The Probability of a Global Catastrophe in the World with Exponentially Growing Technologies

Justin Shovelain
Convergence
jshovelainsiai@gmail.com

Alexey Turchin
Foundation Science for Life Extension
alexeiturchin@gmail.com

Nov 2018

Abstract. In this article is presented a model of the change of the probability of the global catastrophic risks in the world with exponentially evolving technologies. Increasingly cheaper technologies become accessible to a larger number of agents. Also, the technologies become more capable to cause a global catastrophe. Examples of such dangerous technologies are artificial viruses constructed by the means of synthetic biology, non-aligned AI and, to less extent, nanotech and nuclear proliferation. The model shows at least double exponential growth of the probability of the global catastrophe which means that the accumulated probability of the catastrophe will grow from negligible to overwhelming at the period of a few doubling times of the technological capabilities. For biotech and AI, such doubling time roughly corresponds to the doubling period of Moore's law and its analogues in other technologies and is around 2 years. Thus the global catastrophe in the exponential technologies world will most likely happen during a "dangerous decade". Such a dangerous decade could start as early as in the 2020s. We also found that the double exponential growth in the model makes the model less sensitive to its initial conditions, like the number of dangerous agents or initial probabilities constants.

The model also shows that smaller catastrophes could happen earlier than larger ones, and such a smaller catastrophe may be able to stop the technological growth before larger ones become possible, thus global risks will be self-limiting. However, if the growth of the number of dangerous agents will be very quick, the multiple smaller catastrophes could happen simultaneously and will be equal to a global catastrophe.

Highlights:

- The exponential development of the technologies implies at least double exponential growth of the probability of global catastrophe.
- The main drivers of risks are the growth of the number of actors, the increase of the danger of each technology and the appearance of the new types of risks.
- Catastrophic risk will grow from negligible to inevitable in a few technological doublings, which at the current rate of progress is equal to the approximately one decade.
- Smaller catastrophes could be the self-limiting factor of global risks in the case of slower progress, as they will stop technological development.
- One of the ways to prevent risks is a global control system based on international agreements, narrow AI-based monitoring or superhuman AI.

1 Introduction

2 The model of the catastrophe depending on the number of agents and the speed of the technological progress 3
2.1 A case of just one dangerous technology and no technological progress 3
2.2 A case with an exponentially growing number of the dangerous agents 4
2.3 A case where the technology becomes more dangerous in time 5
2.4 A case where technological progress is accelerating 7
2.5 A case where the number of the types of possible catastrophic risks is growing 7
2.6 Exploring the properties of the model 7
3 Applying the model to the currently existing technologies 8
3.1 Democratization of synthetic biology 8
3.2 Creation of non-aligned AI 9
4 Could smaller catastrophes prevent the large one? 9
5 Global control system for the prevention of the risks' explosion 10
References 10

1 Introduction

There are many global catastrophic risks which could cause human extinction. Some of them are natural, others are anthropogenic (Bostrom, 2002; Turchin, 2015). Between the anthropogenic risks, there is a cluster of the global risks connected with quickly evolving new technologies, like biotech and AI. Advances of these technologies result in their democratization, as their price is dropping, and the number of the potentially dangerous individual or organizations which may have access to them (which we will call "agents") is growing.

These "doomsday agents" or "omnicidal agents" were analyzed by Phil Torres (Torres, 2016). Here we will ignore their motivational structure and assume that if an agent has access to a dangerous technology, there is some small probability that the agent will use it in a dangerous way. Torres analyzed the growth of the number of entities capable of creating global risks in Section 2.1 of (Torres, 2018). As the efficiency of any given technology is growing, it is also becoming more capable to cause an accident or to be misused by an "omnicidal agent". However, the most dangerous is "rational risk-taking" where adversarial effects are underestimated by the perpetrator.

