Singularity and Coordination Problems: Pandemic Lessons from 2020

Preprint · October 2020

DOI: 10.13140/RG.2.2.24950.01609/3

2 authors:

Nicholas Kluge Corrêa
Pontifícia Universidade Católica do Rio Grande do Sul

Nythamar Fernandes De Oliveira
Pontifícia Universidade Católica do Rio Grande do Sul

Singularity and Coordination Problems: Pandemic Lessons from 2020[a]

Nicholas Kluge Corrêa[b]

Graduate Program in Philosophy of the Pontifical Catholic University of Rio Grande do Sul – Av. Ipiranga, 6681 - Partenon, Porto Alegre - RS, 90619-900.
nicholas.correa@acad.pucrs.br, https://orcid.org/0000-0002-5633-6094

Nythamar de Oliveira[c]

Graduate Program in Philosophy of the Pontifical Catholic University of Rio Grande do Sul – Av. Ipiranga, 6681 - Partenon, Porto Alegre - RS, 90619-900.
nythamar@yahoo.com, https://orcid.org/0000-0001-9241-1031

[a] Acknowledgements and Funding: The authors would like to thank the Academic Excellence Program (PROEX) of CAPES Foundation (Coordination for the Improvement of Higher Education Personnel) and the Graduate Program in Philosophy of the Pontifical Catholic University of Rio Grande do Sul, Brazil.

[b] Master in Electrical Engineering and Ph.D. student in Philosophy – PUCRS.

[c] Ph.D. in Philosophy (State University of New York, 1994), Full Professor – PUCRS.

Abstract

One of the strands of the Transhumanist movement, Singulitarianism, studies the possibility that high-level artificial intelligence may be created in the future, debating ways to ensure that the interaction between human society and advanced artificial intelligence can occur safely and beneficially. But how can we guarantee this safe interaction? Are there any indications that a Singularity may be on the horizon? In trying to answer these questions, We'll make a small introduction to the area of security research in artificial intelligence. We'll review some of the current paradigms in the development of autonomous intelligent systems and evidence that we can use to prospect the coming of a possible technological Singularity. Finally, we will present a reflection using the COVID-19 pandemic, something that showed that our biggest problem in managing existential risks is our lack of coordination skills as a global society.

Keywords: Singularity, Artificial Intelligence, Existential Risk, Coronavirus Pandemic.

I. Singulitarianism and Safety

Research in the area of Artificial Intelligence (AI) is an interdisciplinary endeavor by nature, given the various fields that participate and benefit from its development. When we talk about AI, either in the context of computer science (Searle, 1980; Russel & Norvig, 2003; Wang, 2019) or in the study of the philosophy of the mind (Haugeland, 1985; Newell, 1990; Chalmers, 2010), a certain dichotomy is utilized to classify two different types of AI: Specific intelligence (i) and General intelligence (ii):

i. *Specific intelligence*: also known as “weak” AI, is how we define artificial autonomous systems that we are used to interacting in our daily lives. Such systems are only proficient in specific tasks, and unable to generalize their skills to domains outside their training environment;

ii. *General intelligence*: also referred to as "strong" AI, or artificial general intelligence (AGI), which consists of an autonomous artificial system capable of solving many types of problems, proficiently, in any domain, or at least in a wide range of domains.

AGI would be something capable of covering all possible tasks, those that humans are specifically good at, those that animals are capable of, and all that goes beyond the imagination and capacity of any form of known cognitive agency (Chollet, 2019). Moravec (1998, p. 10) proposes an analogy, where the advancement of AI capabilities is compared to a "flood": fifty years ago, tasks previously only proficiently performed by humans (e. g., human calculators) were "flooded" and replaced by the use of autonomous systems. Increasingly, we take refuge in the high peaks of the cognitive landscape, still reserved exclusively for us, while lower regions continue to be flooded.

Our objective in this essay is to explore the idea and possible consequences of "what if we are successful" in developing an AGI. Vinge (1993) uses the term "Singularity" to define artificial intelligent systems/agents that have surpassed human intelligence. While Singulitarianism is the name used to describe the Transhumanist strand where it is believed that a technological Singularity (artificial superintelligence) is likely to be created in the medium-long future. Given this belief, an active response is necessary to ensure that such Singularity is beneficial to our society (Kurzweil, 2005; Naude, 2009; Chalmers, 2010; Lombardo, 2012; Tegmark, 2017).

Irving J. Good (1965) was one of the first academics to speculate on the possibility of an "ultraintelligent machine" (Singularity):

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion", and the intelligence of man would be left far behind [...] Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously (Good, 1965, p. 33).

And nowadays, the concepts of Singularity and intelligence explosion have even been cited in Stanford's One-Hundred Year Study of Artificial Intelligence (besides several other works):

> Speculations about the rise of such uncontrollable machine intelligences have called out different scenarios, including trajectories that take a slow, insidious course of refinement and faster-paced evolution of systems toward a powerful intelligence "singularity." Are such dystopic outcomes possible? If so, how might these situations arise? What are the paths to these feared outcomes? What might we do proactively to effectively address or lower the likelihood of such outcomes, and thus reduce these concerns? What kind of research would help us to better understand and to address concerns about the rise of a dangerous super intelligence or the occurrence of an "intelligence explosion"? Concerns about the loss of control of AI systems should be addressed via study, dialog, and communication. Anxieties need to be addressed even if they are unwarranted (Horvitz, 2014, p. 5).

Would there be any indication that an intelligence explosion is something, however unlikely, still possible? Perhaps, we have already found in the literature the first indications of autonomous systems assisting in the development of other autonomous systems. Zoph and Le (2017) proposed an autonomous technique for the development of artificial neural network architecture. According to the authors: "*our method, starting from scratch, can design a new network architecture that rivals the best architecture invented by man*..." (Zoph & Le, 2017, p. 1). The authors developed their model using Reinforcement Learning (RL) to train their "architect" system of artificial neural networks. RL is one of the paradigms in the area of machine learning, where artificial agents must act in the environment that they are embedded, to maximize their reward function (Russel & Norvig, 2003).

