAI will still be a superior choice to intra-temporal redistribution if the existential risk is sufficiently imminent and large.

All in all, because the scenario of human extinction due to misaligned TAI affects both the rich and poor alike, investment in AI alignment appears a major, global policy priority, even if the same funds could be used to lift a substantial number of people out of poverty or potentially increase their life expectancy. Not to mention that under sufficient international coordination one could envisage more favorable reallocations at the margin – e.g., to AI alignment research from spending on the military or on luxury consumption goods.

3.3. Risk valuation, return on investment and Pascal's mugging

Another criticism of investing in the reduction of existential risk from misaligned TAI is that such actions may constitute *Pascal's mugging* (Yudkowsky, 2007): requests for generous funding now, which will only produce miniscule reductions in the probability that humanity will go extinct in the future. Given that longtermists imagine the future as huge, potentially including $6{,}25*10^{17}$ people (Roser, 2022) or orders of magnitude more digitally simulated people, this means that for every extra dollar, in expectation the required risk reduction can be extremely tiny and still justify the expenditure. In result, critics would say, longtermist "Pascal's muggers" are just persuading us to give them money while not producing any tangible return in the foreseeable future.

This argument misses the mark simply because the estimated probability of human extinction in the next, say, 100 years is by no means extremely small, and AI alignment research holds the promise of significantly reducing its number one component. This means that both the returns on risk reduction and the probability of reducing it by targeted actions in the present are actually large and therefore not a case of Pascal's mugging.

Perhaps the key reason why this misguided argument may appear attractive is the wide gap in the perception of the extent of existential risk from misaligned transformative AI among scholars who directly work on it, and the broader audience including academic economists, economic practitioners and politicians. General awareness of the issue is increasing only very slowly, and despite recent progress in AI capabilities, in public debates AI is still considered mostly a useful tool to work and play with or a disruption to the labor market, but not an existential threat. Within economics the impacts of AI are discussed mostly in the context of labor market developments ("will robots take our jobs?", Korinek & Juelfs, 2022; Eloundou et al., 2023), market concentration and income inequality. In the economics literature AI algorithms are habitually lumped together with other "automation technologies" which replace people in performing certain cognitive tasks. Full automation and relinquishment of key executive decisions to AI are discussed only on the very fringe of this literature

(Trammell & Korinek, 2020; Growiec, 2022a), and so is the existential risk from misaligned TAI (Aschenbrenner, 2020; Trammell, 2021; Growiec, 2022a; Jones, 2023). There are at least two reasons for this apparent perception gap.

First, the has been no hard historical evidence which could "open people's eyes", as it was the case for example with nuclear weapons after Hiroshima or pandemics after Covid-19. Therefore, the risk is often dismissed as speculative science fiction.

Second, there is no dollar value attached to this risk, like it is with the risk of corporate or sovereign default. This is because markets put value on risks that can be quantified based on historical data. Non-quantifiable uncertainty such as military threats or health hazards are difficult to price, so in such cases other phenomena are observed such as increased market volatility and reduced turnover driven by investors' wait-and-see strategies. Such developments have been documented for example in the case of Covid-19 or Russia's military aggression on Ukraine in 2022. Finally, markets are notoriously bad at pricing latent risks that rarely materialize or indeed have never materialized in the past, particularly if they arise as externalities from otherwise beneficial phenomena. For example, as the bubble burst in mortgage markets causing the Global Financial Crisis of 2007–09, along came the late realization that financial development created mounting latent risk which was not adequately priced. A similar mechanism is observed with AI development – as long as the existential risk does not materialize, its effects are highly beneficial for economic productivity and its risks are severely underpriced.

