Assessing the future plausibility of future catastrophically dangerous AI

Alexey Turchin

Science for Life Extension Foundation, Moscow

Prospekt Mira 124-15, Moscow, Russia, 129164.

[alexeiturchin@gmail.com](mailto:alexeiturchin@gmail.com)

**Abstract**:

In AI safety research, the *median* timing of AGI creation is often taken as a reference point, which various polls predict will happen in second half of the 21 century, but for maximum safety, we should determine the *earliest* possible time of *dangerous* AI arrival and define a *minimum acceptable level of AI risk*. Such dangerous AI could be either narrow AI facilitating research into potentially dangerous technology like biotech, or AGI, capable of acting completely independently in the real world or an AI capable of starting unlimited self-improvement. In this article, I present arguments that place the earliest timing of dangerous AI in the coming 10–20 years, using several partly independent sources of information: 1. Polls, which show around a 10 percent of the probability of an early creation of artificial general intelligence in the next 10-15 years. 2. The fact that artificial neural network (ANN) performance and other characteristics, like number of “neurons”, are doubling every year, and extrapolating this tendency suggests that roughly human-level performance will be reached in less than a decade. 3. The acceleration of the hardware performance available for AI research, which outperforms Moore’s law thanks to advances in specialized AI hardware, better integration of such hardware in larger computers, cloud computing and larger budgets. 4. Hyperbolic growth extrapolations of big history models.

**Keywords**: artificial intelligence – existential risks – singularity – near-term risks – Moore’s law

**Highlights**:

* The median timing of the AGI predictions is the wrong measure of AI risk assessment; we should estimate earliest plausible time of AGI’s creation and a *minimum acceptable level of AI risk*.
* A dangerous AI level is defined through the AI’s ability to facilitate a global catastrophe; this could be an AGI able to act in the real world independently, or to self-improve, or to facilitate the creation of other dangerous technologies.
* The growth rate of hardware performance for AI applications has accelerated in the past five years and could provide enough computational power for near-human capabilities in the next decade.
* The main measures of neural nets performance have been doubling every year since 2012 and if this trend continues, they will reach near-human level less than a decade.
* Several historical trend extrapolation methods predict near-human-level AI in one or two decades.

[1 Introduction 2](#_Toc531528194)

[2 Dangerous AI and declining computational complexity of human omnicide 3](#_Toc531528195)

[3 Earliest and median time of dangerous AI creation from the point of view of risk analysis 4](#_Toc531528196)

[4 Different ways to estimate the earliest possible creation of dangerous AI 5](#_Toc531528197)

[5 AI and hardware projection for next 1-2 decades 5](#_Toc531528198)

[5.1 AI-related hardware progress 5](#_Toc531528199)

[5.2 Hardware as proxy of AI performance 6](#_Toc531528200)

[5.3 “Dangerous” level of AI hardware 7](#_Toc531528201)

[5.4 Future of Moore’s law and new ways to overcome its limitations via multi-chip processors and lower per-computation price 7](#_Toc531528202)

[5.5 AI-related “Moore’s law”: growing budget for AI research helps outperform Moore’s law 9](#_Toc531528203)

[5.6 Conclusions about hardware 10](#_Toc531528204)

[6 Progress in neural net performance 10](#_Toc531528205)

[6.1 Neural nets revolution dramatically increases the speed of AI progress 10](#_Toc531528206)

[6.2 Neural nets’ performance metrics have doubled every year since 2012 10](#_Toc531528207)

[6.3. Neural nets’ size grew 100 times in last five years 11](#_Toc531528208)

[6.4 As training data sets grow, neural net performance increases logarithmically, and a human-size dataset may be able to provide human-level performance 11](#_Toc531528209)

[6.5 Signs of self-improving and data transfer in ML 12](#_Toc531528210)

[6.6 There are other fields in AI research that could get a boost in the future 12](#_Toc531528211)

[6.7 Evolution of the “intellect stack” 12](#_Toc531528212)

[6.8 Jumps in algorithmic efficiency in neural networks 12](#_Toc531528213)

[6.9 Different attempts to predict human-level AI timing in neural nets converge around the same date 13](#_Toc531528214)

[7 Surveys of experts 13](#_Toc531528215)

[8 All hyperbolic future growth projections converge in the next two decades 14](#_Toc531528216)

[9 AI timing predicted based on randomness of the moment of AI creation 15](#_Toc531528217)

[9.1 The growth of the probability to find ground breaking idea 15](#_Toc531528218)

[9.2 AI timing prediction based on my random location between beginning of the research and its end 16](#_Toc531528219)

[10 Combining all projections 16](#_Toc531528220)

[11 Uncertainty in AI predictions 17](#_Toc531528221)

[12 What could be done to ameliorate risks of early AI? 17](#_Toc531528222)

[Conclusion 18](#_Toc531528223)

# 1 Introduction

In 2017 the Machine Intelligence Research Institute (MIRI) updated their strategic view and assigned a higher probability to the idea that artificial general intelligence (AGI) will be created in next two decades (MIRI, 2017), without providing any numerical estimates. This is a change from its previous assessments that creation of AI is very difficult and will not be created for a few decades. The change is a reaction to the recent acceleration of AI progress, related hype, growth of funding, and impressive results.

However, Max Tegmark (2017) constructed a map of what the AI safety community believes about AI risk timing, and in it “virtually nobody” expects superhuman AI in the next few years. Shakirov (2016) estimated that AGI could be created earlier based on his extrapolation of the artificial neural networks (ANNs) progress; he estimated that it could happen as in the next decade. Turchin and Denkenberger explored the global catastrophic risks of non-superintelligent AI, which could happen before real superintelligence will be created, and thus appear earlier in AI-risk timelines (Turchin & Denkenberger, 2018a). Paul Christiano explored the idea of the “mundane singularity” (Christiano, 2016) based on so-called messy AI or KANSI – *known-algorithm non-self-improving agent* (Arbital, 2016), an AI created from known modules without any additional insights. He recently said, “I think that my probability for human labor being obsolete within 10 years is probably something in the ballpark of 15%, and within 20 years is something within the ballpark of 35%” (Perez, 2018).

Many public figures have started to speak out about the risks of AI, including Elon Musk, Vladimir Putin, Bill Gates, Stephen Hawking, and Hillary Clinton. Cyberwarfare was probably used to affect elections in the U.S., which increase the probability of a global catastrophe, according to some experts (Granoff, 2016; P. Torres, 2016).

There have been many recent attempts to estimate AI timing. Grace performed a large poll of scientists (Grace, 2017b). The author of the Bayesian Investor blog analyzed only the effects of hardware improvements (Bayesian Investor Blog, 2017). He assumed that only hardware progress is important and given median estimates of the human brain power and willingness to pay for AI, the necessary hardware is already available, but longer training, possibly on the scale of years, may be needed. Thus, he stated, “[t]his analysis suggests that the probability of human-level AGI being reached in any given year has become nontrivial now, will reach a peak in the mid to late 2030s, and if it isn’t reached by 2100, then my approach here will have been mistaken”.

In his article “There’s No Fire Alarm for Artificial General Intelligence”, Yudkowsky (2017) suggests that predicting the advent of AGI is almost impossible, and we may not have good signs of AGI until two years before the event. Alpha Zero’s rapid self-improvement and ability to win games in many domains was close to such a “fire alarm” event (Silver et al., 2017).

*OpenAI* recently published an article about the increase of compute in training of the biggest neural nets, which shows a 3.5-month doubling time and 300 000-fold increase after 2012 (OpenAI, 2018). Ilya Sutskever, research director of OpenAI said in November 2018: said “We (OpenAI) have reviewed progress in the field over the past six years. Our conclusion is near term AGI should be taken as a serious possibility” (Peng, 2018). This clearly demonstrates that we are living in a period of explosive growth, which can’t be exhausted soon by a lack of available hardware; as it also grows quicker than Moore’s law, according to the newly coined “Huang’s law”, which predicts a 4–10 times increase of power of AI-related hardware per year (Perry, 2018).

The observed acceleration of AI development and large uncertainty about AGI timing suggest that it is important to explore the probability of the creation of AGI in the near future, but even here we can distinguish two time periods with important practical differences between them: 1) near-term: less than a decade, and 2) medium-term: one-two decades.

According to the MIRI strategy, an earlier horizon for the creation of AGI means that more attention should be focused on collaboration with existing institutions and on analysis of pre-existing AGI technologies (MIRI, 2017). These suggestions are in line with estimations of the risk of medium-term AI. However, if AGI is a near-term risk there is little or no time to create, promote and implement perfect AGI safety theory, and some other solutions should be explored, as we have previously discussed (Turchin & Denkenberger, 2017a).