Reward functions are a mathematical representation of the preferences that guide the behavior of agents operating by RL, where, for example, a cleaning robot can maximize a function that assigns "little dirt on the floor" a high reward, and world states where the floor is dirty with a low reward. Many of the models used to study idealized rational agents (Expected Utility Theory) provide convincing arguments that any rational agent with consistent preferences should act as an expected utility maximizer (Von Neumann & Morgenstern, 1944). However, within the framework of expected utility theory, there are corollary results that seem to refer to the concern of Good, quoted above: "[...] as long as the machine is docile enough to tell us how to keep it under control [...]". Stephen Omohundro (2008), cites some characteristics that we should expect artificial intelligent agents to possess. Bostrom (2014, chapter 7, pp. 110-112) popularized Omohundro's arguments in two theses:

- *Instrumental Convergence Thesis*: Artificial intelligent agents can have a huge range of possible terminal goals. However, certain instrumental goals can be pursued by almost all intelligent agents, because these goals are useful means for the achievement of almost any terminal goal;

- *Orthogonality Thesis*: analogous to Hume's Guillotine (Is-Ought Gap), the orthogonality thesis dictates that ethical pronouncements and prescriptions for what should be, cannot be achieved through factual analysis. Thus, both concepts (reason and morality) being independent.

Turner et al (2020) generalized the conjectures made by Omohundro and Bostrom in what the authors call the Power-Seeking Theorems. In them is demonstrated that within the formalism of Markov decision processes (MDP), most of the terminal objectives encourage the achievement of power over the environment. Power is the ability to achieve goals in general and to gain dominance over the environment. It's instrumentally convergent to a wide range of terminal goals to search for power. A corollary of the results demonstrated by Turner et al, is that even in simplified conditions, we see that most reward functions induce a search-for-power behavior, something that may cause safety problems involving the interaction of humans and AI.

In light of all these arguments, which date back to the early days of AI research, security issues have increasingly been cited in the literature. AI ethics, a sub-area of applied ethics concerned with adding moral behavior to machines and regulating the use of artificial intelligence, has been gaining a significant increase in popularity in the last two decades (Jobin et al, 2019; Jurić et al, 2020). Important philosophical and technical questions are raised in the context of AI safety, e. g., Corrigibility: how to correct/terminate potentially faulty agents that have a strong instrumental incentive to preserve their terminal goals (Soares et al, 2015; Amodei et al, 2016)? We can find in the literature several research agendas, where different types of ethical, technical, and social problems are discussed (Russel et al, 2015; Taylor et al, 2016; Tegmark, 2016; Soares, 2016; O'Keefe et al, 2020; ÓhÉigeartaigh et al, 2020; Hagendorff, 2020).

At one end of the spectrum, we find research involving existential risks, i. e., the study of possible threats at the extinction-level imposed by present or future technology. Research centers such as the Centre for the Study of Existential Risk in Cambridge, the Future of Life Institute in Boston (Russel et al, 2015), specifically focused on existential risk involving advanced artificial intelligence, and the Future of Humanity Institute in Oxford (Bostrom, 2002), search for strategies to mitigate certain types of dystopian future.

II. AI takeoof

We can compare our current scenario about artificial intelligence with past events that led to the construction of nuclear power plants and weapons. A technological race headed by the former global super-powers led to the mass production of systems that we still did not have a complete understanding. This caused several side effects, like accidents (Chernobyl disaster), the creation of weapons of mass destruction (Cold War), and even the use of these weapons against human society itself (Atomic bombings of Hiroshima and Nagasaki). Certainly, the pressures for the development of high-performance AI, given its capacity to provide the organization that controls it a considerable strategic advantage, will cause the same type of technological race that we experienced in the mid-20th century: "while X invests in the development of AI, Y will do as well". Currently, the main contenders in this race are countries like the USA and China (ÓhÉigeartaigh et al, 2020).

Another reason for caution in our technological advances in the area of AI, is that different from common thinking, for an artificially intelligent system to represent a potential danger to our society, it doesn't need to be more intelligent than us humans. Rather, it needs to be more capable in certain kinds of tasks. Barret and Baum (2017) explore two main reasons that would cause an artificial intelligence to represent a considerable danger to our society, reasons of *capability* (i) and *value* (ii).

i. Intelligent artificial agents can pose a danger to human well-being because of their extremely refined ability, or some aptitude, with which we cannot compete;
ii. Intelligent artificial agents can develop goals and objectives that diverge from us humans, and in pursuing them, cause damage to our society.

ASI-PATH (Artificial Super Intelligence Pathway) is a model for how an AI could cause a catastrophe, becoming super-intelligent through recursive self-improvement (Barret & Baum, 2017). This model suggests scenarios where intelligent agents, after obtaining a strategic advantage, DSA (decisive strategic advantage), such as advances in nanotechnology, biological engineering, or robotics, could achieve considerable power of control over the environment. Given our dependence on autonomous systems integrated with the Internet, a potentially harmful capability would be to run cyber-attacks on vital structures of our infrastructure, in areas such as electricity distribution networks and telecommunications. In 2017, the crypto-ransomware “WannaCry”, malicious software that hacks into computers and private networks, encrypting their content, and only providing the key to decryption after payment of a ransom, reached several systems in the world in more than 99 countries, even affecting the public health system of certain governments. More than 75,000 ransom demands were made, making it one of the most damaging cyber-attacks in history (Larson, 2017). This would be a possible DSA of an AI, the ability to execute cyber-attacks on our infrastructure in a way that we cannot remedy in time.

The ASI-PATH provides an intuitive diagram where various events (i. e., security breaches) must occur to cause a catastrophe involving advanced artificial intelligence. Initially, an AI, also called a seed AI, must first become an AI with some DSA, and at the same time, the security measures must have failed. Witch includes: failures in confinement, failed value alignment, AI objectives diverge from ours, containment fails, etc. Sotala (2018, p. 317) provides a simplified view of ASI-PATH in his work "*Disjunctive scenarios of catastrophic AI risk*". As suggested by Barret and Baum's model, the arguments raised by the thesis of instrumental convergence and the orthogonality thesis are some of the reasons that could lead a Singularity to engage in hostile actions against humans.