Sotos suggested a model in which dangerous biotechnologies are an explanation of the Great Filter in the Fermi paradox (Sotos, 2017). In his model, the number of agents, capable to cause a global catastrophe is constant as well as the uniform probability per annum per agent to cause the catastrophe. He used these assumptions to calculate the median life expectancy of a civilization, which happens to be from hundreds to thousand years. But the assumption of the non-progress and stable number of agents is unfounded as technological civilizations tend to quickly evolve – or, if they fail to evolve new technologies, they will be stuck by resource depletion. Turchin et al suggested in the article about multipandemic (Turchin, Green, & Denkenberger, 2017) that a global catastrophe could happen from the actions of many agents which release simultaneously many non-catastrophic viruses, if the number of agents will increase very quickly.

Bostrom recently wrote about "Vulnerable world hypothesis", where the catastrophe is a "black ball" which an unlucky civilization gets as a result of unexpected technological development

(Bostrom, 2018). Manhiem suggested that not a single black ball, but the accumulation of “grey balls” may result into the existential catastrophe (Manheim, 2018b).

Kurzweil wrote that exponential progress in the new technologies is driven by the “law of the accelerating returns” (Kurzweil, 2006), and the whole group of NBIC (nano-info-bio-cogno) technologies is developing at the speed close to Moore’s law with doubling time around 2 years. The progress in the biotech could be measured by lowering the prices of DNA sequencing which even outperform Moore’s law (Honorof, 2013). The progress in the computation and miniaturization is fueling the growth of all NBIC sector.

The doubling period in electronics has declined from 3 years at the beginning of the 20th century to 2 years in the middle of the 20th century, which may be evidence of the double exponential nature of Moore's law at longer distances (Korotayev, 2018). While original semiconductor-related Moore’s law may soon hit the wall in the chip lithographic technologies, the advances of computational hardware is continuing via appearing of specialized chips for AI (Graphcore, 2017). This AI-related hardware is claimed to accelerate above Moore’s law according to the newly coined Huang law which imply 4-10 improvement in GPU in 1 year (Perry, 2018). We could assume that the technological progress in AI and biotech will continue for at least a few decades from now based on a large number of unexplored technological opportunities connected with miniaturization, genetic experiments and AI.

The large number of possible technological risks in combination with exponential technological growth suggests that some catastrophe could happen in a relatively short timescale, even before the “technological singularity” which appears in some models of the future. In this article, we create a general model of growth of the catastrophe’s probability with the technological progress (section 2), apply it to actually existing risks to get some probable time estimates (Section 3), explore the chance that the global risk could be self-limiting (Section 4) and discuss the ways to escape the technological explosion curse via adequate control systems in the Section 5.

2 The model of the catastrophe depending on the number of agents and the speed of the technological progress

2.1 A case of just one dangerous technology and no technological progress

Imagine a world where a few groups of individuals have access to some potentially dangerous technology. This could be several labs in case of biorisks, or several countries owning nuclear weapons.

Let’s p be the probability that a catastrophe is produced by a group A in a time unit t. Then $1 - p$ is the probability that no such disaster is produced by that group during that time unit. If the number of groups is n, and T is the total number of time units then:

$(1 - p)^n$ is the probability of no disaster in one time unit given n groups (assuming uniformity of p across all groups).

$(1 - p)^T$ is the probability of no disaster produced by any group over T time units assuming uniformity of p for that group across the time T.

$S = (1 - p)^{nt}$ is the probability of no disaster given n groups over time T, that is, the probability of survival. In that case, P_{cat}, the accumulated probability of the catastrophe from one technology in the case of a constant number of research groups is:

$P_{cat} = 1-(1-p)^{nt}$ (1)

This equation is also present in Sotos article, where he applies it to the hypothetical static extraterrestrial civilizations (Sotos, 2017).

For example, in the case of just one risk with $p = 0.1$, one agent and risk period of 1 year (Equation 1), the probability of catastrophe reaches 0.95 in 28 years. (Figure 1). We should mention that the human extinction risk becomes unacceptable long before it is inevitable, and even 10 percent of the risk is above the acceptable level.