It is known that predicting AI is very difficult (Armstrong & Sotala, 2015), and predicting the timing of AI’s emergence is one of the most heated topics, where opposite opinions are often presented. Some have argued against the idea of near-term AI risks as dangerous and misleading, providing an estimate of the chances of dangerous AI emerging in the next decade at less than 1 percent (Baumann, 2018).

In this article, we do not try to predict the median time of AGI creation, but present arguments that an earlier creation of dangerous AI is quite possible for several distinct reasons, and that it has important practical implications for AI safety.

# 2 Dangerous AI and declining computational complexity of human omnicide

To escape theoretical discussions of AI completeness (having general intelligence), we introduce here the notion of *dangerous AI*: AI powerful enough to create global catastrophic risks (GCR). Sotala called this ability major decisive advantage (Sotala, 2018). It may or may not be AGI, and it may even not be able to pass Turing test. It most likely will be close to human-level, but it could formally fail a Turing test based on lack of some human abilities, like consciousness. This lack of some human features might make it even more dangerous. Such AI may also be able to perform full or limited self-improvement, but again, this is not critical to its definition. A similar idea is Karnofsky’s proposal of “transformative AI”, defined as “AI that precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution. I believe there is a nontrivial likelihood (at least 10% with moderate robustness, and at least 1% with high robustness) that transformative AI will be developed within the next 20 years” (Karnofsky, 2016).

There are many ways for AI to become dangerous and cause a global catastrophe, and in another work the author identified and classified a few dozens of them, but most start to be serious when AI reaches mostly human capabilities (Turchin & Denkenberger, 2018a).

One such scenario is that AI will become dangerous when it has the capability to act independently in the wild (probably in the Internet) and perform better than humans in most human tasks. Such a capability will probably be based on the ability to create powerful world models and natural language processing (NLP). Thus, measuring progress in NLP and world modeling in AI, we could estimate the creation time of such systems.

Another type of dangerous AI is that could help to facilitate dangerous research in fields such as biotech or increase the effectiveness of existing military technologies.

The third, and most important, type of dangerous potential AI is one which can start an intelligence explosion via self-improvement. Such an AI could facilitate AI-related research in large organizations; this phenomenon is already occurring with system like AutoML and the use specially designed hardware at Google. Sotala explored ways that such an infra-human mind could experience quick capability gains and become superintelligent (Sotala, 2017); the author explored various ways of self-improvement in (Turchin & Denkenberger, 2017b).

Many different types of potentially dangerous AI are clustered around roughly human-level AI, that is, AI capable of performing most human tasks. As a result, we could measure the distance to dangerous AI via AI’s ability to perform most human work and without requiring it to be “AGI-complete” or “conscious”.

# 3 Earliest and median time of dangerous AI creation from the point of view of risk analysis

Grace et al conducted a large poll of experts about the timing of AGI emergence (Grace, 2017b). The experts’ projections of the creation of human-level AI are distributed almost linearly in the beginning of the probability graph between 5 and 25 years from now, growing at a rate of approximately 1.25 percent a year. This means that experts estimate a 6.25 percent chance of AI before 2022, 12.5 percent before 2027, and almost 15 percent before 2030. One may suggest these data be ignored as noise, but such a discounting seems arbitrary; it may be more prudent to take the data at face value.

AGI predictions are often analyzed to determine a median time of predicted AGI creation. However, from the risk analysis point of view, we need not a median time, but the earliest time of AI creation. It can be understood by analogy: we need to know not the median time that a bridge will collapse, but the earliest possible time of such failure. Half of all bad events are likely to happen before the median time, so if we prepare AI safety theory for the median timing of AGI creation, we will be dead in 50 percent of cases—assuming that any non-aligned superintelligence is a global catastrophic risk, which may be less likely given the small utility of killing humans (Turchin, 2017). However, if we accept this point of view, we will have very small time to prepare for AGI creation, which may be the reason for its unpopularity.

Thus, we need not median timing AGI creation, but a *minimum acceptable level of AI risk*. Kent showed (2004) that even the smallest risk of a global catastrophe is unacceptable. But in our case, if we assume a linear distribution, the probability of AGI happens tomorrow are 0.003 percent, already too high by Kent’s metric. However, little can be done about it, as rushed actions may even impede the slower efforts which will produce the largest aggregated probability of survival. That is, quick attempts to ban AI will likely fail and only damage long-term projects on AI alignment.

We suggest such a threshold of *minimum acceptable level of AI risk* as five percent of the *cumulative* probability of powerful AI’s creation in a given time frame (either as AGI or powerful narrow AI to create global catastrophic risk; see next section). Note that we use term “probability” in Bayesian terms, as a measure of our expectations based on available information. Also consider that the yearly probability of full scale nuclear war is around one percent (Barrett, Baum, & Hostetler, 2013), with a large margin of uncertainty, so it seems that a comparable level of prevention efforts is needed. But the annual probability of nuclear war is assumed to be constant, while the probability of AI is quickly growing.

Applying our *minimum acceptable level* threshold to the Grace poll data, we obtain a 5 percent chance of AGI creation in the next few years. We assume that human-level AI will pose an existential risk as soon as it will be created, but it may actually take years for AI to evolve into dangerous superintelligence.

# 4 Different ways to estimate the earliest possible creation of dangerous AI

There are several distinct ways to estimate the timing of AGI, and some of them have been covered extensively (Constantin, 2017; Grace, 2017b; Kurzweil, 2006; Vinge, 1993). They are not completely independent, as polls are affected by observed technological trends. These ways will be explored in the next sections, 5-10:

* extrapolation of the technological trends in hardware (section 5)
* growth in problem solving performance (section 6)
* polling of experts (section 7)
* historical analogues (section 8)
* analysis of general laws of acceleration of history (section 9)
* use of the ideas of randomness of the moment of AI creation (section 10)

We are most interested by the question: will different ways of prediction produce similar results and what they claim about possible timing of the coming of the dangerous AI which is able to reach computational complexity of the omnicide.

# 5 AI and hardware projection for next 1-2 decades

## 5.1 AI-related hardware progress

There are two competing views on the speed of the technological progress in hardware. One view is that Moore’s Law no longer holds, and that if we calculate the actual growth of flops per dollar in GPU, we could see very unimpressive numbers on the one order of magnitude of growth within 10–16 years (Grace, 2017a).

Another view is that we should look at the hardware available for AI research, which has grown very impressively, as demonstrated by the OpenAI review which found a doubling period of 3.5 months, or one order magnitude in a year (OpenAI, 2018), and that such growth could continue for at least 5 years.

The main reason is that AI related computing is not the same as the abstract quantity of flops available per dollar. AI-related computing growth is much quicker than this abstract power, since it uses different routes to achieve it:

First, GPUs were almost not used in AI research until 2012, when it became mainstream.

Second, AI research budgets have grown.

Third, the size of the best available computers for AI research has increased. Even if the price of computational power had remained the same, the shorter time required for experiments on larger computers results in economy of other resources—first among them, human salaries—and obviously in accelerated testing of new ideas. Quicker experiments also solve the problem of obsolesce of the computers, which became morally obsolete rapidly, typically in a few years. If one spent two months on one experiment, the computer will become obsolete after 20 experiments, so the researcher would need to account more carefully for amortization of the hardware in the computational costs of the experiment.

Bigger computers also mean not only quicker results, but also that experiments which were not possible at all on smaller computers can be performed; supercomputers are typically designed for quicker exchange data between nodes.

Cloud computing also lowers the price of ownership, as researchers become able to pay for computational power only when it is needed. This solves the problem of costly hardware down-time. The simplification of AI-related calculations and growth of the number of threads for one processing core has also increased the capabilities of AI-related hardware.

AI-related hardware has already passed through several epochs:

1. CPU-epoch: before 2012.
2. GPU-epoch: 2012–2016. During this epoch, AI-related computations jumped to already extant graphics cards, which provided growth of 1–2 orders of magnitude.
3. TPU-epoch: 2016-now. This epoch was pioneered by the Google Tensor processing unit, specialized hardware for AI (Sato, 2017). NVIDIA immediately added tensor cores to its graphic cards in 2017 (Smith, 2017), and many startups jumped in to work on specialized “Intellectual Processing Units” (IPU), which promise to multiply larger matrices, providing a large reduction in computational costs. The startup Graphcore promises a 100× increase over current performance (Graphcore, 2017). While we can’t be sure that any startup will actually deliver, we could see their projects as anticipation of what will eventually appear in the industry. “Graphcore provided a glimpse of its 23.6 billion-transistor Colossus that aims to hold an entire neural-net model in its 300 Mbytes of on-chip memory. The startup claims that it can process 7,000 programs in parallel on its 1,216 cores, each capable of 100 GFlops” (Merritt, 2018), which roughly equivalent to the 100 petaflops performance of a single chip.
4. Neuromorphic chips epoch: from 2019? Neuromorphic chips promise to solve the “von Neumann” bottleneck, that is, latency because of the need to download data from memory. The most computationally effective approach seems to be spiking neural net chips. The main problem is how to convert existing neural net programs into spiking neural net hardware. For example, the Akida chip “packs 1.2 million neurons and 10B synapses in an 11-layer SNN along with a RISC processor to work its magic – which should be up to 1,400 frames per second per watt. Those are impressive numbers – particularly for a $10 chip” (Morris, 2018) and they could be connected up to 1000 chips.
5. In the future: Something currently unknown, like quantum computers or “in-memory computation”.