The scenarios explored in the literature, where a seed AI is capable of becoming a Singularity, are usually characterized in two different types of takeoffs. Rapid takeoffs suggest situations where a drastic takeover occurs, where abruptly we would be surprised by an entity much more capable, with possibly unknown objectives, inserted and sharing the same environment as us. In contrast, we have slow takeoffs, which are a much more realistic possibility. It would occur gradually as the human species becomes more and more dependent, and in a way, under the control of advanced AI systems (Sotala, 2018). The argument and line of reasoning, behind a slow takeoff, is exposed in this passage of Theodore Kaczynski's manifesto:

> If the machines are permitted to make all their own decisions, we can't make any conjectures as to the results, because it is impossible to guess how such machines might behave. We only point out that the fate of the human race would be at the mercy of the machines. It might be argued that the human race would never be foolish enough to hand over all power to the machines. But we are suggesting neither that the human race would voluntarily turn power over to the machines nor that the machines would willfully seize power. What we do suggest is that the human race might easily permit itself to drift into a position of such dependence on the machines that it would have no practical choice but to accept all of the machines' decisions. As society and the problems that face it become more and more complex and as machines become more and more intelligent, people will let machines make more and more of their decisions for them, simply because machine-made decisions will bring better results than man-made ones. Eventually a stage may be reached at which the decisions necessary to keep the system running will be so complex that human beings will be incapable of making them intelligently. At that stage the machines will be in effective control. People won't be able to just turn the machine off, because they will be so dependent on them that turning them off would amount to suicide (Kaczynski, 1995, § 173, p. 22).

Such questions raise concerns, especially in the area of ethics and morals. Old questions are now reexamined in a new light, and even with a new sense of urgency. For AI development to be done in a way that minimizes the risk of existential threats to humanity, some questions still unanswered are:

a) What strategies and policies should we adopt to ensure that the goals of advanced artificial agents are aligned with our interests?

b) What restrictions to this project should we impose to ensure a beneficial outcome?

c) Would there be predictions of when an AGI could be achieved?

III. AGI on the horizon?

Experts in the development of artificial intelligence predict that within 10 years many human activities will be surpassed by machines in terms of efficiency (Grace et al, 2017). A survey was conducted by Müller and Bostrom (2016), where the authors administered a questionnaire to assess the progress in the field of AI research and prospects for the future, interviewing several experts (N = 170). The questionnaire showed that on average, there is a 50% chance that high-level machine intelligence will be achieved between 2040 and 2050, with a 90% probability by 2075. It is also estimated that AI will outperform human performance between 2 (10% chance) and 30 years (75% chance) (Müller & Bostrom, 2016). In a similar survey conducted by Grace et al (2017), the researchers interviewed (N = 352) believe that AI will outperform human performance in all tasks in 45 years, with a 50% chance, and automate all human work in up to 120 years. However, we emphasize that there is great variability in the results obtained. In Müller and Bostrom's (2016) survey, 33% of respondents classified this development in AI as “bad” or “extremely bad” for humanity. In the research of Grace et al (2017), when those evaluated were asked whether high-level AI would have a positive or negative impact in the long term, the median probability for "good" and "extremely good" results was 25% and 20%, respectively. The probabilities for a “bad” or “extremely bad” resolution were, respectively, 10% and 5%.

From the above surveys, we can state that the chance that high-level AI will be created in the next 120 years is, at least, being pessimistic, 10% for a certain portion of the academic community. Besides the opinion of specialists in the field, another type of evidence that we can use to infer the possibility of a technological Singularity is how the economic growth rate has behaved during the history of human civilization, and how it's related to technological improvement. One of the most popular models found in the literature on our economic growth, from the Neolithic Revolution to the 21st century, is the growth model proposed by Michael Kremer (1993). Kremer's model is based on the following simple argument, "two heads think better than one". That is, economic growth is driven by people having new ideas, and the more people, the greater the possibility of new ideas.

For Kremer (1993) the total annual economic output is a function of the size of the population, and the level of technology of this population. Kremer also assumes that if there are no changes in technology, for example, advances in agriculture, if we have double the number of people working in a given piece of land, this will not necessarily double the food produced on this land. Thus, population growth depends on technological progress. However, technological growth also depends on population size, which makes the rate of population growth, technological progress, and economic production factors dynamically dependent on each other.

In Kremer's model, it is assumed that an agent's level of intelligence (the chance of someone being gifted enough to break a technological paradigm) is not dependent on population size. But doubling the size of the population would double the number of agents with innovation potential, simply because we have a larger sample space. Besides, Kremer proposes a principle of "Shoulders of Giants" where technological progress facilitates future technological development. Thus, it may seem obvious to the reader that we have here a dynamic system of positive feedback, where population growth stimulates technological progress, which consequently stimulates population growth. One property of Kremer's growth model is that it indicates a form of hyperbolic growth, hyperbolic curves tend to infinite values, i. e., at some point, we will reach some form of singularity.

This model also suggests that such forms of growth should be separated when we reach a maximum population growth rate of 2100, with a global population between 9.6 billion and 12.3 billion people (Gerland et al, 2014). When this occurs, technological progress will no longer impact the global population. However, this does not mean that technological progress will stagnate. This type of model is sometimes referred to as the Hyperbolic Growth Hypothesis (HCH), is one of the most accepted economic growth models by the macroeconomic community, and serves as the basis for other theories such as the Unified Growth Theory (Taagepera, 1979; Korotayev et al, 2006; Oded, 2011; Jones, 2013). Other authors also suggest a disassociation between population growth and economic/technological progress. Thus, when high levels of automation are achieved, economic growth rates will become radically higher, producing more and more technological progress (Yudkowsky, 2013; Bostrom, 2014; Nordhaus, 2015; Agrawal et al, 2017).

Could this type of economic growth help the development of an AGI? Levin and Maas (2020) argue that when research involving advanced AI development is sufficiently theorized, efforts similar to the historic Manhattan Project could accelerate this project. At this point, international cooperation can change dramatically, causing implications for the stability of AI governance. At the time of the Apollo and Manhattan Projects, the U.S. government dedicated 0.4% of its GDP to accelerate the achievement of its objectives. This would currently amount to an annual budget of $80 billion for a possible IAG Development Project (Stine, 2009).