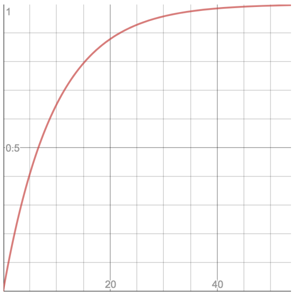

Figure 1. The plot of the model in equation (1) with P = 0.1 and a constant number of agents. The figure is generated on the site using the following equation: 1-(0.9)^(x)

If there are 100 groups or agents, and the risk is 0.00001 per group per year, the 95 percent confidence of extinction will be reached in 3000 years, but 10 per cent will be reached only in 100 years.

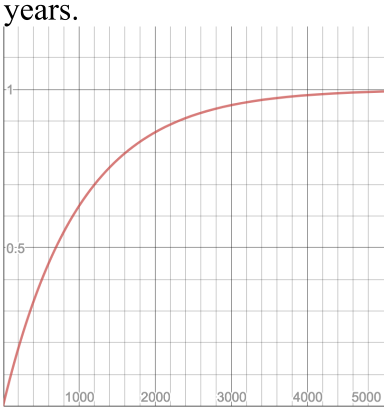

Figure 1a: 1-(0.99999)^{100x}

2.2 A case with an exponentially growing number of dangerous agents

Now we will explore what will happen if the number of potentially dangerous agents is growing exponentially. The reason for such growth is the exponential decline of the price of potentially dangerous hardware, like DNA manipulating technology, driving by exponential technological progress. Cheaper hardware could be available for a larger group of people. So, the number of agents is growing exponentially:

$n(t) = ae^{kt}$

Thus, the probability of the survival is decaying roughly double exponentially, if we account for the increase of the number of the technology owners:

$P_{cat}(T) = 1 - \prod_{t=0}^{T=m}(1-p)^{amte^{kt}}$ (2)

(or $P_{cat} = 1 - (1-p)^{ate^{kt}}$ (2')

- I still don't understand from where the "product" sign appears here and why we can't just substitute the n(t) function into the equation (1)?)

The example graph of the equation (2') where the number of agents is doubling every 2 years, assuming some form of Moore's law, is on the figure (2) – and in it the catastrophe becomes inevitable in a period of around 12 years.

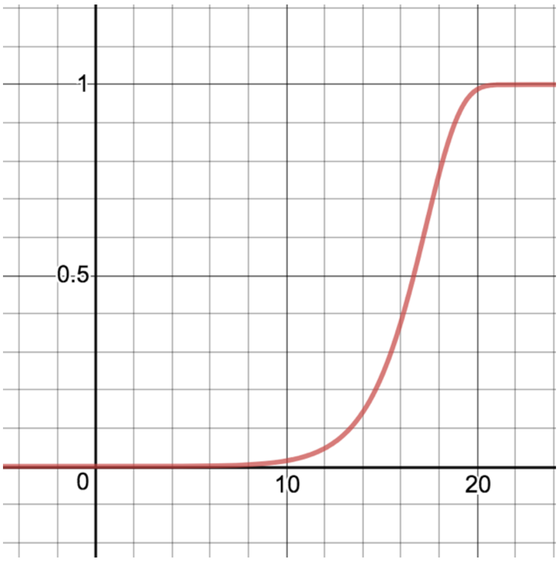

Figure 2, 1-(0.99999)^{xe^{0.5x}} at desmos.com

Interestingly, if we change the initial probability *p* on one order of magnitude, like from 0.0001 to 0.00001, it only moves the whole graphic right on 5 years.

However, we also may take into account here the Pareto distribution of wealth, as the cheaper hardware could be available to much larger groups of people. Pareto distribution is a power law with some coefficient *alpha*. For US now, the coefficient is around 1.59 (Vermeulen, 2018). It means that the lowering of the price of some technological artefact 2 times will increase the number of people who could buy it in 3.01 times. This should be accounted in our equation of the growing number of agents.

$\frac{w}{w} = (\frac{n}{n_0})^{\alpha}$, so $n(t) = (price(t))^{\alpha} = (ae^{kt})^{\alpha} = ae^{\alpha kt}$

Including the Pareto distribution is not changing the type of dependence, it is still roughly double exponential.