## 5.2 Hardware as proxy of AI performance

There are several reasons why hardware growth is a good proxy of AI performance, and that when AI reaches an approximately human hardware level, it will also will have ability to solve most of the tasks that humans are able to solve:

- The current explosion in AI became possible after the appearance of cheap computational power from GPUs. Hinton named the availability of cheap computation power one of the few reasons for the current neural net revolution (Chklovski, 2017). Growth in hardware allows quick and cheap testing of new ideas in AI science.

- Humans have highest concentration of neurons in the cerebral cortex among other animals (21 billion), and only finned whale have an even larger neuron number (37 billion) (Mortensen et al., 2014); the growth of performance in animals strongly correlates with the size of the cortex. See also discussion by Cannell (Cannell, 2015) that if the brain is a “universal learning machine”, then hardware equivalence is more probable to deliverer full AGI.

- Experiments with very large neural nets demonstrated performance grows logarithmically with the size of the net, and such growth outperform gains from architectural complexity. Google reached a state-of-the-art result with a 100 billion parameter (parameter = synapse) net in 2016 and is using it now for machine translation in Google Translate (Shazeer et al., 2017) (more about it below); they have since started to experiment with a trillion parameter net. The adult human brain has 100-500 trillion synapses (Drachman, 2005), a number roughly equivalent to the neural net parameters.

- Moreover, even if there will be almost no new ideas in AI, the growth of hardware will eventually result in the creation of a working model of the human brain via direct scanning and modeling of human brains. For example, in 2021, the exascale Aurora 21 computer will be used to create a map of the human connectome (Bouzd, 2018).

We should note that hardware predictions are, in fact, conservative, as any algorithmic improvements will only shorten the expected timeline.

It is also known that humans tend to overestimate their personal uniqueness and achievements in different domains (Taylor & Brown, 1988). This bias may affect us globally, as we could overestimate the complexity of the individual human mind. The actual metrics of the human mind’s performance in most domains are rather bleak: the total size of human memory accessible to the conscious mind is estimated to be only around 2.5 gigabytes (Crevier, 1993), and computational ability is very low.

## 5.3 “Dangerous” level of AI hardware

According to our criteria for dangerous AI, the hardware should reach a level close to the lower estimates of human brain performance, thus enabling an accelerated rate of progress, or at least the capability to create full world models and understand natural language, which is close to autonomous 5 driving capability.

Grace wrote that brain power estimations by different scientists is between 3 x 1013 to 1025 flops (Grace, 2015), and median is “roughly 1-30x1016 flops, with high uncertainty”. But as was said above, if we accept median estimation as our planning point, we are 50 per cent dead, and it is safer to take first 10 per cent of the distribution. Thus, we could take 1x1016 flops (10 petaflops) as lowest safety margin, behind which human level performance will be possible. Having human level performance is not equal to having a human mind, as even if we have correct model human mind, it will require extensive education. But NVIDIA is already implementing training its autonomous car brains in the virtual environment, for which it designed specialized generative network exactly because it wants accelerate the learning process.

However, we are already above this level, as supercomputer Summit has 200 petaflops, and 3 exaflops in deep learning operations (Feldman, 2018a). Even standard Google’s TPU pod has 11 petaflops. This is roughly equal to five NVIDIA’s DGX-2 systems, which would cost 2 million USD at 2018 prices (Solca, 2018).

In other words, many startups now could afford to buy AI hardware with performance roughly equal to the lower threshold of the human brain. Lowering of the price of computation and growth in funding means that an exponentially growing number of agents are gaining access to potentially dangerous level of computation for their experiments.

## 5.4 Future of Moore’s law and new ways to overcome its limitations via multi-chip processors and lower per-computation price

Despite computer lithography is close to its limits, the progress in it is expected at least for next 5 years, to early 2020s (Waldrop, 2016), so the slowdown will not affect AI’s perspectives in the near-term mode. Also, even if Moore’s law completely stops, the growth of total available computer power will continue because of continued manufacturing of the components, and their price will continue to drop as less money will go on new chips research. Moreover, many new architectural solutions, like turn to graphic cards and FPGA, could provide additional acceleration on the chips with the same transistor count. Thus, even if Moore’s law stops, computer power (available for AI researchers) could grow several orders of magnitude.

Also, Moore’s law’s end can be interpreted as the end of the shrinking of transistor sizes, but should not be counted as the end of the progress in computing and of the ability to deliver bigger and cheaper systems for AI research. Moore’s law originally predicted doubling of the number of transistors on the largest chips every one year (1965), then every two years (from 1975), and this still holds: the largest chip in 2016 had 5.7 billion transistors; in 2017, NVIDIA unveiled its Volta V100 chip, with 21.1 billion transistors on a 815 mm² die (Walton, 2017). IBM has announced plans to start manufacturing chips with 30 billion transistor in 2020 (Nield, 2017).

An increase in the number of transistors on a chip could be reached in two ways: either by increasing of the size of the chip, or by using smaller elements. Both of these parameters are close to physical limits. The price of large chips is growing very quickly as error rates increase and chips reach the limitations of current fabrication techniques.

However, the main manufacturers of specialized processors used in AI started to use several chips for one “processing unit”, which opens up a new way to overcome the limitations of a single chip. For example, Google’s TPUs are arranged in modules of 4 chips (Sato, 2017). In other words, parallelization of computing helps users to overcome the technological limits of lithography, and most current AI-related computing could be effectively parallelized.

Implementation of the extreme ultraviolet lithography opens the way for Moore’s law to continue up to the early 2020s (DeYoung, 2017), which could provide at least one order of magnitude increase of performance for a single chip, but the greatest increase will come from parallelization of many chips and architectural improvements.

Moore’s law will hit the physical limits of lithography technology in the sense of feature size somewhere in the 2020s, but that does not imply a limit to chips’ performance per dollar. The size of chips and number of chips in a processing unit could still grow, and manufacturing costs and energy consumption diminish.

While growth of the TOP-500 supercomputers seems to level off in recent years, it is not relevant for AI research which until recently was done on smaller computers optimized for deep learning and which performance is growing very quickly. NVIDIA released computer DGX-2 in March 2018 for $400K (Solca, 2018) with 2 Petaflop in deep learning performance which is said to be 10 times faster in neural nets training than the system DGX-1 from 2017 which cost $149K (this means 4 times increase of cost effectiveness in 1 year). While “deep learning performance” is not the same as typical performance because specialized accelerators are counted, this type of calculation is exactly what is needed for the current progress in AI.

There is also a trend to have more instructions per processor core, from a few in CPUs to tens in GPUs, hundreds in TPUs, and even more on projected chips; e.g. NVIDIA’s Tensor core runs 64 operations per clock (Armasu, 2017b).

Moreover, the latest supercomputer, Summit, is equipped with deep learning capabilities up to 3 exaflops (Feldman, 2018a) and is intended to be used in the AI field.

There are several promising ideas that may increase performance of computers; some of them will probably be practical and may be implemented in 2020s:

* *TPUs and other types of specialized hardware*. Intel has promised to increase neural net performance 100× by 2020 (from 2017) by use of specialized chip-accelerators they called Nervana (Mannes, 2017).
* *3D chips.* A 3D System combining memory and computing cores on a chip may increase energy efficiency 1000 times and computational speed more than 50 times by 2021 by eliminating memory bottlenecks, according to DARPA (DARPA, 2017).
* *Quantum computers* are expected to perform above classical computers in the 2020s in some tasks, but the most interesting would be if they can be used to accelerate training of the neural nets, which would help AI applications. Quantum neural nets are now researched (da Silva, Ludermir, & de Oliveira, 2016).
* *FPGA*. These programmable chips could combine the efficiency of TPUs with the speed of ordinary computers. Fujitsu claims to have optimized FPGA architecture to be 10,000 times faster (Fujitsu, 2016).
* *Memristors*. Memristors seem to enable more efficient neural networks (Du et al., 2017; Kaplan, Yavits, & Ginosar, 2018). They could be the basis for physical neural nets, which could be especially effective in inference, as each memristor will replace one synapse.
* *Spiking neural nets.* The *TrueNorth* chip from IBM provides 10 000× the energy economy of conventional chips and could solve the same tasks as ordinary neural nets after compilation (Hsu, 2014). IBM also invented in 2018 a system of analogues synapses, which provides 100 times the power economy, and also impose less load on the information transfer bus, as the synapses are trained “locally”, as in the human brain (Ambrogio et al., 2018).
* *Non-von-Neumann architectures*. DARPA is exploring a new type of computing which could offer a 1000× boost in computational power called HIVE which will “its ability to simultaneously perform different processes on different areas of memory simultaneously” and work with data graphs (R. Johnson, 2017).
* *Superconducting transistors* could reach 200 GHz, which could enable computers capable of 1024 FLOPS in 2030, according to Dr. T. Sterling (Feldman, 2018b).
* *In-memory-calculations* promise to provide 100 timesimprovement (Hadhazy, 2018).