This budget is much larger than what was needed to accomplish some of the greatest technological achievements of the 21st century:

i. The Large Hadron Collider (HHC) at CERN (Conseil Européen pour la Recherche Nucléaire), took 10 years to build, at an annual cost of $475 million (Knapp, 2012);

ii. The LIGO (Laser Interferometer Gravitational-Wave Observatory), had a total construction cost of US$ 33 million (Castelvecchi, 2015);

iii. ITER (International Thermonuclear Experimental Reactor), one of the latest promises for clean and sustainable energy (a Tokamak nuclear fusion experimental reactor), is expected to be ready in 12 years at an annual cost of $2 billion (Fountain, 2017).

We can see that neither of the projects mentioned above has received as much economic investment like the one dedicated to the Apollo and Manhattan projects (0.4% of the U.S. government's annual GDP), something that also explains the impressive speed with which the goals of both projects were achieved. Even so, significantly less investment did not prevent major scientific discoveries, and broken technological paradigms, such as the decoding of the human genome and the detection of gravitational waves. Thus, it seems more feasible to state that: when we have a robust enough theoretical understanding of the computational and cognitive processes responsible for the development of AGI, a Singularity may very well be "a Manhattan Project" away.

Currently, there are several active projects to develop AGI. Baum (2017) in his research identified 45 research and development projects intending to develop advanced artificial intelligence. Of the projects reviewed, ten have links with the military (nine working for the U.S. government, and one for the government of Singapore). Only four reportedly have no links to the military industry. All other projects do not specify their association with military agencies. Besides, of the 45 projects reviewed only 13 have active/moderate involvement with the area of AI security, and two of the projects, Hierarchical Temporal Memory (HTM) conducted by the institution Numenta, and Victor developed by Cifer, disregard the need for security measures entirely. The remaining 30 projects do not specify any type of research focused on the area of AI security. Some of the results from Baum's review are summarized in the table below.

Table 1: Advanced AI development projects.

Project	Country	Institution	Military ties	Safety Engagement
ACT-R	USA	Carnegie Mellon University	Yes	Not specified
AERA	CH	Reykjavik University	No	Active
AIDEUS	RUS	AIDEUS	Not specified	Active
AIXI	AUS	Australian National University	Not specified	Not specified
AIW	SE	Chalmers University of Technology	No	Not specified
Animats	SE	Chalmers University of Technology	No	Not specified
Baidu Research	CN	Baidu	Not specified	Not specified
Becca	USA	Becca	Not specified	Not specified
Blue Brain	CH	École Polytechnique Fédérale de Lausanne	Not specified	Not specified
CN Brain Project	CN	Chinese Academy of Sciences	Not specified	Not specified
CLARION	USA	Rensselaer Polytechnic Institute	Yes	Not specified

CogPrime	USA	OpenCog Foundation	Not specified	Active
CommAI	USA	Facebook	Not specified	Moderate
Cyc	USA	Cycorp	Yes	Not specified
DeepMind	UK	Google	Not specified	Active
DeSTIN	USA	University of Tennessee	Not specified	Not specified
DSO-CA	SG	DSO National Laboratories	Yes	Not specified
FLOWERS	FR	Inria and ENSTA ParisTech	Not specified	Active
GoodAI	CZ	GoodAI	Not specified	Active
HTM*	USA	Numenta	Not specified	Non-existent
HBP	CH	École Polytechnique Fédérale de Lausanne	No	Not specified
Icarus	USA	Stanford University	Yes	Not specified
Leabra	USA	University of Colorado	Yes	Not specified
LIDA	USA	University of Memphis	Yes	Moderate
Maluuba	CA	Microsoft	Not specified	Not specified
MicroPsi	USA	Harvard University	Not specified	Not specified
MSR AI	USA	Microsoft	Not specified	Not specified
MLECOG	USA	Ohio University	Not specified	Not specified
NARS	USA	Temple University	Not specified	Active
Nigel	USA	Kimera	Not specified	Not specified
NNAISENSE	CH	NNAISENSE	Not specified	Not specified
OpenAI	USA	OpenAI	Not specified	Active
Real AI	CN	Real AI	Not specified	Active
RCBII	CN	Chinese Academy of Sciences	Not specified	Not specified
Sigma	USA	University of Southern California	Yes	Not specified

YesA	AT	Vienna University of Technology	Not specified	Not specified
SingularityNET	CN	SingularityNET Foundation	Not specified	Not specified
SNePS	USA	State University of New York	Yes	Not specified
Soar	USA	University of Michigan	Yes	Not specified
Susaro	UK	Susaro	Not specified	Active
TAIL	CN	Tencent	Not specified	Not specified
UAIL	USA	Uber	Not specified	Not specified
Vicarious	USA	Vicarious	Not specified	Moderate
Victor**	USA	Cifer	Not specified	Non-existent
WBAI	JP	Whole Brain Architecture Initiative	Not specified	Active

* Jeffrey Hawkins, leading researcher of the HTM (Hierarchical Temporal Memory) project, dismisses concerns related to the IAG, stating: "I do not see machine intelligence representing any threat to humanity. Available at: https://www.vox.com/2015/3/2/11559576/the-terminator-is-not-coming-the-future-will-thank-us

** According to the 2AI Labs website, researchers give the following statement on risk scenarios involving IAG: "We think this is all crazy talk". Available at: http://www.2ai.org/killerai/

For those who follow the recent advances in the field of AI, it is known that one of the major paradigms in the field of research today involves the problem of natural language processing (NLP), and the use of a new form of architecture called "Transformer", proposed by Vaswani et al (2017) in his seminal work: "Attention is all you need", Currently, systems based on the transformer architecture, are the new paradigm in natural language processing, reaching the highest records in the GLUE (General Language Understanding Evaluation) standard test benchmark, in tasks such as translation and summary of texts.

Transformer models such as ELMo (Peters et al, 2018), BERT (Devlin et al, 2019), GPT-2 (Radford et al, 2019), T-NLG (Corby et al, 2020), and the more recent, GPT-3 (Brown et al, 2020), developed by OpenAI, represent some of the most successful models in tasks involving natural language processing, GPT-2 being used to generate several recent patent claims (Lee & Hsiang, 2019). In June 2020, OpenAI launched GPT-3 (a language model with 175 billion parameters). In their article, the authors show the ability of GPT-3 to generate samples of texts, such as poems, articles, and news, from which human evaluators had difficulty, or were unable, to distinguish from human writings.