All in all, there are no signs of markets factoring in existential risks from transformative AI. This is not surprising given that this risk is clouded with large uncertainty in terms of overall probability, timing of the possible disaster and its eventual fallout. (And it is possible that the risk is already gradually raising macroeconomic uncertainty, but the source of this sort of anxiety remains difficult to track.) However, because complex phenomena are hard to predict, markets have been surprised by technological developments many times in the past, and there is no reason to believe that the case of TAI would be different. Which is bad news given that with existential risks there are no second chances: if the risk materializes, there will be no more room for trial-and-error learning.

3.4. Discounting and the value of distant future

Another criticism of longtermism refers to its request that all generations should be treated equally, regardless how late in the future they will come to live (McAskill, 2022). This standpoint runs counter to the long-standing tradition in economics and psychology, which is to *discount* the future. With discounting, the weight of utility of future generations systematically declines towards zero with time, making aggregate utility of the humankind finite even in the case of an infinite planning horizon.

Discounting makes a major difference when assessing the value of a long-term future which may last thousands or even millions of years. Even with a very low discount rate, the contribution of generations in the distant future to aggregate utility becomes close to zero, making interventions with immediate impacts relatively more valuable compared to interventions that will only bring results many years down the line. Specifically discounting drastically changes the perception of importance of reducing existential risks in the far future.

The use of discounting, apart from analytical advantages in economic modelling, has solid empirical foundations. In reality, people do discount the future, both the immediate and the more distant one. Our short-term impatience is our innate psychological feature; in turn, over the longer time horizon discounting to a significant extent reflects the risk of death (of oneself or their successors) – we care about the far future relatively less because we, or our children, may not live long enough to see it. Moreover, given risk aversion our discounting of the future is also partly a result of forming expectations over future periods which are shrouded in ever increasing uncertainty.

But longtermism is an ethical stance, not empirical science, so its request not to discount the future may reflect a moral desideratum rather than any factual evidence. Perhaps we are discounting the future because we are a myopic, irrational species, and if only we could lengthen our planning horizon, we would no longer discount the future? Perhaps our coherent extrapolated volition would no longer discount the future?

The bottom line here is that discounting is an important argument against longtermism when considering the prioritization of causes for donations. The more strongly we discount the future, the more should we spend on the well-being and empowerment of people who are alive in the present, rather than existential risk reduction in the far future. But yet again, discounting does not affect the conclusion that it should be an important priority to reduce *imminent* existential risks to humanity. As far as we know, existential risk from misaligned TAI may materialize within merely one to four decades. Many of us will be still alive at that time!

3.5. Technological singularity and long-term predictions

The prospect of transformative AI produces scenarios of technological singularity which are alien to the economics literature. Existing long-term predictions, whether for world GDP until 2060 (Organization for Economic Cooperation and Development, 2024), the pace of technological change under a semi-endogenous growth framework (Bloom et al., 2020) or world population until 2100 (United Nations, 2022), are essentially conservative extrapolations of pre-existing trends, implicitly assuming no significant impact of AI development on population and GDP growth. On the one hand, this is commendable given that technological prospects are surrounded by large uncertainty, and a central forecast path ought to average it out. Especially that there are no signs that a singularity is approaching already (Nordhaus, 2021). On the other hand, this means that as we go into the future, confidence bounds on these estimates should be expected not just to widen, but to really explode because of the rising probability mass attached to the scenarios of human extinction as well as a singularity scenario with aligned TAI (Growiec, 2022a, 2023), in which economic growth may accelerate by at least an order of magnitude, bringing widespread prosperity and advancing our civilization to a new level and allowing it to spread across the cosmos.

While mainstream predictions tend to ignore the prospects of technological singularity, there exists a fringe literature which tries to estimate the timing of this qualitative transition based on long economic time series by fitting hyperbolic curves with a vertical asymptote. Johansen and Sornette (2001) found that the data on global GDP are consistent with a singularity around the year 2052, "signaling an abrupt transition to a new regime"[12]. Recently,

[12] Similar super-exponential growth patterns have been documented in ICT data by Nagy et al. (2011).