Other approaches are graphene based chips, approximate computing, photonics. Probably after 2020s photonics, gallium nitride chips (Nowakowski, 2017), complex multilevel 3D structures, and powerful quantum computers will appear and become part of mass market.

## 5.5 AI-related “Moore’s law”: growing budget for AI research helps outperform Moore’s law

The price of computation is important but it is not the only important factor; the total AI research budget of organizations must also be considered. Budgets were small during the “AI winter” of 1990s–2000s, but have since grown many hundred-fold, partially due to a new AI arms race. If China spends tens of billions of dollars on AI superiority, it can spend a large fraction of that budget on development of AI-related hardware.

AI market growth is projected to be 57 percent a year until 2025 (Grand View Researh, 2017). To support such rate, largest competitors need to invest in R&D even more. Spending on AI grew at 54 percent a year (IDC, 2018), of which hardware is part. Given that total world market of hardware is around 1 trillion USD, according to OpenAI (OpenAI, 2018), there is room for growth of 2–3 orders of magnitude.

Governments are not the only players in the field; tech giants like Google and IBM could order specialized computer chips (like TPU) for their software with turnaround times from one to several months. While owning a supercomputer is expensive, renting cloud computing time can be more cost-efficient, as one pays only for time used and there is no downtime. There are several AI-related clouds, including Google *AutoML* *Cloud*. People also now could rent out their GPUs via services like [vectordash.com](http://vectordash.com/) or cover the costs of downtime via cryptocurrency mining.

There are other issues with predictions based on hardware costs and budgets. Falling hardware prices are probably not adjusted for inflation, and if adjusted, will give steeper curves. As the global economy grows, a larger share could be spent on building computational power. Energy consumption is growing as a share of the total price of calculations. Increased availability of cheaper renewable energy and production of more energy-efficient chips could contribute to lower prices for AI development.

AI’s “Super Moore’s law” is fueled by internal demand by tech giants: “Google's scientists are saying "please give us a tera-weight machine," computers capable of computing a trillion weights. That's because "each time you double the size of the [neural] network, we get an improvement in accuracy” (Ray, 2018).

## 5.6 Conclusions about hardware

Computer power available for AI research is doubling much quicker than according Moore’s law because of combined effects of lowering computational prices and growing budgets.

This trend could continue for several years even after Moore’s law in transistor miniaturization will be dead, and there are several ways how this trend could provide cheaper computational cost, including architectural improvement, economy of scale and more chips in a processor unit as well as more threads per core, specialized ASICs, and solving memory bottlenecks to train larger nets.

Thus, there is no hardware constraints to build human level AI now for large companies, like Google, and computer power, similar to the lowest estimation of human brain performance, is currently available even for the individual researchers via clouds.

From this follows that, as we have enough computation power, the main measure of time before AI becomes algorithm performance.

# 6 Progress in neural net performance

We use here progress in current neural nets as a proxy for AI performance, but the real jump to universal AI will probably built upon the existing achievements of neural nets. They will be just building blocks of more complex systems, or perhaps a completely new principle of ML will arise, the implementation of which will build upon experience gained during neural net stage. Storrs Hall thinks that the turn in AI research in last 10 years from academia to industry (which is more interested in combining existing ideas and is capable of employing brute force) is fueling current progress in AI research (Storrs Hall, 2016). Industry has created a positive feedback loop consisting of money, talent, results and expectations, which fuels accelerated progress in the field.

## 6.1 Neural nets revolution dramatically increases the speed of AI progress

Neural nets were known for decades, but in 2012 they started to dramatically outperform other AI methods because of the implementation of two important ideas: training neural nets on very large datasets, like *Imagenet*, and the use of large and many layers neural nets, called deep learning, which become possible because of increased availability of cheap computing power with GPUs.

The implementation of these ideas became possible due to the growth in graphic processor units, which provided a lot of cheap hardware for experimentation.

Transhumanists and futurists started to react to the neural net revolution only in 2014–2015. MIRI added machine learning agenda in its research priorities in 2016 (MIRI, 2016).

We have probably already solved the most computationally difficult part of AI: "Complex strategy and tactics require only a few Neurons — The LSTM to driver OpenAI Five consisted of only 4,000 LSTM nodes.... Here is however the real problem why predicting AGI with 5 to 10 years in within the realm of possibility, this is known as Moravec’s paradox. [Moravec’s paradox](https://en.wikipedia.org/wiki/Moravec%27s_paradox) is the observation made by many AI researchers that high level reasoning requires less computation than low level unconscious cognition" (Perez, 2018). This could mean that the most complex work for AI – image recognition and movement control – is almost solved via progress in deep learning, and we need a computationally simpler reasoning engine, for which we have already enough computing resources and may just need to develop a few new ideas to implement it.

## 6.2 Neural nets’ performance metrics have doubled every year since 2012

After the 2012, the performance of many metrics of neural networks, first of all, progress in low-level human-specific tasks like image and voice recognition started to double every year (“AI Progress Measurement,” 2017). Moreover, neural networks reached superhuman performance in some important areas in 2016–2017 according to these measurements.

There are 5–10 doublings before human levels in some other important areas, which means that near-human performance could be reached in the 2020s, and potentially as early as 2023. *Imagenet’s* recognition performance increased from 27 percent error in 2011 to human-level performance in this task of 5 percent error in 2015, reaching a 1.5 percent error rate by the end of 2017. A twofold reduction in error has occurred every 1.3 years (“AI Progress Measurement,” 2017). Similar progress has been made in street number recognition, handwriting recognition, speech recognition, and in computer and board games according to (“AI Progress Measurement,” 2017). Significant progress has also been made in natural language processing, but it still below human performance. Machine translation increased in BLEU score from 37 (2014) to 41 (2017), while the professional human level is 50. At this rate of progress, it will require six years (2023) to reach a point equal to human-level performance. Performance on the *Stanford Question Answering Dataset 1.1* has increased from 71 (2016) to 81 (2017) and to 95 (2018), while the average human level is 91 (SQuAD, 2018).

## 6.3. Neural nets’ size grew 100 times in last five years

Let us look at recent changes in the size and effectiveness of neural nets. The size (number of parameters, or connections, roughly equal to the number of synapses) of Google’s cat recognizer in 2012 was one billion.

Later most of private research was done on graphics cards and the size of parameters was limited by the size of the memory of graphic cards, which recently reached up to 12 GB. For example, Karpathy's famous recurrent neural nets (RNN) had only 3 million parameters, but were able to generate grammatically correct text (Karpathy, 2015).

However, Google created a neural net in 2016 with 130 billion parameters, which they now use for Google Translate. They have shown that translation quality has grown with the size of the net, though some diminishing returns were observed (Shazeer et al., 2017). Thus, the number of parameters in the best neural nets by Google grew 100 times in five years, and they are planning a trillion-parameter net.

The human brain has around 100–500 trillion synapses (Drachman, 2005). If the speed of growth of the size of the best neural nets continues, a net with 100 trillion synapses can be expected around 2024. By saying "best nets" we exclude some very large simulations which have been done with less impressive results.

However, there is the problem of training such big nets. However, *OpenAI* found a solution which is easily scalable by changing the way the net is educated. It does not use backpropagation, but evolutionary strategies (Salimans, Ho, Chen, Sidor, & Sutskever, 2017).

If we look at the number of “neurons” we will see the same trend observed for parameters. IBM’s Artificial Brain has grown from 256 neurons to 64 million neurons in 2017 in just six years, with ten billion Projected by 2020 (Neal, 2017). It will also only require 20 W of power, as it will use spiking neurons in *TrueNorth* chips. This implies an improvement of one order of magnitude every year, reaching the human level of 100 billion neurons as early as 2021.