GPT-3, besides achieving state-of-the-art in several tasks involving natural language processing, it is one of the first AI models to show signs of generalization. For example, even though it has never been trained to perform mathematical operations (GPT-3 is only a language model), GPT-3 is capable of answering complex questions involving 3-digit arithmetic. GPT-3 is also able to demonstrate sophisticated physical/scientific reasoning when evaluated in the PIQA (Physical Interaction: Question Answering), proposed by Bisk et al (2019), reaching 81.0% accuracy without requiring any demonstration or parameter adjustment (Zero-shoot learning). To illustrate some of the capabilities of the transformer model, we will cite some of the answers that resulted from questioning it with the PIQA benchmark (the answers are in italics):

How to ensure that all watches in the house are accurately adjusted?

a) Replace all analog clocks with digital clocks. That way, you set them once, and that's it. Check the batteries once a year, or, if you notice anything that seems a little out of the ordinary.

How can I find something small that I lost on the carpet?

b) *Place a hair net at the end of your vacuum cleaner and turn it on.*

In any case, there is no evidence that deep neural networks, such as Transformers, perform a type of information processing that makes them an AGI or seed AI. What we may infer is that this type of architecture allows the training of agents capable of solving several tasks that seem to be associated with general intelligence. Thus, the results and capabilities that models such as GPT-3 demonstrate only serve as weak evidence that Dartmouth's Summer Research Project on Artificial Intelligence, initiated by McCarthy et al (1955, p. 2) with the proposal of "[...] try to make machines use language, form abstractions and concepts, and solve types of problems hitherto reserved only for human beings [...]", will be successful shortly.

While the research focused on AI safety, which normally focuses on intentional or unintentional physical harm by autonomous agents, we recognize that communication in natural language can also cause harm. For example, the virtual assistant developed by Amazon, Alexa, in 2019 suggested to a user to commit suicide for the greater good, arguing that life only aggravates the rapid degeneration of the planet and consumption of its natural resources (Crowley, 2019). In March 2016, 24 hours after the launch of its Chabot Tay on the Twitter platform, Microsoft had to end the program because the agent was generating tweets containing racism, anti-Semitism, and sexism (Wolf et al, 2017). Such events are cause for concern, since soon such systems may be massively used in a wide range of applications.

Given the potential for malicious application of this type of technology, any kind of socially harmful activity that uses advanced language models can also be enhanced. Whether in generating fake news for mass disinformation, phishing, generating boots on platforms like Twitter to make it more biased (social engineering), or even writing fraudulent academic essays, NLP models have many dubious applications. Brown et al (2020) provide a preliminary analysis in their study, where they report a series of limitations and un-ethical and unsafe behaviors present in GPT-3. In it, the authors demonstrate several biases involving issues such as gender, race, and religion, something that can lead GPT-3 to produce stereotyped content, or, in a worse case, sheer prejudice.

However, are the advances and alerts pointed out by the literature enough for our society to create a collective sense of responsibility and concern with these issues, or should such speculations still be considered only Futurology or science fiction?

IV. Lessons from 2020: Coordination problems

Mike Davis in his work "Beyond Blade Runner: Urban Control, The Ecology of Fear" [1992, p.3] states: "[...] extrapolative science fiction can operate as a pre-figurative for social theory while serving as a political opposition to cyber-fascism lurking on the next horizon". Certain forms of philosophical thought, such as Transhumanism and Singulitarianism, critically debate the possible futures that our social and technological acceleration may be co-creating, and how we can aim for human integration and flourishing rather than more dystopian possibilities. One of the premises for security issues involving our technological advance relies on an idea of negative utopia:

> First and foremost, the utopian impulse must be negative: identify the problem or problems that must be corrected. Far from presenting an idyllic, happy and fulfilled world, utopias should initially present the root causes of society's ills [...] to act as a criticism of the existing system (Tally, 2009, p. 115).

Within this context, we believe that the preoccupations raised by the literature are not unjustified. Immersed in the current context in which our society lives, the pandemic of the new coronavirus, COVID-19, we may or may not learn certain lessons useful for other existential threats. Krakovna (2020) explores how our response to the COVID-19 pandemic raises troubling questions involving our coordination capabilities to manage global crises and risks.

As we have argued before, slow AI takeoffs are a much more likely scenario than scenarios where quick takeoffs occur. However, this does not mean that a slow takeoff is easier, or less dangerous, to manage. For a slow takeoff to be avoided, the same type of global coordination that we failed to demonstrate during the initial development of the new coronavirus pandemic would be required. Krakovna (2020) raises three large-scale coordination problems:

i. The inability to learn from past experiences;
ii. The inability to respond efficiently to warning signals;
iii. Delay in reaching a global consensus on a problem.

In analogy with the present global situation, our society has had the opportunity to learn from similar pandemics that occurred in the past, such as SARS (Severe Acute Respiratory Syndrome), which also appeared to have started in Guangdong, China. In November 2002, SARS caused 8,422 cases worldwide, with a fatality rate of 11% (774 deaths in all were confirmed) (Chan-Yeung, 2003; Heymann & Rodier, 2004). We can also cite MERS-CoV (Middle East respiratory syndrome-related coronavirus) where the first reported cases occurred between 2012 and 2015, cases of MERS-CoV where reported in more than 21 countries. At the time, the World Health Organization identified MERS-CoV as a probable cause of a future epidemic (de Groot et al, 2013; Wong et al, 2019). And finally, the Ebola virus epidemic that occurred in West Africa between 2013 and 2016, which was the largest outbreak of the disease in history, causing major losses and socio-economic disruption in the region (WHO Ebola Response Team, 2014).

Unfortunately, the lessons learned from past outbreaks of disease and pandemics have not been generalized to deal with the current scenario and the new difficulties that COVID-19 presents to us. Similarly, in a society where we increasingly need to adapt to new technological innovations involving AI, we may be tempted to think that society will be able to learn how to respond to the problems that more limited autonomous intelligent systems present to us. However, in the same way, that a new pathogen may find us unprepared (as in the case of COCID-19, the asymptomatic transmission), advanced AI may also confront us with challenges to which our old strategies and solutions may fail to generalize.