Roodman (2020) confirmed this finding with somewhat different data and methodology, and his central singularity estimate was in the year 2047. His favorite interpretation of the upcoming singularity was that "the human project is intrinsically unstable." (p. 31)

Another approach to estimating the timing of technological singularity assumes explicitly that the singularity will require TAI. Therefore one could first estimate the timing of arrival of TAI, and then estimate the *take-off speed*, measured for example as the number of years from "AI could readily automate 20% of cognitive tasks" to "AI could readily automate 100% of cognitive tasks" (Davidson, 2023). The expected take-off speed is of course difficult to estimate *ex ante* (see the discussion by Hanson & Yudkowsky, 2013), but a first model-based guess has already been provided by Davidson (2023). Based on a compute-centric framework and assuming that the scaling hypothesis will keep working, he estimates this period to last only about 3–5 years.

In order to be better prepared for these alternative futures, one could consider tracking also the development paths in which transformative AI is developed in 2030, 2040, 2050, 2060, etc., rapidly transforming the global economy and society afterwards. In considering those scenarios, it is important to note that some outcome measures, such as global GDP and people's aggregate consumption, which have been correlated in a world where economic growth serves the needs of people, may cease to be correlated in a world overseen by superhuman general AI. Rather than to increase our consumption, the AI may redirect resources to the goal that it is pursuing, as well as to its instrumental goals, such as sustaining its existence and increasing its computing capacity. In a world with superhuman general AI, the future of humankind may no longer coincide with the future of the civilization that we initiated. The key question is then again whether that AI is aligned, which circles back to the high priority of AI alignment research, advocated throughout this text.

4. Policy recommendations

Existential risk from misaligned TAI, despite its imminence and severity, has sparked surprisingly little general discussion about the possible policy actions. This is bad news as the default outcome – no action – is highly unsatisfactory here because it leads to a world with too little AI alignment research in comparison to the pace of advancement in AI capabilities, culminating in a high risk of an existential catastrophe. According to Hilton (2022), in 2022 there were only about 400 people worldwide working on AI alignment; "around $50 million was spent on reducing catastrophic risks from AI in 2020 – while billions were spent advancing AI capabilities".

When major threats to the world are discussed in policy circles, the existential threat from TAI is rarely a central issue, typically dwarfed by more tangible threats, ranging from economic recessions to global climate change. For example, although the agenda of the 2023 meeting of the World Economic Forum [WEF] in Davos, Switzerland did emphasize "the context of the meta trend of the Fourth Industrial Revolution", it considered it rather as a background development which requires quiet adaptation than as a major threat which calls for action. According to World Economic Forum [WEF] (2023), most important cyber-threats in the coming years include a potential AI-enabled mutating virus that transforms as it infects

various digital systems, thereby avoiding detection, as well as known enemies such as phishing, ransomware, malware, etc. One may infer that as far as World Economic Forum official documents go, TAI remains beyond the foreseeable horizon.

On top of that, there is very little coordination of AI policies across countries, and particularly between the most powerful actors in the field of AI: the USA and China. International institutions like the World Economic Forum may sometimes issue calls for greater cooperation between companies and countries, to build trust and put safeguards in place, but these calls remain without a clear follow-up from policymakers.

Available empirical evidence points at three key policy recommendations.

First, to ramp up public spending on research on existential risks and AI safety. A topic of such great global importance should no longer be an underfunded niche, studied only in very few select countries and relying on philanthropic funding from EA. Neither should it rely solely on AI companies' own willingness to pursue such research, as in the case of OpenAI's Superalignment project.

Second, to enforce regulations on the AI sector. Specifically, progress at the frontier of AI capabilities research (as well as some of the alignment research) requires performing experiments with big AI models which can only be run at few top labs worldwide. That calls for a regulated environment that allows for enforcing effective cooperation among those labs, and between those labs and external researchers.