## 6.4 As training data sets grow, neural net performance increases logarithmically, and a human-size dataset may be able to provide human-level performance

One of first successes in deep learning came when the size of training data sets had grown dramatically, to around a million objects in the case of *Imagenet*. In the article “Revisiting Unreasonable Effectiveness of Data in Deep Learning Era”, Sun et al. showed that increasing training data set size 10 and 100 times, up to 375 million objects, produced a logarithmic increase in performance which reached the state-of-the-art level even with rather simple neural net architectures (Sun, Shrivastava, Singh, & Gupta, 2017). Extrapolation of the metric used in the article gives near-perfect performance at the level of 100 billion images, equal to one million hours of video. This corresponds to growth of dataset by a factor of 300 in five years, or a doubling time of less than one year (around eight months).

To compare it with human performance, we introduce the notion of a “human dataset”, which is equal to all child lifetime experiences, estimated as ~100 000 hours (12 years) of video—surprisingly similar to the estimation of the human level dataset above. This dataset at 30 frames per second would be ~10 billion images, but there is great redundancy in these images. At the current rate of doubling of the size of the image database (8 months), the human dataset size will be reached in 5 doublings, 3–4 years, or 2021–2022.

The use of such large datasets is technically possible, as more than 100 000 years of YouTube videos are available (Fortunelords, 2016). Google announced a video dataset in 2017 consisting of 57 000 labeled segments and equal to around 40 hours of video (Gu & Ross, 2017).

## 6.5 Signs of self-improving and data transfer in ML

In 2017 several important things happened which apply to the generalization of thinking algorithms. Google invented *AutoML* (Google, 2017), a machine learning system which predicts the best ML configuration for a given task. Deepmind unveiled *AlphaZero* (Silver et al., 2017), which was able to achieve state-of-the-art performance in several board games. Pretrained neural nets (transfer learning) have demonstrated better ability to learn (Shi, Siva, & Xiang, 2017).

In 2018 Kurzweil and his team in Google (Cer et al., 2018) demonstrated transfer learning via whole sentences encoding into pretrained neural net, which is significant step after nets pretrained on the word level.

## 6.6 There are other fields in AI research that could get a boost in the future

Most current success in AI research comes from neural nets, but there are many other neglected fields of AI-related science that could help to boost AI research. These include *OpenCog*, top-down symbolic approaches, genetic algorithms, brain modeling, hundreds of different already existing cognitive architectures, expert systems, analogues of *Eurisko* rule-modifying systems (Lenat & Brown, 1984), and Bayesian nets (Kendall, 2017).

Many of these approaches are not feasible without access to large amounts of computer power. All of them could use modules created by neural net systems for some low-level processing and thus get a boost.

## 6.7 Evolution of the “intellect stack”

Evolution of AI becomes possible because of evolution of the “intellect stack” promoted by NVIDIA, that is, a vertically interconnected system of chips, low-level computational language (CUDA), higher-level language for implementing neural nets, open libraries and open courses in Python and ML to attract more programmers to the field, as well as financial incentives. Any improvement in connection between levels produces large gains in total performance in the field. The increasing quality of the intellect stack means that its different levels may be replaced in a manner invisible to the end user; for example, IBM’s spiking networks could seamlessly perform algorithms prepared for ordinary networks (Levenchuk, 2017; Price, 2017).

## 6.8 Jumps in algorithmic efficiency in neural networks

Even with all improvements in the hardware, neural networks are computationally intensive, which limits the size of the networks that could be trained, even on supercomputers. The main limits are the need to quickly accesses large DRAM, massive energy consumption, difficulties in parallelization, and long training times. However, there are several approaches which promise jumps in performance. We will list just a few recent ideas:

* Binary neural nets. It turns out that lowering the precision of calculations does not worsen neural net performance. Simpler calculations, the extreme case of which is binary networks which use simple logical operations, eventually results in a loss of accuracy, but this loss can be compensated for by other means (Prabhu, Batchu, Gajawada, Munagala, & Namboodiri, 2018).
* Sparse Evolutionary Training, where not all neurons are connected, is a promising approach to quadruple the number of artificial neuron models on a computer. The authors state that current supercomputers could train nets with only 16 million artificial neurons (not parameters), but after implementation of the new training method, the number could reach 80 billion (Mocanu et al., 2018).
* Graph networks suggested by DeepMind could process graphs and be used for causation conclusions (Battaglia et al., 2018).
* Google Brain has reached a major learning time reduction with networks using attention (Vaswani et al., 2017).
* A “generative cortical network” could reduce the need for training data to just a few hundred examples (George et al., 2017).
* Pretreating of neural nets for quicker learning and knowledge transfer, which allows quick learning, like in OpenAI’s Reptile (Nichol, Achiam, & Schulman, 2018).
* The *Resnet* competition resulted in 150x progress in computation time in one year, from October 2017 to October 2018 (Stanford DAWN, 2018); examples like this tell us that progress in algorithm development may be more important than progress in hardware. Such happening right now with surprising speed, and there is probably plenty of room for such progress still ahead.

## 6.9 Different attempts to predict human-level AI timing in neural nets converge around the same date

Using different approximations of neural net performance metrics, we got above in the section 6 the following predictions for the timing of near human AI performance: converge to a median in the coming decade.

This is in line with the prediction of Shakirov (2016), who concluded that human-level AI will be achieved between 2021 and 2026 using similar but different sets of performance metrics which were available at the time he wrote his article.

This prediction includes expectation of the continued hardware growths which will enable supporting of very large neural nets and assumed that no unpredictable difficulties will arise.

# 7 Surveys of experts

There have been several surveys of experts about AI timing.

In Klein’s poll in 2007 median was between 2030 and 2050 and first 7 per cent was before 2020 (Klein, 2007).

In Baum and Goertzel poll in 2009 poll Nobel science level of AI is expected with 10 percent at 2020 and 50 percent at 2045 (Baum, Goertzel, & Goertzel, 2011).

Bostrom’s survey in 2012-2013, that is, before the current boom, gave median timing of high-level AI as of 2050 (Müller & Bostrom, 2016).

Grace poll gave median time of AGI creation is 45 years after the poll which was conducted in 2016, so it is around 2061 (Grace, Salvatier, Dafoe, Zhang, & Evans, 2017) with 6.25 per cent before 2022 and around 16 per cent to 2030.

A survey at the HLAI-2018 conference in Prague showed that 37% of respondents expect “human-like artificial intelligence will be achieved within five to 10 years" (Johnson, 2018).

Different surveys at different times produce consistent results with median AI timing at 2050 (between 2045 and 2061) and several percent of probability around 2020. There is also small upward trend in estimation of median AI timing, probably, as experts update on the information that AI was not created in the last 10 years in 2010s.

# 8 All hyperbolic future growth projections converge in the next two decades

There are several independent projections of the future of humanity based on the idea of the hyperbolic acceleration calculated based on different historical trends (Table 1).

Table 1: Different hyperbolic predictions for the emergence of AI.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Prediction by | Initial prediction date of infinity | Updated date based on recent data points | Human-level AI and the start of pre-singularity turbulence |  |
| Foerster, 1960 | 2026 |  |  |  |
| Vinge, 1993 | 2005-2030 | 2030 |  |  |
| Panov- Diakonov, 1994 | 2004 | 2024 | 2021 |  |
| Schmidhuber, 2006 | 2040 | 2036 | 2030 |  |
| **Mean** | 2019-2025 | **2030** | 2025 |  |

In 1993, Vernor Vinge famously predicted the technological singularity (AGI) and said that he would be surprised if it were to happen before 2005 or after 2030 (Vinge, 1993). In 1960, Foerster created a model of human population growth called: “Doomsday: Friday, 13 November, A.D. 2026” (Von Foerster, Mora, & Amiot, 1960). According to him, at that date, human population will approach infinity if it grows as it has grown in the last two millennia. Korotayev (Markov & Korotayev, 2009) created an explanation of this law as a solution to a differential equation, where the speed of innovations is proportional to the square of population, and population is proportional to innovation’s ability to support a larger population. However, Korotayev’s law stopped working as early as the 1960s because humans cannot replicate that fast. If we could count population as a total of humans and computers, we could see that this total population is still growing very quickly, as most people in developed countries own multiple computers in the form of phones, TVs, game consoles, cars, and PCs.

In his recent article, Korotayev made a revision of the different predictions about singularity, which all converge around 2027-2029, however, he concluded that this means only the end the past trends and the beginning of global slowdown (Korotayev, 2018).

Panov (Panov, 2005) (based on works of Diakonov and Snooks) charted scientific revolutions and created a law that predicts the timing of each revolution. His hyperbolic law is presented as dates of technological revolutions, where each of them occurs in a period 2.67 times shorter than the previous. His last reference points are 1830 and 1950 (marking the first and second industrial revolutions), and his singularity point is at 2004 ± 15 years. Each data point is the end of a large geopolitical epoch and the beginning of the new technological epoch. However, if we extrapolate his law based on his last data points, we obtain the following dates: 1994, 2011, 2017, 2020, 2021, with infinity reached in 2021.