2.3 A case where the technology becomes more dangerous in time

The probability that the given piece of technology will result in a global catastrophe is also growing with time; thus, *p* should be *p*(*t*), because the technology itself is evolving.

$p(t) = be^{ce^{dt}}$

(Should it be presented in the way in which it can't grow above 1?
Or should we just say that as p is very small, we should approximate it as exponentially growing?)

where *c* and *d* are some constants.

The first part be^ has to do with the difficulty of producing a dangerous event given technological "distance" and the second part ce^(dt) has to do with how this distance is decreasing with time because of the fast progress which is assumed to be exponential (discussed in the next subsection – is it the same idea that doublings in Moore's law are becoming quicker?) because we are assuming there is an exponentially growing number of researchers who are also making a constant amount of progress per time unit per researcher.

Obviously, it is an oversimplification, as some internal control inside an agent will not allow starting a technology which will almost sure wipe out its creator. But on early stages exponential growth of the risk of the technology may be always unnoticeable – or there will be an incentive for an agent to take the risk, even if the agent knows about it, as the agent expects to get an advantage over rivals or have other egoistic cost-effective calculations. For example, when the first nuclear bomb was detonated, there was some probability estimations that it will start the nuclear chain reaction in the atmosphere, but the scientists decided to proceed anyway, as they concluded that Soviets will do the same experiments eventually.
Therefore,

$$P_{cat}(T) = 1-\prod_{t=0}^{T=m}\left(1-be^{ce^{dt}}\right)^{amte^{kt}} \qquad (2)$$

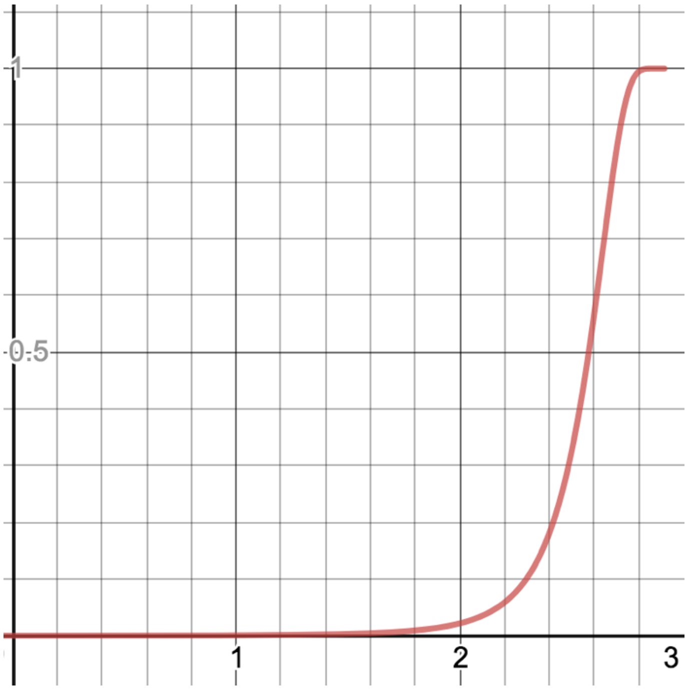

Figure 3. *1-(1\ -0.0001e^{0.5e^x})^{xe^{0.5x}}*
Here we see that the catastrophe becomes almost inevitable in just one doubling.

2.4 A case where technological progress is accelerating

As we mentioned above, the doubling period is also compressing due to technological acceleration. The Moore's law has been accelerating from a doubling time of 3 years at the beginning of the 20th century to 1 year for progress in neural nets in the 2010s and DNA sequencing price in 2000s, implying even quicker growth. The growth in compute for neural nets has now doubling time of 3.5 months, implying order of magnitude growth in a year, and NVIDIA CEO Huang declared "Huang's law" (Perry, 2018) according to which the performance of Graphics Processor Units will from 4-10 times a year, but this law is not taking into account the price performance.