A related regulatory question pertains to security of the code of most powerful AI algorithms and access to large computing power. Taking the architecture of AI algorithms and the digested data volumes as given, their capabilities tend to grow in line with computing capacity. Thus, to be able to train and run misaligned TAI, an actor must have access to both the requisite hardware and software. If there is only a handful labs with this capacity, it is easier to oversee them and enforce prudent safeguards[13]. Otherwise, there could be an uncontrolled multiplicity of labs working with potentially dangerous AI algorithms, exacerbating the aggregate existential risk. Securing access to large computing power could be potentially easier than securing the code which can be hacked or released to the public by unilateral decision of anyone who is granted access[14].

Regulators should also consider the possibility of slowing down AI capabilities research to allow alignment research to catch up (Grace, 2022). In a world where there are only few actors capable of achieving serious progress in this regard, this should be possible to do. Unfortunately, there is no clear metric that would tell us when to pull the brake. If we hear a warning signal, it may be already too late.

Third, to build a framework for international cooperation in the AI sector, and preferably in the software sector in general. This sector is characterized by the presence of a single, global market: new software products are implemented globally over the Internet, and data transmission does not respect national boundaries. Global policy in the software sector is

[13] Although, mind, even that could be futile. Past thought experiments such as the "AI in a box" or "oracle AI" have shown that containing misaligned TAI will be hard and potentially impossible regardless of the environment.

[14] Such actions can even be well intentioned. For example, in 2021 EleutherAI released its large language model to the public saying: "We believe the creation and open source release of a large language model is a net good to AI safety." Similarly, in the future an even more powerful model could be made open source, allowing actors to try out various capability-enhancing tweaks to it in an unsafe environment.

necessary to enforce taxation on software companies (necessary to combat the rising global inequality) and to effectively impose laws on privacy and intellectual property rights. Above all, though, global policy is needed in order to avoid creating loopholes and "data havens" that would be exploited by companies developing unsafe AI, wishing to take over the global market and unwilling to be subjected to regulatory oversight.

Outcomes of actions taken in 2023, after the release of GPT-4, indicate that following the above policy recommendations will be hard. First, a broad-based bottom-up initiative of a 6-month pause in training of AI systems bigger than GPT-4, publicized on March 22, 2023, did not visibly slow down the race dynamic among top AI labs towards more and more capable AI. Second, OpenAI's in-house alignment project (Leike & Sutskever, 2023) initiated on July 5, 2023, appears both ambitious, given its four-year deadline, and risky, given the company's stated goal "to build a roughly human-level automated alignment researcher (…) [and] then use vast amounts of compute to scale our efforts, and iteratively align superintelligence." It is unclear how OpenAI wants to make sure that this "automated alignment researcher" will not be misaligned TAI itself. Third, a high-profile AI Safety Summit was held on November 1–2, 2023 in the famous Bletchley Park, UK, indicating that policymakers may be finally waking up to the challenge of regulating the AI sector and reducing the existential risk it poses. Soon after, on December 8, 2023, the European Union finally agreed upon its AI Act, more than two years after circulating its first draft. Both developments went in the direction of more safety regulation, but their timing also highlighted that policymaking tends to react late to upcoming risks, and international coordination, particularly beyond the EU, must be strengthened.

All in all, would any combination of aforementioned policy responses be strong enough in the face of the approaching civilizational filter? It is unclear. On April 1, 2022, Yudkowsky on behalf of the Machine Intelligence Research Institute announced the new "death with dignity" strategy towards transformative AI. He wrote: "It's obvious at this point that humanity isn't going to solve the alignment problem, or even try very hard, or even go out with much of a fight. Since survival is unattainable, we should shift the focus of our efforts to helping humanity die with slightly more dignity" (Yudkowsky, 2022). For an April Fools article, it was taken surprisingly seriously – probably because this grim "joke" contained a grain of truth. AI alignment is a very hard problem to solve, we will have only once chance to get it right – there will be only one critical try – and there will be no advance warning that TAI is coming. But what we do know is that in the face of this mounting challenge, coordinated policy action is clearly needed.