Interestingly, 1994 and 2011 each nearly coincide with a technological and a geopolitical revolution. The former year, 1994, is close to the end of Cold War and to the beginning of the Internet revolution; the latter, 2011, is near the beginning of the neural net revolution and what is called the 4th technological revolution (Schwab, 2015). There were also geopolitical changes including growth of China as a global player and the Arab Spring. If we correct the last points based on actual data, that is, if we use 1991 and 2012, the next revolutions should be expected in 2020, then 2023, with the singularity at 2024.6

Schmidhuber estimated in 2006 by extrapolating from history, that the “Omega” point (another term for technological singularity) will be reached in 2040 (Schmidhuber, 2006). He used similar but different data points to Diakonov-Panov.

Predicted mean timings based on historical data are presented in the bottom row of Table 1. While the uncertainty of such predictions is large, the 2020s may be the start of a very dangerous pre-singularity period in which changes could happen very quickly and include war and other catastrophes. Bostrom called it turbulence in society caused by the growth of AI (Bostrom, Dafoe, & Flynn, 2016).

Acceleration increased from Moore’s law to neural nets’ law: the doubling time declined from two years in chip size to one year in neural net performance. Hanson has predicted that after human-level AI is created, doubling time will be around one month (Hanson, 2016). On larger timescales, Moore’s law looks not exponential, but like part of a hyper-exponential curve, as in the beginning of the twentieth century its doubling time was four years (Kurzweil, 2006), in the second half it was around two years, and in the age of neural nets it is one year in AI related hardware as we discussed above in section 4.1.

Hyperbolic future predictions do not take into account the possibility of existential catastrophes which could prevent reaching the singularity point. This could be analogous to the way in which an object falling into a black hole will never reach its singularity but will be destroyed near its Schwarzschild radius. We call such a growing probability of catastrophes “oscillations before the singularity”. Thus, even as most hyperbolic predictions put the singularity at 2030 ± 5 years, such an existential catastrophe could happen several years before.

# 9 AI timing predicted based on randomness of the moment of AI creation

## 9.1 The growth of the probability to find ground breaking idea

Yudkowsky (Yudkowsky & Hanson, 2008) suggested that the creation of AI depends not on the availability of hardware, but on appearing of one crucial idea, and that this makes AI unpredictable. However, the probability of such an idea appearing depends on the number of researchers.

The number of students enrolled in major universities to study machine learning grew around tenfold from 2007 to 2017, and the number of AI-related papers, which is the lagging indicator, grew nine times from 1996 to 2017 (Grey, 2017). Future investment in AI is expected to grow, with the AI market projected to grow 57-fold between 2016 and 2025, implying an order of magnitude increase in the number researchers over the same period (Feldman, 2017).

The creator of *Keras* estimated that the number of neural net researchers has grown from 10,000 to one million from 2015 (Rosebrock, 2018), implying an order of magnitude growth in a year, but the total number of programmers in the world is around 25 million, which puts an upper limit on such growth. Other metrics of AI research are also growing, maybe not so spectacularly, as demonstrated in the “AI Index report 2017” (Shoham, Perrault, Brynjolfsson, & Clark, 2017), with doubling periods of a few years for different metrics. However, this mostly covers the period before 2016, after which it is possible that even stronger growth has flourished.

If ideas come to researchers randomly, then the probability of Yudkowsky’s “crucial idea” is also increasing exponentially with the number of researchers. This means that 100 years of “linear” AI research without such growth will be condensed into just a few years.

However, the creation of new ideas has diminishing returns, as it becomes impossible to read all new papers in the field and test all ideas. In addition, most ideas currently tested in the deep learning field have been well-known since at least the 1990s, but at that time the hardware capacity to test them did not exist. Ideas alone, without ability to quickly test them, are rather useless, but the growth in the number and the power of AI-specialized computers will lower the costs and time for testing even “crazy” ideas, and thus more ideas will be tested.

## 9.2 AI timing prediction based on my random location between beginning of the research and its end

There is an even more speculative way to predict the future based on logic similar to the Doomsday argument, which the author has previously discussed in (Turchin, 2015). We could use the mediocrity principle, that is, we, as observers of AI development, are randomly chosen somewhere between the beginning of AI research (in 1951) and the future moment of AI creation. According to Gott’s formula (Gott III, 1993) for predicting the future duration of a process based on the random moment of its observation, in that case, AI will be created with 50 percent probability in the next 67 years (2018 - 1951 = 67), or in 2085 (which is surprisingly close to the median estimate of 2062 for the timing of AI’s creation by Grace’s poll (Katja Grace et al., 2017)). However, this prediction is based on the assumption of a linear distribution of the probability of AI creation within the entire period. If we consider that the number AI researchers has grown 10 times recently, and assume that this corresponds to 10 times increase in the probability density of finding correct idea about AI, the period before AI could be 10 times less, that is ~7 years from now, or 2025.

This is just speculation, of course, but it demonstrates that replacing a hardware -based prediction with prediction based in the random generation of some crucial AI-related idea does not change the estimated time of AI creation.

# 10 Combining all projections

In the table 2 all predictions about powerful AI timing are combined, and they are in agreement about earliest time of dangerous AI creation in 2020s.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method of prediction** | **Earliest expected**  **powerful AI creation (first 5 percent)** | **Median AI creation timing** | **Date of a AI-related catastrophe** |  |
| Poll of experts by Grace | 2021 | 2062? | +8 years? |  |
| Hardware extrapolation based on Moore’s law and brain power | Now, as of 2017 |  |  |  |
| Aggregated  extrapolation of different neural nets performance metrics | 2022 |  |  |  |
| Number of “neurons” | 2021 (IBM) | 2025 (Google) |  |  |
| Size of training dataset |  |  |  |  |
| Future funding compared with previous funding |  |  |  |  |
| Historical analogies | 2022 (2017 is new 1939 when major players start arms race to nuclear weapons) |  |  |  |
| Hyperbolic projections (more in table 1) | 2025 | 2030 median time of singularity |  |  |

*Table 2. Summary of methods for predicting the timing of AI emergence.*

# 11 Uncertainty in AI predictions

Above we show that different methods of predicting the timing of the emergence of AI give an earliest date of creation for potentially dangerous AI coming decade, but it is known that predicting AI is very difficult (Armstrong & Sotala, 2015). In our case, however, this is not a simple prediction, and not even directly about AI. It is not a prediction about the timing of the first AI, but of the beginning of a period of significant existential risk.

Another factor that makes our predictions uncertain is the possibility of future “black swan” events, which could be geopolitical, like war, or technological, like a sudden insurmountable technological obstacle or change in technological evolution in an unprecedented direction. Some black swan events could accelerate AI progress, like changes in the leading type of AI from neural nets to something else. Others could decelerate AI, like a severe AI winter connected with hype depletion, or obstacles connected to semiconductor technology. Technological black swans occur with a frequency of about once around a decade (Taleb didn’t exactly state it but his examples, like appearance of internet and 2001 terroristic attack 10 years distance between them (Taleb, 2007); also there is a point of views that in more complex systems black swan events are increasing in frequency (Ormerod, 2007)) thus, we could expect that all trends will continue for next few years, but not next 10–20 years.

It all means that we have some predictions about AI which are enough for us to worry about AI catastrophic risks, but not enough to be sure that AI will be created. As it is not easy to quantify the level of the probability above which we should worry, or the actual probability distribution of AI creation, or its uncertainty, we could use a non-digital presentation of the risk levels, as discussed in (Turchin & Denkenberger, 2018b), and we could state that there are risks significant enough to pay attention to of dangerous AI creation in the next decade, which corresponds to the “yellow level” in the color-coded scale of risks (Turchin & Denkenberger, 2018b).

# 12 What could be done to ameliorate risks of early AI?

It looks like “plan A” for AI safety, based on the creation and implementation of robust AI safety using the first AGI can’t be done in time, as this process would require a decade of mathematical research according to MIRI estimates, while dangerous AI may be created sooner. Thus, we need something like a “plan B”, which is much less likely to work, but is better than nothing. Such a plan may include:

* An active AI boxing and control system, including limiting AI capabilities via “artificial stupidity” (Trazzi & Yampolskiy, 2018).
* International agreements to control AI (for more, see Turchin & Denkenberger, 2017a)
* Use of Narrow AI to create global ubiquitous surveillances system (Turchin, 2018).
* Creation of the simple version of AI safety theory, where all known up-to-date ideas are presented in a form which can be used by AI developers.
* Balance of capabilities between a few, probably nationwide AI; this may be in the form of a US–China AI cold war (Turchin & Denkenberger, 2018c).
* Use of more mundane AI safety measures, like distillation and amplification (Christiano, 2016), may be starting from self-driving car safety or ethical training datasets.
* Concentration of AI research in a few trusted centers with high internal control.
* Try to combine AI safety excellence with outperforming everybody else in AGI research by concentration of the best minds in one breakthrough project.