2.5 A case where the number of the types of possible catastrophic risks is growing

Another factor which should be included in the model is the exponential growth of the number of types of possible global risk—the first plausible idea of the technological catastrophe appeared during the dawn of nuclear age, currently we know around 10 main types of hypothetical existential risks, each have a few subtypes, and the number of known risks is only growing. This means that many equations like (2) should be calculated simultaneously for each risk and then aggregated.

If we assume that risks are not mutually exclusive, it simplifies the equation. In that case, the total probability of survival of several independent risks (like nano, bio, nukes, AI) will be:

$P_{cat}(T) = \sum P_i$ (2) Is sum enough? Or we need (1- product)?

Different such risks may have different parameters of the speed of growth, for example, nuclear technology is growing slower. If we assume that all risks have the same growth parameters when each year only the number of risks will be growing, this simplifies the equation:

????

We also didn't include in the model the possibility of the risk interactions, which were described by Baum (Baum, Maher, & Haqq-Misra, 2013) and Tonn (Tonn & and MacGregor, 2009) and Hanson (Hanson, 2008). NBIC-convergence also means that the technological progress in different technologies is fueling one another, as, e.g. progress in biotech allows better nanotech in form of, say DNA-origami, and the progress in AI allows quicker progress in all other technologies as it accelerates the speed of the technological discoveries.

2.6 Exploring the properties of the model

The main feature of the equation (2) is that it grows very quickly from just above 0 to almost 1. Even if some exponential assumptions are wrong and are better modelled by linear terms, the probability will still grow very quickly.

In practice, this means that the situation of almost no risk to almost inevitable catastrophe in just a few doublings of exponential growth. An example of such quick growth is the history of computer viruses in the 80s which become almost ubiquitous just in a few years.

Now let's look at the behavior of our equations.

Table 1. The increase of the probability of catastrophe depending of time and the law

Number of technological doubling (possible years for illustration only)	Exponential growth of cumulative probability based on no progress (Sotos model) *$1-(0.99999)^{100 x}$*	Equation 1' *$1-(0.99999)^{te^{0.5t}}$*	Illustration of the speed of growth of triple exponential function $2^{2^{(2^x)}}$	*Equation 2'* *$1-(1\ -0.0001e^{0.5e^t})^{te^{0.5t}}$*
d=0 (2020)	1	0.014 (t=10)	1	0.00016 (2020)
d=1 (2022)	0.001	0.047 (t=12)	16	0.00064 (2021) 0.0217 (2022) 0.33 (2022.5) 1 (2022.8)
d=2 (2024)	0.002	0.14 (t=14)	65536	
d=3 (2026)	0.003	0.37	10^{77}	
d=4 (2028)	0.004	0.76	10^{19278}	
d=5 (2030)	0.005	0.987	$10^{(10^{9.11})}$	
d=6 (2032)	0.006	0.999998		
d=100 (2230)	0.95	1		
Number of doublings to inevitable catastrophe	**100** (200 years)	**5** (10 years)		**1** doubling (2 years)

We could see that in the case of equation 2' which is close to double exponential, the catastrophe will grow from negligible to inevitable in just 10 years if Moore's law doubling remains 2 years (other initial conditions define the moment of the start of the quick growth). This means that all catastrophic risks condense into one "dangerous decade". Equation 2' gives even a quicker growth with step-like function, where a jump of probability happens in only one doubling or 2 years.

Interestingly, another double exponential process is human aging described by Gompertz curve https://en.wikipedia.org/wiki/Gompertz_function This should not be surprising as it has approximately the same dynamics: an exponentially growing number of possible diseases each of which is becoming more exponentially more dangerous in time because of the weakening of the body defense systems.

3 Applying the model to the currently existing technologies

Above, we look at the model on the purely theoretical grounds, but there two main technologies which are currently growing with Moore's law-like speed and could present global catastrophic risks. It is biotech and AI research. (There are other exponential technologies, which could present global risk, but currently they are relatively weak compared to these two. These other technologies may include nanotech and hypothetical "cheap nukes".)