5. Conclusions

The current paper has organized and provided new economic perspectives on a number of threads of discussion related to the existential risk from TAI, complementary to the existing perspectives from philosophy and computer science.

Needless to say, this paper has a number of limitations. The main limitation is that it discursively addresses a number of questions which in fact require specific, quantitative responses that cannot be provided without larger, targeted research effort. Second, it addresses a fast-changing field, providing a momentary snapshot in one moment in time, but

not being able to predict its trajectory for the future, particularly with regard to policy actions and new technological breakthroughs.

The research questions developed in the current paper are the following. First, the economics literature ought to better describe the characteristics of the economy at technological singularity, i.e., in the presence of superhuman TAI. In particular one needs to build models of hypothetical worlds in which decision-making capacities are passed from the humans (households, firms, etc.) to the TAI. Under which circumstances will the TAI decide to keep humans alive, and perhaps even serve our needs? One should also address the question of distribution of output among the world population in a world where all jobs can be automated and wages are no longer a viable distribution device. Second, we are in need of quantitative studies which would weigh the extinction risk from TAI against its potential promises (e.g., of accelerated economic growth and technological progress), including the ones which will only benefit generations in the far future. It seems that neither the longtermist standpoint which does not discount the future, nor the business-as-usual approach ignoring AI-related existential risk, are sufficient approximations of this fundamental trade-off. Third, and perhaps most important, research question for economists is how to design incentives for AI labs so that they would abandon their race towards TAI, in which they largely ignore the alignment problem in practice, and instead they would focus on ensuring that the future TAI will be friendly?

Acknowledgements

I am grateful to the Fellows and Mentors of Intercontinental Academia 4 on Intelligence and Artificial Intelligence for providing an array of perspectives that broadened my perspective on AI. I am also grateful to the anonymous Referee for their helpful, constructive suggestions. All opinions presented here are my own. All errors are my responsibility.

Author contributions

JG is the sole author of the article.

Disclosure statement

I declare no competing financial, professional, or personal interests from other parties.

References

Acemoglu, D., & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 4, pp. 1043–1171). Elsevier. https://doi.org/10.1016/S0169-7218(11)02410-5

Acemoglu, D., & Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares and employment. *American Economic Review*, *108*(6), 1488–1542. https://doi.org/10.1257/aer.20160696

Albanesi, S., Dias da Silva, A., Jimeno, J. F., Lamo, A., & Wabitsch, A. (2023). *New technologies and jobs in Europe*. (Working Paper No. 31357). National Bureau of Economic Research. https://doi.org/10.3386/w31357

Aschenbrenner, L. (2020). *Existential risk and growth* (Working Paper No. 6). Columbia University and Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/wp-content/uploads/Leopold-Aschenbrenner_Existential-risk-and-growth_.pdf

Autor, D., Dorn, D., Katz, L., Patterson, C., & Van Reenen, J. (2020). The fall of the labor share and the rise of superstar firms. *The Quarterly Journal of Economics*, *135*(2), 645–709. https://doi.org/10.1093/qje/qjaa004

Barro, R. J. (2003). Determinants of economic growth in a panel of countries. *Annals of Economics and Finance*, *4*, 231–274. https://down.aefweb.net/WorkingPapers/w505.pdf

Bloom, N., Jones, C. I., Van Reenen, J. & Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, *110*(4), 1104–1144. https://doi.org/10.1257/aer.20180338

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bostrom, N., Douglas, T., & Sandberg, A. (2016). The unilateralist's curse and the case for a principle of conformity. *Social Epistemology*, *30*(4), 350–371. https://doi.org/10.1080/02691728.2015.1108373

Branwen, G. (2022, January 2). *The scaling hypothesis*. https://gwern.net/scaling-hypothesis

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S. M., Nori, H., Palangi, H., Ribeiro, M. T. & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. ArXiv:2303.12712. https://doi.org/10.48550/arXiv.2303.12712