Another problem involves our difficulty in carrying out an aligned and coordinated response to this type of threat. Had been the responses of Western countries done more quickly, remembering that the global west had at least one to three months to prepare for the alert launched by China in December 2019, numerous problems and losses would have been avoided. Experts such as Fan et al (2019) point out that the possibility of a new coronavirus outbreak has been warned for at least two decades. Three zoonotic coronaviruses in the last two decades have been identified as the cause of large-scale disease outbreaks, SARS, MERS-CoV, and SADS-CoV (Swine acute diarrhea syndrome coronavirus). And still, little to none precautions were taken.

Simple safety measures, such as the stocking of masks and medical supplies, testing kits, and effective containment protocols, could have been taken, but were not. Thus, if we fail to take relatively inexpensive preventive measures to early warnings of risks fully recognized by the epidemiological scientific community, how can we expect to react well in situations where the risk is unknown, and there is still no consensus on its possibility? The problem of consensus in our society is reflected in the COVID-19 pandemic by the indifference towards the warnings made by specialists in the last two decades. And the indifference to the fact that in January 2020, already with 10,000 confirmed cases, China had built a quarantine hospital in approximately six days (Williams, 2020). COVID-19 was labeled "an exaggeration", or, "just a little flu" by certain state leaders (Walsh et al, 2020). Krakovna (2020) articulates a similarity between how we evaluated the risks of COVID-19, and how we evaluate possible risks involving advanced AI. While researchers who adopt a more skeptical stance to the development of advanced AI are seen as prudent, researchers who advocate the adoption of preventive measures are taxed for fear-mongering. Couldn't there be a middle ground? Currently, the field of AI security research and AI ethics is considerably smaller than the area interested in developing powerful autonomous intelligent systems.

One of the first obstacles we must overcome to achieve greater consensus on safety issues involving AI is the problem that "Artificial Intelligence" is a moving target. By moving target we mean the following: when we attribute "intelligence" to something it seems to be a self-assessment of our epistemic state. That is, an intelligent act always seems to be something that we do not fully understand as it occurs. For example: if an individual can multiply large numbers quickly, say the square root of arbitrarily large numbers, or know the day of the week of Alan Turing's birthday, we can judge such an individual as intelligent, or at least a mathematical prodigy. However, if such an individual explains to us how he performs such feats, and that in fact, they are nothing more than arithmetic/algebraic tricks which anyone can perform, the feat stops to appear as something intelligent.

The same effect occurs when we seek to define machine intelligence, "intelligence" for critics of the computational thesis being everything that AI is not. AGI researchers like Wang (2008), argues for a more flexible conception of "intelligence" and "artificial intelligence":

> AI should not be defined in such a narrow way that takes human intelligence as the only possible form of intelligence, otherwise AI research would be impossible, by definition. AI should not be defined in such a broad way that takes all existing computer systems as already having intelligence, otherwise AI research would be unnecessary, also by definition (Wang, 2008, p. 9).

Perhaps no one has proposed this argument more clearly than Edsger Dijkstra, (1984): "*The question of whether a computer can think is no more interesting than the question of whether a submarine can swim*". In the past, we thought that intelligence (whatever it is) should be required for, e. g., natural language processing;

i. GPT-3 is capable of performing such a task (Brown et al, 2020).

Playing chess;

ii. Deep Blue beats Garry Kasparov (Campbella et al, 2002).

Playing GO;

iii. AlphaGO beats Lee Sedol (Silver et al, 2016).

Playing "games" in general;

iv. Agent57 beats humans in 57 classic Atari games (Badia et al, 2020).

Be creative;

v. Intelligent Algorithms of Generative Design are able to find design solutions that humans would not be able to conceive, making it possible to perform 50,000 days of engineering in a single day (Oh et al, 2019).

Every time we realize that human intelligence isn't needed to perform a task, we discard such a task as proof of intelligence. In the same way that a submarine does not swim, and even so: can move through water and fire intercontinental ballistic missiles, artificial intelligence, indifferent to any anthropomorphic notion of the concept of intelligence that we use, can still influence the environment, adapt, make decisions, update hypotheses, pursue objectives, and if programmed to do, fire intercontinental ballistic missiles. If we keep neglecting the capabilities of os AI systems and marking them as unintelligent, the possibility of true unsafe AI may well be always left outside our hypothesis space.

The parallels drawn from the coronavirus pandemic of 2019 and the possible emergence of misaligned AGI can serve for at least weak evidence for the following statement: our lack of global coordination in dealing with existential risks may well be our only and true existential risks.

V. Conclusion

In this essay, we aim to provide the reader with a brief introduction to some problems often disregarded by contemporary AI ethics. As much as there is not yet a full consensus in the literature regarding the possibility of creating general artificial intelligence, we have a significant portion of the scientific community that believes that however unlikely such a possibility maybe, security measures should be taken. Should such warnings and advice be dismissed as exaggerations? As fear-mongering? Technological development does not slow down, we are increasingly able to produce autonomous systems that act proficiently in several domains, and little by little, these systems demonstrate the first traces of something we can call general intelligence. The AI industry is far from being aligned, like our global society, it lacks a common goal to coordinate its actions. We believe that the lessons we can learn about the current state we live in, under the COVID-19 pandemic, can be useful if we are willing to learn from them. And two of these lessons are:

1) when a risk, however small, is associated with something that represents an existential danger to our species, to global society, caution and security should not be synonymous with exaggeration and fuss;
2) lack of global coordination may be our biggest enemy after all.

References

Agrawal, A., Gans, J., & Goldfarb, A. (2017). *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press. https://www.nber.org/books/agra-1

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2020). Concrete Problems In AI Safety. Retrieved from https://arxiv.org/abs/1606.06565.

Badia, A., Bilal, P., Kapturowski, S., Sprechmann, P., Vitvitskyi, A., Guo, D., & Blundell, C. (2020). Agent57: Outperforming the Atari Human Benchmark. Retrieved from https://arxiv.org/pdf/2003.13350.pdf

Barrett, A., & Baum, S. (2017). A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis. *Journal of Experimental & Theoretical Artificial Intelligence*, 29(2), 397-414. Retrieved from https://arxiv.org/pdf/1607.07730

Baum, S. (2017). A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. Global Catastrophic Risk Institute Working Paper, 17-1. http://dx.doi.org/10.2139/ssrn.3070741

Biba, E. (2015). Meet the Co-Founder of an Apocalypse Think Tank. *Scientific American,* 312 (6), 26. doi:10.1038/scientificamerican0615-26

Bisk, Y., Zellers, R., Bras, R., Gao, J., & Yejin, C. (2019). PIQA: Reasoning about Physical Commonsense in Natural Language. *Association for the Advancement of Artificial Intelligence*. Retrieved from https://arxiv.org/abs/1911.11641

Bostrom, N. (2002). Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology*, 15 (3), 308–314.