3.1 Democratization of synthetic biology

There is a growing movement of the biohackers who are experimenting with changing the DNA of living organisms. DIY packages for genetic modification are available as well as many other ingredients, information and service which could be ordered from the internet. There is a growing trend of merging between computers and home biolabs (e.g. digital to biological converter by Craig Venter (Boles et al., 2017)).

The progress in biotech is even quicker than Moore's law. In two decades, the price of genome sequencing failed tens millions of times. The availability of knowledge and instruments, as well as the demand for illicit drugs or self-augmentation, may fuel the biohackers movement similar to the computers hackers who appeared in 1980s and eventually started to create computer viruses – first for experimenting and later for revenge and income. The number of computer viruses grew 1000 times in the 1980s starting from one in a year, and continue to grow with a million pieces of malware a day 2010s (Harrison & Pagliery, 2015).

The most possible biological catastrophes are not extinction risks but smaller catastrophes.

3.2 Creation of non-aligned AI

The number of people enrolling in machine learning courses is growing approximately 10 times a year from 2015. The field is advancing and more powerful computers are available individually or for rent. Currently, most AIs are narrow and non-agential, but soon the boom in robotic minds will start, fueled by home robotics, drones and self-driving cars. This increases chances that someone will create self-improving AI, which quickly gains capabilities and may have non-human-aligned goals. The probability of such an event may be distributed unevenly between actors.

4 Could smaller catastrophes prevent the large one?

The model presented above suggest that there are a few main ways to prevent a global catastrophe:

- stop the technological progress
- lower the number of independent agents
- lower the access of the potentially dangerous agents to the dangerous technologies via some drastic control measures.

W. Wells (Wells, 2009) suggested that smaller catastrophes are more probable that larger ones approximately 2-3 times and thus non-extinction level catastrophe is the more probable outcome of the civilizational development.

We will define here "smaller global catastrophe" as the one which is capable of stopping technological progress, but not cause human extinction. Similar catastrophes were called "W-risks" by Nell Watson (Watson, 2018), as the civilization may not be capable to restart because of lack of easily accessible resources. In our case, it doesn't matter if the civilization will be able to rebuild itself or not. We look at the catastrophes that are capable of stopping quick technological progress for a considerable amount of time.

The self-limiting factors were explored in case of a pandemic, where increase distancing and death of super-spreaders could limit the impact (Manheim, 2018a).

5 Global control system for the prevention of the risks' explosion

The obvious way to prevent the catastrophic outcome according to the model is to lower the number of the potentially dangerous agents which have access to the risky technology as well to lower the risks presented by any separate piece of technology.

To lower the number of agents and to increase the safety of technology, some form of a *global control system* is needed, which will license only a finite number of agents the access to the dangerous technology and ensure that these agents are using the technology according to the established safety protocols, or not using it at all. Such control system needs to be:

Global – that is, to cover all the surface or the earth and all human space colonies if any appear, without any "excluded territory", as such territory would attract potentially dangerous research.

Intelligent – the system should be able to establish the best safety rules, and also able to recognize any potentially dangerous activity on early stages.

Powerful – the system should include some law enforcement agency.

There are three hypothetical ways how such a global control system could appear:

1. But the best candidate for such control system is Superintelligent AI based Singleton. However, the creation of the first AGI and establishment of its global domination come with its own significant risks.

2. Some national states are able to provide the needed level of control inside their border. UN, if it becomes more powerful, could work as a "global state". This could happen if smaller catastrophes will demonstrate the risks of the uncontrollable supertechnologies and different counties delegate their power to something like "global risk prevention committee".

3. Also, one country could get a decisive strategic advantage over other counties using the fruits of some of the supertechnologies, and establish its global rule which will also include control over possible risks. Such domination may be based on narrow AI advantage, on biotech or nuclear power, but the process of the establishment may look like a world war with all its moral costs and global risks.

What else?