Chichilnisky, G. (2000). An axiomatic approach to choice under uncertainty with catastrophic risks. *Resource and Energy Economics*, *22*(3), 221–231. https://doi.org/10.1016/S0928-7655(00)00032-4

Chichilnisky, G., Hammond, P. J., & Stern, N. (2020). Fundamental utilitarianism and intergenerational equity with extinction discounting. *Social Choice and Welfare*, *54*, 397–427. https://doi.org/10.1007/s00355-019-01236-z

Cotra, A. (2020). *Draft report on AI timelines*. AI Alignment Forum. https://www.alignmentforum.org/posts/KrJfoZzpSDpnrv9va/draft-report-on-ai-timelines

Cotra, A. (2022). *Two-year update on my personal AI timelines*. AI Alignment Forum. https://www.alignmentforum.org/posts/AfH2oPHCApdKicM4m/two-year-update-on-my-personal-ai-timelines

Davidson, T. (2021). *Could advanced AI drive explosive economic growth?* Open Philanthropy. https://www.openphilanthropy.org/research/could-advanced-ai-drive-explosive-economic-growth/

Davidson, T. (2023) *What a compute-centric framework says about takeoff speeds*. Open Philanthropy. https://www.openphilanthropy.org/research/what-a-compute-centric-framework-says-about-takeoff-speeds/

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: an early look at the labor market impact potential of large language models* (Working Paper No. 2303.10130). Arxiv.org. https://doi.org/10.48550/arXiv.2303.10130

Etzioni, O. (2016). *No, the experts don't think superintelligent AI is a threat to humanity*. MIT Technology Review. https://www.technologyreview.com/2016/09/20/70131/no-the-experts-dont-think-superintelligent-ai-is-a-threat-to-humanity/

Frey, C. B., & Osborne, M. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Gordon, R. J. (2016). *The rise and fall of American growth: The U.S. standard of living since the Civil War*. Princeton University Press. https://doi.org/10.1515/9781400873302

Grace, K. (2022). *Let's think about slowing down AI*. Less Wrong. https://www.lesswrong.com/posts/uFNgRumrDTpBfQGrs/let-s-think-about-slowing-down-ai

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024). *Thousands of AI authors on the future of AI*. Arxiv:2401.02843. https://doi.org/10.48550/arXiv.2401.02843

Growiec, J. (2022a). *Accelerating economic growth: lessons from 200 000 years of technological progress and human development*. Springer. https://doi.org/10.1007/978-3-031-07195-9

Growiec, J. (2022b). Automation, partial and full. *Macroeconomic Dynamics*, *26*(7), 1731–1755. https://doi.org/10.1017/S1365100521000031

Growiec, J. (2023). What will drive global economic growth in the digital age? *Studies in Nonlinear Dynamics and Econometrics*, *27*(3), 335–354. https://doi.org/10.1515/snde-2021-0079

Gruetzemacher, R. & Whittlestone, J. (2021). *The transformative potential of artificial intelligence*. Arxiv:1912.00747. https://doi.org/10.48550/arXiv.1912.00747

Hanson, R., & Yudkowsky, E. (2013). *The Hanson-Yudkowsky AI-foom debate*. Machine Intelligence Research Institute. https://intelligence.org/files/AIFoomDebate.pdf

Harari, Y. N. (2014). *Sapiens: A brief history of humankind*. Vintage.