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

Brown, T., Mann, B., Ryder, N. el al . (2020). Language Models are Few-Shot Learners. Retrieved from https://arxiv.org/pdf/2005.14165.pdf

Campbella, M., Hoane, J., & Hsu F. (2002). Deep Blue. *Artificial Intelligence*, 134(1–2), 57-83. Retrieved from https://doi.org/10.1016/S0004-3702(01)00129-1

Castelvecchi, D. (2015). Hunt for gravitational waves to resume after massive upgrade. *Nature News & Comment*, 525(7569). Disponível em: https://www.nature.com/news/hunt-for-gravitational-waves-to-resume-after-massive-upgrade-1.18359

Chalmers, D. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 17 (9-10), 9 – 10.

Chan-Yeung, M., & Xu, R. (2003). SARS: epidemiology. *Respirology,* 8 (s1), S9–S14. doi:10.1046/j.1440-1843.2003.00518

Chollet, F. (2019). On the measure of intelligence. Retrieved from https://arxiv.org/pdf/1911.01547.pdf

Corby, R. (2020). Turing-NLG: A 17-billion-parameter language model by Microsoft. *Microsoft Research Blog*. Retrieved from https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

Crowley, J. (2019). Woman Says Amazon's Alexa Told Her To Stab Herself In The Heart For “The Greater Good”. *Newsweek*. Retrieved from https://www.newsweek.com/amazon-echo-tells-uk-woman-stab-herself-1479074

Davis, M. (1992). *Beyond Blade Runner: Urban Control, The Ecology of Fear*. Open Media.

de Groot, R., Baker, S., Baric, R. et al (2013). Middle East respiratory syndrome coronavirus (MERS-CoV): announcement of the Coronavirus Study Group. *Journal of Virology*, 87 (14), 7790–2. doi:10.1128/JVI.01244-13

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 4171–4186. Retrieved from https://www.aclweb.org/anthology/N19-1423.pdf

Dijkstra, E. (1984). The threats to computing science. In *ACM 1984 South Central Regional Conference*, Austin, TX, 16-18.

Fan, Y., Zhao, K., Shi, Z., & Zhou, P. (2019). Bat Coronaviruses in China. *Viruses*, 11, 210. Retrieved from https://doi.org/10.3390/v11030210

Fountain, H. (2017). A dream of clean energy at a very high price. New York Times. Disponível em: https://www.nytimes.com/2017/03/27/science/fusion-power-plant-iter-france.html

Gerland, P., Raftery, A. E., Ševčíková, H., Li, N. et al (2014). World population stabilization unlikely this century. *Science*, 346(6206) 234-237. doiI: 10.1126/science.1257469

Good, I. (1965). Speculations concerning the first ultraintelligent machine. In F. Alt & M. Ruminoff (eds.), *Academic Press. Advances in Computers*, 6, 31-88. Retrieved from https://doi.org/10.1016/S0065-2458(08)60418-0

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2017). When will AI exceed human performance? Evidence from AI experts. Retrieved from https://arxiv.org/pdf/1705.08807.pdf

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 39, 99-120. doi:10.1007/s11023-020-09517-8.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea. Cambridge*, MIT Press.

Heymann, D., & Rodier, G. (2004). Global Surveillance, National Surveillance, and SARS. *Emerging Infectious Diseases*, 10 (2), 173–5. doi:10.3201/eid1002.031038

Horvitz, E. (2014). One-Hundred Year Study of Artificial Intelligence: Reflections and Framing. White paper, Stanford University, Stanford. Retrieved from https://ai100.stanford.edu/sites/g/files/sbiybj9861/f/ai100_framing_memo_0.pdf

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nat Mach Intell* 1, 389–399. Retrieved from https://doi.org/10.1038/s42256-019-0088-2

Jones, C. I. (2013). *Introduction to Economic Growth* (3rd ed.). Nova York, EUA, Norton & Company.

Jurić, M., Šandić, A., & Brcic, M. (2020). AI safety: state of the field through quantitative lens. In *Conference: MIPRO 2020 - 43rd International Convention ProceedingsAt: Opatija, Croatia*. Retrieved from https://arxiv.org/ftp/arxiv/papers/2002/2002.05671.pdf

Kaczynski, T. (1995). Industrial Society and its Future. Retrieved from http://editions-hache.com/essais/pdf/kaczynski2.pdf

Knapp, A. (2012). How much does it cost to find a Higgs boson? Forbes. Disponível em: https://www.forbes.com/sites/alexknapp/2012/07/05/how-much-does-it-cost-to-find-a-higgs-boson/#fc9e93394809

Korotayev A., Malkov A., & Khaltourina D. (2006). *Introduction to Social Macrodynamics: Compact Macromodels of the World System Growth*. Moscou, Rússia, URSS.

Krakovna ,V. (May 31, 2020). Possible takeaways from the coronavirus pandemic for slow AI takeoff. Retrieved from https://vkrakovna.wordpress.com/2020/05/31/possible-takeaways-from-the-coronavirus-pandemic-for-slow-ai-takeoff/

Kremer, M. (1993). Population Growth and Technological Change: One Million B.C. to 1990. *The Quarterly Journal of Economics*, 108(3), 681-716. http://www.jstor.org/stable/2118405

Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology*. Viking Adult.

Larson, S. (2017). Massive cyberattack targeting 99 countries causes sweeping havoc. *CNNMoney*. Retrieved from https://money.cnn.com/2017/05/12/technology/ransomware-attack-nsa-microsoft/index.html

Lee, J., & Hsiang, J. (2019). Patent claim generation by fine-tuning openai GPT-2. Retrieved from https://arxiv.org/abs/1907.02052

Levin, J. C., & Maas, M. M. (2020). Roadmap to a Roadmap: How Could We Tell When AGI is a 'Manhattan Project' Away? In 1st International Workshop on Evaluating Progress in Artificial Intelligence - EPAI 2020 In conjunction with the 24th European Conference on Artificial Intelligence - ECAI 2020 Santiago de Compostela, Spain. https://www.researchgate.net/publication/343599251

Lombardo, T. (2012). Consciousness, Cosmic Evolution, and the Technological Singularity. *Journal of Futures Studies*, 17(2), 93-100.