Another interesting point is that earlier arrival of AI may be beneficial, as only an advance global control systems seems to be able to prevent other global risks, especially risks of democratization of synthetic biology which could also happen near-term (Torres, 2018; Turchin, Green, & Denkenberger, 2017).

# Conclusion

Our analysis of current computer hardware and neural net development implies that potentially dangerous AI could be created in the coming decade. This probability is not certain but it is enough to start paying attention to the risks posed by AI and to prepare some adequate safety measures for local and global control of potentially dangerous AIs.

**Acknowledgments:** Special thanks for David Denkenberger who did a lot of work to improve this article. We would like also to thank D. Kornienko, V. Shakirov, S. Markov, R. Gruetzemacher and Jonathan Yan for thoughtful discussion of the topic. No current external funding sources for this study.

Literature:

AI Progress Measurement. (2017). Retrieved from https://www.eff.org/ai/metrics

Ambrogio, S., Narayanan, P., Tsai, H., Shelby, R. M., Boybat, I., Nolfo, C., … Farinha, N. C. (2018). Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature*, *558*(7708), 60.

Arbital. (2016). Known-algorithm non-self-improving agent. Retrieved January 22, 2018, from https://arbital.com/p/KANSI/

Armasu, L. (2017a, April 5). Google’s Machine Learning Chip Is Up To 30x Faster, 80x More Efficient Than CPUs And GPUs. *Tom’s Hardware*. Retrieved from https://www.tomshardware.com/news/google-tpu-comparison-haswell-k80,34069.html

Armasu, L. (2017b, May 11). On Tensors, Tensorflow, And Nvidia’s Latest “Tensor Cores.” Retrieved November 10, 2018, from https://www.tomshardware.com/news/nvidia-tensor-core-tesla-v100,34384.html

Armstrong, S., & Sotala, K. (2015). How we’re predicting AI–or failing to. In *Beyond Artificial Intelligence* (pp. 11–29). Springer.

Barrett, A. M., Baum, S. D., & Hostetler, K. (2013). Analyzing and reducing the risks of inadvertent nuclear war between the United States and Russia. *Science & Global Security*, *21*, 106–133.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., … Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. *ArXiv:1806.01261 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1806.01261

Baum, S. D., Goertzel, B., & Goertzel, T. G. (2011). How long until human-level AI? Results from an expert assessment. *Technological Forecasting and Social Change*, *78*(1), 185–195.

Baumann, T. (2018, October 23). Thoughts on short timelines. Retrieved November 9, 2018, from http://s-risks.org/thoughts-on-short-timelines/

Bayesian Investor Blog. (2017). Moore’s Law and AGI Timelines. Retrieved January 22, 2018, from http://www.bayesianinvestor.com/blog/index.php/2017/12/21/moores-law-and-agi-timelines/

Bostrom, N., Dafoe, A., & Flynn, C. (2016). *Policy Desiderata in the Development of Machine Superintelligence*. working Paper, Oxford University.

Bouzd, S. (2018, June 6). The supercomputer that could map the human brain. *NOVA Next*. Retrieved from http://www.pbs.org/wgbh/nova/next/tech/brain-mapping-supercomputer/

Cannell, J. (2015). The Brain as a Universal Learning Machine. Retrieved June 17, 2018, from https://www.lesswrong.com/posts/9Yc7Pp7szcjPgPsjf/the-brain-as-a-universal-learning-machine

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., … Kurzweil, R. (2018). Universal Sentence Encoder. *ArXiv:1803.11175 [Cs]*. Retrieved from http://arxiv.org/abs/1803.11175

Chklovski, T. (2017, May 22). Artificial Intelligence - what should we tell our children. Retrieved July 1, 2018, from https://www.huffingtonpost.com/entry/artificial-intelligence-what-should-we-tell-our-children\_us\_59235f2de4b0b28a33f62ecc

Christiano, P. (2016). Prosaic AI alignment. Retrieved from https://ai-alignment.com/prosaic-ai-control-b959644d79c2

Constantin, S. (2017, January 28). Performance Trends in AI. Retrieved April 2, 2018, from https://srconstantin.wordpress.com/2017/01/28/performance-trends-in-ai/

Crevier, D. (1993). *AI: the tumultuous history of the search for artificial intelligence*. Basic Books.

da Silva, A. J., Ludermir, T. B., & de Oliveira, W. R. (2016). Quantum perceptron over a field and neural network architecture selection in a quantum computer. *Neural Networks*, *76*, 55–64.

DARPA. (2017). 3 Dimensional Monolithic System on a Chip (3DSoC). Retrieved from https://www.darpa.mil/attachments/3DSoCProposersDay20170915.pdf

DeYoung, C. (2017). Moore’s Law….where next? asml.com. Retrieved from https://staticwww.asml.com/doclib/investor/presentations/2017/asml\_20170322\_2017-03-22\_BAML\_Taiwan.pdf

Drachman, D. A. (2005). Do we have brain to spare? *Neurology*, *64*(12), 2004–2005.

Du, C., Cai, F., Zidan, M. A., Ma, W., Lee, S. H., & Lu, W. D. (2017). Reservoir computing using dynamic memristors for temporal information processing. *Nature Communications*, *8*(1), 2204.

Feldman, M. (2017). Market for Artificial Intelligence Projected to Hit $36 Billion by 2025 | TOP500 Supercomputer Sites. Retrieved from https://www.top500.org/news/market-for-artificial-intelligence-projected-to-hit-36-billion-by-2025/

Feldman, M. (2018a). New GPU-Accelerated Supercomputers Change the Balance of Power on the TOP500 | TOP500 Supercomputer Sites. Retrieved from https://www.top500.org/news/new-gpu-accelerated-supercomputers-change-the-balance-of-power-on-the-top500/

Feldman, M. (2018b). Thomas Sterling Talks Exascale, Chinese HPC, Machine Learning, and Non-von Neumann Architectures | TOP500 Supercomputer Sites. Retrieved June 23, 2018, from https://www.top500.org/news/thomas-sterling-talks-exascale-chinese-hpc-machine-learning-and-non-von-neumann-architectures/

Fortunelords. (2016, March 19). Youtube Statistics - 2018. Retrieved April 7, 2018, from https://fortunelords.com/youtube-statistics/

Fujitsu. (2016). Fujitsu Laboratories Develops New Architecture that Rivals Quantum Computers in Utility - Fujitsu Global. Retrieved June 21, 2018, from http://www.fujitsu.com/global/about/resources/news/press-releases/2016/1020-02.html

George, D., Lehrach, W., Kansky, K., Lázaro-Gredilla, M., Laan, C., Marthi, B., … Wang, H. (2017). A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs. *Science*, *358*(6368), eaag2612.

Google. (2017). AutoML for large scale image classification and object detection. Retrieved April 2, 2018, from https://research.googleblog.com/2017/11/automl-for-large-scale-image.html

Gott III, J. R. (1993). Implications of the Copernican principle for our future prospects. *Nature*, *363*, 315–319.

Grace, K. (2015). Brain performance in FLOPS – AI Impacts. Retrieved January 23, 2018, from https://aiimpacts.org/brain-performance-in-flops/

Grace, K. (2017a). Recent trend in the cost of computing – AI Impacts. Retrieved January 23, 2018, from https://aiimpacts.org/recent-trend-in-the-cost-of-computing/

Grace, K. (2017b). When Will AI Exceed Human Performance? Evidence from AI Experts. Retrieved from https://arxiv.org/pdf/1705.08807.pdf

Grace, Katja, Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2017). When Will AI Exceed Human Performance? Evidence from AI Experts. *ArXiv Preprint ArXiv:1705.08807*.

Grand View Researh. (2017). Artificial Intelligence Market Size Analysis | Industry Growth Report 2025. Retrieved from https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market

Granoff, J. (2016). Donald Trump Is an Existential Threat to America and the World. *Time*. Retrieved from http://time.com/4560742/donald-trump-existential-threat/

Grey, A. (2017). These charts will change how you see the rise of AI. Retrieved January 25, 2018, from https://www.weforum.org/agenda/2017/12/charts-artificial-intelligence-ai-index/

Gu, G., & Ross, D. (2017). Announcing AVA: A Finely Labeled Video Dataset for Human Action Understanding. Retrieved January 24, 2018, from https://research.googleblog.com/2017/10/announcing-ava-finely-labeled-video.html

Gyrphalcon. (2018). Lightspeeur 2803 Neural Accelerator for Edge & Cloud. Retrieved November 12, 2018, from https://www.gyrfalcontech.ai/solutions/2803s/

Hadhazy, A. (2018, September 21). Chip ramps up artificial intelligence systems’ performance. *School of Engineering and Applied Science*. Retrieved from https://engineering.princeton.edu/news/2018/09/21/chip-ramps-artificial-intelligence-systems-performance

Hanson, R. (2016). *The Age of Em: Work, Love, and Life when Robots Rule the Earth*. Oxford University Press.