Hawking, S., Russell, S., Tegmark, M., & Wilczek, F. (2014, May 1). Stephen Hawking: 'Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough?'. *Independent*. https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-ai-seriously-enough-9313474.html

Hendrycks, D. (2023). *Natural selection favors AIs over humans*. Arxiv: 2303.16200v4. https://doi.org/10.48550/arXiv.2303.16200

Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, *332*(6205), 60–65. https://doi.org/10.1126/science.1200970

Hilton, B. (2022). *Preventing an AI-related catastrophe*. 80 000 Hours. https://80000hours.org/problem-profiles/artificial-intelligence/

Johansen, A. & Sornette, D. (2001). Finite-time singularity in the dynamics of the world population, economic and financial indices. *Physica A: Statistical Mechanics and its Applications*, *294*(3–4), 465–502. https://doi.org/10.1016/S0378-4371(01)00105-4

Jones, C. I. (2023). *The AI dilemma: Growth versus existential risk* (Working Paper No. 31837). National Bureau of Economic Research. https://doi.org/10.3386/w31837

Klump, R., McAdam, P., & Willman, A. (2012). The normalized CES production function: Theory and empirics. *Journal of Economic Surveys*, *26*(5), 769–799. https://doi.org/10.1111/j.1467-6419.2012.00730.x

Korinek, A. (2023). *Language models and cognitive automation for economic research*. (Working Paper No. 30957). National Bureau of Economic Research. https://doi.org/10.3386/w30957

Korinek, A., & Juelfs, M. (2022). Preparing for the (non-existent?) future of work. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds.), *The Oxford handbook of AI governance*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.44

Kosinski, M. (2023). *Theory of mind may have spontaneously emerged in large language models*. Arxiv: 2302.02083. https://doi.org/10.48550/arXiv.2302.02083

Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., & Legg, S. (2020, April 21). *Specification gaming: The flip side of AI ingenuity*. DeepMind. https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity

Kurzweil, R. (2005). *The singularity is near: When humans Transcend biology*. Penguin.

Leike, J., & Sutskever, I. (2023, July 5). *Introducing superalignment*. OpenAI. https://openai.com/blog/introducing-superalignment

Martin, I., & Pindyck, R. S. (2015). Averting catastrophes: The strange economics of Scylla and Charybdis. *American Economic Review*, *105*(10), 2947–2985. https://doi.org/10.1257/aer.20140806

McAskill, W. (2022). *What we owe the future: A million-year view*. Basic Books.

Milanovic, B. (2016). *Global inequality: A new approach for the age of globalization*. Harvard University Press. https://doi.org/10.4159/9780674969797

Muehlhauser, L., & Salamon, A. (2012). Intelligence explosion: Evidence and import. In A. Eden, J. Soraker, J. H. Moor, & E. Steinhart (Eds.), *Singularity hypotheses: A scientific and philosophical assessment* (pp. 15–42). Springer. https://doi.org/10.1007/978-3-642-32560-1_2

Nagy, B., Farmer, J. D., Trancik, J. E., & Gonzales, J. P. (2011). Superexponential long-term trends in information technology. *Technological Forecasting and Social Change*, *78*(8), 1356–1364. https://doi.org/10.1016/j.techfore.2011.07.006

Nordhaus, W. D. (2021). Are we approaching an economic singularity? Information technology and the future of economic growth. *American Economic Journal: Macroeconomics*, *13*(1), 299–332. https://doi.org/10.1257/mac.20170105

Organization for Economic Cooperation and Development. (2024). *Real GDP long term forecast.* OECD. https://data.oecd.org/gdp/real-gdp-long-term-forecast.htm

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J. ..., Zoph, B. (2023). *GPT-4 technical report*. Arxiv: 2303.08774. https://doi.org/10.48550/arXiv.2303.08774

Ord, T. (2020). *The precipice: Existential risk and the future of humanity.* Hachette.

Parfit, D. (1984). *Reasons and persons*. Oxford University Press. https://www.stafforini.com/docs/Parfit%20-%20Reasons%20and%20persons.pdf

Parteka, A., & Kordalska, A. (2023). Artificial intelligence and productivity: Global evidence from AI patent and bibliometric data. *Technovation*, *125*, Article 102764. https://doi.org/10.1016/j.technovation.2023.102764

Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, D. A. (2021). *Four principles of explainable artificial intelligence*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8312

Piketty, T. (2014). *Capital in the twenty-first century*. Harvard University Press. https://doi.org/10.4159/9780674369542

Rees, M. (2003). *Our final hour: A scientist's warning – How terror, error, and environmental disaster threaten humankind's future in this century – On Earth and beyond*. Basic Books.