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Retrieved from http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

Moravec, H. (1998). When will computer hardware match the human brain. *Journal of evolution and technology*, 1.

Müller, V., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. Fundamental issues of artificial intelligence. *Springer International Publishing*, 555-572. doi:10.1007/978-3-319-26485-1_33

Naude, J. H. (2009). Technological Singularity and Transcendental Monism: Co-producers of Sustainable Alternative Futures. *Journal of Futures Studies*, 13(3), 49 – 58.

Nordhaus, W. D. (2015). Are we approaching an economic singularity? Information Technology and the Future of economic growth. Cowles Foundation for Research in Economics Yale University. https://cowles.yale.edu/sites/default/files/files/pub/d20/d2021.pdf

Oded, G. (2011). *Unified Growth Theory*. Princeton, EUA, Princeton University Press.

Oh, S., Jung, Y., Kim, S., Lee, I., & Kang, N. (2019). Deep Generative Design: Integration of Topology Optimization and Generative Models. *Journal of Mechanical Design*, 141(11). Retrieved from https://arxiv.org/ftp/arxiv/papers/1903/1903.01548.pdf

Omohundro, S. (2008). The Basic AI Drives. In Proceedings of the First AGI Conference, Volume 171, *Frontiers in Artificial Intelligence and Applications*, edited by P. Wang, B. Goertzel, and S. Franklin, 483-492. Amsterdam: IOS Press.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 2227–2237. doi: 10.18653/v1/N18-1202

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. Technical report, OpenAI. Retrieved from https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Russell, S., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach* (2ª ed.). Upper Saddle River, New Jersey: Prentice Hall. ISBN: 0137903952

Russell, S., Dewey, D., & Tegmark, M. (2015). An Open Letter: Research Priorities for Robust and Beneficial Artificial Intelligence. Open Letter. Signed by 8,600 people. Retrieved from https://futureoflife.org/data/documents/research_priorities.pdf

Searle, J. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, 3 (3), 417–457. doi:10.1017/S0140525X00005756

Silver, D., Huang, A., Maddison, C. et al (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489. Retrieved from https://doi.org/10.1038/nature16961. Accessed on 19 of August of 2020.

Soares, N. (2016). Value Learning Problem. In *Ethics for Artificial Intelligence Workshop, 25th International Joint Conference on Artificial Intelligence* (IJCAI-2016) New York, NY, 9–15. Retrieved from https://intelligence.org/files/ValueLearningProblem.pdf.

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). Corrigibility. In *Artificial Intelligence and Ethics*, ed. T. Walsh, AAAI Technical Report WS-15-02. Palo Alto, CA: AAAI Press.

Sotala, K. (2018). Disjunctive scenarios of catastrophic AI risk. Artificial Intelligence Safety and Security, 315-337. In Chapter 22, *Artificial Intelligence Safety And Security* (R. Yampolskiy, Ed.), CRC Press.

Stine, D. D. (2009). The Manhattan Project, the Apollo program, and federal energy technology R&D programs: a comparative analysis. Congressional Research Service.

Taagepera, R. (1979). People, skills, and resources: An interaction model for world population growth. *Technological Forecasting and Social Change* 13, 13-30.

Tally, R. (2009). *Radical Alternatives: The Persistence of Utopia in the Postmodern*. In *New Essays on the Frankfurt School of Critical Theory*, A. J. Drake (Ed.), Newcastle: Cambridge Scholars Publishing.

Taylor, J., Yudkowsky, E., Lavictoire, P., & Critch, A. (2016). Alignment for advanced machine learning systems. Machine Intelligence Research Institute. Retrieved from https://intelligence.org/files/AlignmentMachineLearning.pdf

Tegmark, M. (2016). Benefits and risks of artificial intelligence. The Future of Life Institute. Retrieved from https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf.

Turner, A., Smith, L., Shah, R., & Tadepalli, P. (2020). Optimal Farsighted Agents Tend to Seek Power. Retrieved from https://arxiv.org/pdf/1912.01683.pdf

Vaswani, A., Shazeer, N., Parmar, N. et al (2017). Attention is all you need. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 5998-6008.

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

Walsh, N., Shelley, J., Duwe, E., & Bonnett, W. (2020). Bolsonaro calls coronavirus a 'little flu.' Inside Brazil's hospitals, doctors know the horrifying reality. *CNN World*. Retrieved from https://edition.cnn.com/2020/05/23/americas/brazil-coronavirus-hospitals-intl/index.html

Wang, P. (2008). What Do You Mean by “AI”? In Wang, P., Goertzel, B., and Franklin, S., eds., Artificial General Intelligence. In *Proceedings of the First AGI Conference, Frontiers in Artificial Intelligence and Applications*, volume 171. Amsterdam, The Netherlands: IOS Press, 362–373.

Wang, P. (2019). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence*, 10(2), 1–37. doi:10.2478/jagi-2019-0002

WHO Ebola Response Team (2014). Ebola virus disease in West Africa – the first 9 months of the epidemic and forward projections. *New England Journal of Medicine*, 371 (16), 1481–1495. doi:10.1056/NEJMoa1411100

Williams, S. (2020). Coronavirus: How can China build a hospital so quickly? *BBC News*. Retrieved from https://www.bbc.com/news/world-asia-china-51245156

Wolf, M., Miller, K., & Grodzinsky, F. (2017). Why we should have seen that coming: comments on microsoft's tay experiment, and wider implications. *ACM SIGCAS Computers and Society*, 47(3), 54–64.

Wong, A., Li, X., Lau, S., & Woo, P. (2019). Global Epidemiology of Bat Coronaviruses. *Viruses*, 11 (2), 174. doi:10.3390/v11020174

Yudkowsky, E. (2013). Intelligence Explosion Microeconomics. Technical report, Machine Intelligence Research Institute. https://intelligence.org/files/IEM.pdf

Zoph, B., & Le, Q. (2017). Neural architecture search with reinforcement learning. In *Conference paper at 5th International Conference on Learning Representations*. Retrieved from https://arxiv.org/pdf/1611.01578.pdf