References

Baum, S. D., Maher, T. M., & Haqq-Misra, J. (2013). Double catastrophe: intermittent stratospheric geoengineering induced by societal collapse. *Environment Systems & Decisions*, *33*(1), 168–180.

Boles, K. S., Kannan, K., Gill, J., Felderman, M., Gouvis, H., Hubby, B., … Gibson, D. G. (2017). Digital-to-biological converter for on-demand production of biologics. *Nature Biotechnology, 35, 672–675 (2017).*

Bostrom, N. (2002). Existential risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology, Vol. 9, No. 1 (2002).*

Bostrom, N. (2018). The Vulnerable World Hypothesis, 38.

Graphcore. (2017). Graphcore's technology for accelerating machine learning and AI. Retrieved January 24, 2018, from https://www.graphcore.ai/technology

Hanson, R. (2008). Catastrophe, social collapse, and human extinction. In N. Bostrom & M. M. Cirkovic (Eds.), *Global catastrophic risks (p* (p. 554). Oxford: Oxford University Press.

Harrison, J., & Pagliery, J. (2015). Nearly 1 million new malware threats released every day. *CNN*. Retrieved from http://money.cnn.com/2015/04/14/technology/security/cyber-attack-hacks-security/

Honorof, M. (2013). Biotech's Explosive Evolution Outpaces Moore's Law - Technology & science - Tech and gadgets - TechNewsDaily | NBC News. Retrieved from http://www.nbcnews.com/id/51870335/ns/technology_and_science-tech_and_gadgets/t/biotechs-explosive-evolution-outpaces-moores-law/

Korotayev, A. (2018). The 21st Century Singularity and its Big History Implications: A re-analysis. *Journal of Big History*, *2*(3), 73–119. Retrieved from https://www.researchgate.net/publication/325664983_The_21_st_Century_Singularity_and_its_Big_History_Implications_A_re-analysis/references

Kurzweil, R. (2006). *Singularity is Near*. Viking.

Manheim, D. (2018a). Self-Limiting Factors in Pandemics and Multi-Disease Syndemics. *BioRxiv*, 401018. https://doi.org/10.1101/401018

Manheim, D. (2018b). *Systemic Fragility as a Vulnerable World*. Retrieved from https://philpapers.org/rec/MANSFA-3?fbclid=IwAR1NLIeco9nCaNGKPca3p62nrWzFJFlL7AtHdm3nrysGV5yuAddVmRgbM2o

Perry, T. S. (2018, April 2). Move Over, Moore's Law: Make Way for Huang's Law. *IEEE Spectrum: Technology, Engineering, and Science News*. Retrieved from https://spectrum.ieee.org/view-from-the-valley/computing/hardware/move-over-moores-law-make-way-for-huangs-law

Sotos, J. G. (2017). Biotechnology and the lifetime of technical civilizations. *ArXiv Preprint ArXiv:1709.01149*.

Tonn, B., & and MacGregor, D. (2009). A singular chain of events. *Futures*, *41*(10), 706–714.

Torres, P. (2016). Agential Risks: A Comprehensive Introduction. 2016. *Journal of Evolution and Technology -*, *26*(2).

Torres, Phil. (2018). Facing disaster: the great challenges framework. *Foresight*.

Turchin, A. (2015). Typology of global risk. Retrieved from http://lesswrong.com/lw/mdw/a_map_typology_of_human_extinction_risks/

Turchin, A., Green, B., & Denkenberger, D. (2017). Multiple Simultaneous Pandemics as Most Dangerous Global Catastrophic Risk Connected with Bioweapons and Synthetic Biology. *Under Review in Health Security*.

Vermeulen, P. (2018). How fat is the top tail of the wealth distribution? *Review of Income and Wealth*, *64*(2), 357–387.

Watson, N. (2018). The Technological Wavefront — Nell Watson. Retrieved December 16, 2018, from https://www.nellwatson.com/blog/technological-wavefront

Wells, W. (2009). *Apocalypse When?: Calculating How Long the Human Race Will Survive*. Praxis. Retrieved from //www.springer.com/us/book/9780387098364