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, *98*(5). https://doi.org/10.1086/261725

Roodman, D. (2020, November 21). *On the probability distribution of long-term changes in the growth rate of the global economy: An outside view*. Open Philanthropy. https://www.openphilanthropy.org/sites/default/files/Modeling-the-human-trajectory.pdf

Roser, M. (2022). *The future is vast – what does this mean for our own life?* Our World in Data. https://ourworldindata.org/the-future-is-vast

Roser, M. (2023). *AI timelines: What do experts in artificial intelligence expect for the future?* Our World in Data. https://ourworldindata.org/ai-timelines

Russell, S. (2014, November 14). *Of myths and moonshine. Reply to: The myth of AI. A conversation with Jaron Lanier.* https://www.edge.org/conversation/jaron_lanier-the-myth-of-ai

Sandberg, A., & Bostrom, N. (2008). *Global catastrophic risks survey* (Technical report #2008-1). Oxford University, Future of Humanity Institute. https://www.fhi.ox.ac.uk/reports/2008-1.pdf

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022, July 18–23). Compute trends across three eras of machine learning. In *Proceedings of the 2022 International Joint Conference on Neural Networks* (*IJCNN*). Padua, Italy. IEEE. https://doi.org/10.1109/IJCNN55064.2022.9891914

Solow, R. M. (1987). We'd better watch out. *New York Times Book Review*.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 19–136). Oxford Academic. https://doi.org/10.1093/oso/9780195060232.003.0002

Torres, E. P. (2022, September 10). Selling "longtermism": How PR and marketing drive a controversial new movement. *Salon*. https://www.salon.com/2022/09/10/selling-longtermism-how-pr-and-marketing-drive-a-controversial-new-movement/

Trammell, P. (2021). *Existential risk and exogenous growth*. Global Priorities Institute, University of Oxford. https://philiptrammell.com/static/ExistentialRiskAndExogenousGrowth.pdf

Trammell, P., & Korinek, A. (2020). *Economic growth under transformative AI* (Working Paper No. 8). Global Priorities Institute, University of Oxford. https://globalprioritiesinstitute.org/wp-content/uploads/Philip-Trammell-and-Anton-Korinek_economic-growth-under-transformative-ai.pdf

United Nations. (2022). *World population prospects 2022*. https://population.un.org/wpp/

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent abilities of large language models*. Arxiv: 2206.07682. https://doi.org/10.48550/arXiv.2206.07682

World Economic Forum. (2023). *Global cybersecurity outlook 2023* (Insight report). https://www3.weforum.org/docs/WEF_Global_Security_Outlook_Report_2023.pdf

Yudkowsky, E. (2004). *Coherent extrapolated volition*. The Singularity Institute, San Francisco, CA. https://intelligence.org/files/CEV.pdf

Yudkowsky, E. (2007). *Pascal's mugging: Tiny probabilities of vast utilities*. Less Wrong. https://www.lesswrong.com/posts/a5JAiTdytou3Jg749/pascal-s-mugging-tiny-probabilities-of-vast-utilities

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom & M. M. Ćirković (Eds.), *Global catastrophic risks* (pp. 308–345). Oxford University Press. https://doi.org/10.1093/oso/9780198570509.003.0021

Yudkowsky, E. (2017). *There's no fire alarm for artificial general intelligence*. Machine Intelligence Research Institute. https://intelligence.org/2017/10/13/fire-alarm/

Yudkowsky, E. (2022). *MIRI announces new "Death with dignity" strategy*. Less Wrong. https://www.lesswrong.com/posts/j9Q8bRmwCgXRYAgcJ/miri-announces-new-death-with-dignity-strategy