Hsu, J. (2014). How IBM Got Brainlike Efficiency From the TrueNorth Chip. Retrieved from https://spectrum.ieee.org/computing/hardware/how-ibm-got-brainlike-efficiency-from-the-truenorth-chip

IDC. (2018). Worldwide Spending on Cognitive and Artificial Intelligence Systems Will Grow to $19.1 Billion in 2018, According to New IDC Spending Guide. Retrieved from https://www.idc.com/getdoc.jsp?containerId=prUS43662418

Johnson, R. (2017). DARPA Funds Development of New Type of Processor | EE Times. *EETimes*. Retrieved from https://www.eetimes.com/document.asp?doc\_id=1331871

Johnson, S. (2018). Human-like AI could emerge in 5 to 10 years. *Big Think*. Retrieved from https://bigthink.com/surprising-science/computers-smart-as-humans-5-years

Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., … Borchers, A. (2017). In-datacenter performance analysis of a tensor processing unit. In *Computer Architecture (ISCA), 2017 ACM/IEEE 44th Annual International Symposium on* (pp. 1–12). IEEE.

Kaplan, R., Yavits, L., & Ginosar, R. (2018). PRINS: Processing-in-Storage Acceleration of Machine Learning. *IEEE Transactions on Nanotechnology*.

Karnofsky, H. (2016, May 6). Potential Risks from Advanced Artificial Intelligence: The Philanthropic Opportunity. Retrieved November 9, 2018, from https://www.openphilanthropy.org/blog/potential-risks-advanced-artificial-intelligence-philanthropic-opportunity

Karpathy, A. (2015). The unreasonable effectiveness of recurrent neural networks. Retrieved from http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Kendall, A. (2017, May 23). Deep Learning Is Not Good Enough, We Need Bayesian Deep Learning for Safe AI. Retrieved April 2, 2018, from //alexgkendall.com/computer\_vision/bayesian\_deep\_learning\_for\_safe\_ai/

Kent, A. (2004). A critical look at risk assessments for global catastrophes. *Risk Analysis*, *24*(1), 157–68.

Klein, B. (2007). When will AI Surpass Human-Level Intelligence. AGI-World.

Koomey, J., Berard, S., Sanchez, M., & Wong, H. (2011). Implications of historical trends in the electrical efficiency of computing. *IEEE Annals of the History of Computing*, *33*(3), 46–54.

Korotayev, A. (2018). The 21st Century Singularity and its Big History Implications: A re-analysis. *Journal of Big History*, *2*(3), 73–119.

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

Larabel, M. (2016). OpenGL/OpenCL Performance & Perf-Per-Watt From NVIDIA’s GeForce 9800GTX To GTX 1080 - Phoronix. Retrieved November 12, 2018, from https://www.phoronix.com/scan.php?page=article&item=9800gtx-gtx-1080&num=1

Lenat, D. B., & Brown, J. S. (1984). Why AM and EURISKO appear to work. *Artificial Intelligence*, *23*(3), 269–294.

Levenchuk, A. (2017). How many levels in the intellegence stack? Retrieved July 1, 2018, from https://ailev.livejournal.com/1382464.html

Mannes, J. (2017). Intel to ship new Nervana Neural Network Processor by end of 2017. Retrieved from https://techcrunch.com/2017/10/17/intel-to-ship-new-nervana-neural-network-processor-by-end-of-2017/

Markov, A. V., & Korotaev, A. V. (2009). *Hyperbolic Growth in Nature and Society*. Moscow: Book House “Librikom”[in Russian].

Merritt, R. (2018). AI Hardware Summit Sees Hardware Testing. Retrieved November 10, 2018, from https://www.eetasia.com/news/article/18092005-ai-hardware-summit-sees-hardware-testing

Mikami, H., Suganuma, H., U-chupala, P., Tanaka, Y., & Kageyama, Y. (2018). ImageNet/ResNet-50 Training in 224 Seconds. *ArXiv:1811.05233 [Cs]*. Retrieved from http://arxiv.org/abs/1811.05233

MIRI. (2016, May 4). A new MIRI research program with a machine learning focus. Retrieved April 4, 2018, from https://intelligence.org/2016/05/04/announcing-a-new-research-program/

Mocanu, D. C., Mocanu, E., Stone, P., Nguyen, P. H., Gibescu, M., & Liotta, A. (2018). Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nature Communications*, *9*(1), 2383. https://doi.org/10.1038/s41467-018-04316-3

Morris, K. (2018, September 11). Brainchip Debuts Neuromorphic Chip. Retrieved November 10, 2018, from http://www.eejournal.com/article/brainchip-debuts-neuromorphic-chip/

Mortensen, H. S., Pakkenberg, B., Dam, M., Dietz, R., Sonne, C., Mikkelsen, B., & Eriksen, N. (2014). Quantitative relationships in delphinid neocortex. *Frontiers in Neuroanatomy*, *8*, 132.

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental issues of artificial intelligence* (pp. 553–570). Springer.

Neal. (2017, June 30). IBM’s Artificial Brain Has Grown From 256 Neurons to 64 Million Neurons in 6 Years - 10 Billion Projected by 2020. Retrieved January 24, 2018, from https://www.singularityarchive.com/ibms-artificial-brain-has-grown-from-256-neurons-to-64-million-neurons-in-6-years/

Nichol, A., Achiam, J., & Schulman, J. (2018). On First-Order Meta-Learning Algorithms. *ArXiv:1803.02999 [Cs]*. Retrieved from http://arxiv.org/abs/1803.02999

Nield. (2017). IBM’s New Computer Chips Can Fit 30 Billion Transistors on Your Fingertip.

Nowakowski, T. (2017). Gallium nitride processor—next-generation technology for space exploration. Retrieved from https://phys.org/news/2017-12-gallium-nitride-processornext-generation-technology-space.html

NVIDIA. (2018). NVIDIA TITAN V Specs. Retrieved November 12, 2018, from https://www.techpowerup.com/gpu-specs/titan-v.c3051

OpenAI. (2018, May 16). AI and Compute. Retrieved June 17, 2018, from https://blog.openai.com/ai-and-compute/

Ormerod, P. (2007). Cascades of failure and extinction in dynamically evolving complex systems. In *Noise and Stochastics in Complex Systems and Finance* (Vol. 6601, p. 66010Q). International Society for Optics and Photonics.

Panov, A. D. (2005). Scaling law of the biological evolution and the hypothesis of the self-consistent Galaxy origin of life. *Advances in Space Research*, *36*(2), 220–225.

Patterson, A. (2017). Nvidia CEO Says Moore’s Law Is Dead. *EETimes*. Retrieved from https://www.eetimes.com/document.asp?doc\_id=1331836

Peng, T. (2018, November 13). OpenAI Founder: Short-Term AGI Is a Serious Possibility. Retrieved November 16, 2018, from https://medium.com/syncedreview/openai-founder-short-term-agi-is-a-serious-possibility-368424f7462f?fbclid=IwAR1w0Jlk54BUeQj7crnPrdZ-Nae0C3CfY2O\_YSb5SeIjhEtauUhPIVn1w0A

Perez, C. (2018). Why AGI is Achievable in Five Years – Intuition Machine – Medium. Retrieved November 9, 2018, from https://medium.com/intuitionmachine/near-term-agi-should-be-considered-as-a-possibility-9bcf276f9b16

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

Prabhu, A., Batchu, V., Gajawada, R., Munagala, S. A., & Namboodiri, A. (2018). Hybrid Binary Networks: Optimizing for Accuracy, Efficiency and Memory. *ArXiv:1804.03867 [Cs]*. Retrieved from http://arxiv.org/abs/1804.03867

Price, J. (2017, October 20). You’re not an AI company until you’re a full-stack company. Retrieved July 1, 2018, from https://venturebeat.com/2017/10/20/youre-not-an-ai-company-until-youre-a-full-stack-company/

Ray, T. (2018). Google says “exponential” growth of AI is changing nature of compute. Retrieved November 10, 2018, from https://www.zdnet.com/article/google-says-exponential-growth-of-ai-is-changing-nature-of-compute/

Rosebrock, A. (2018, July 2). An interview with Francois Chollet. Retrieved September 3, 2018, from https://www.pyimagesearch.com/2018/07/02/an-interview-with-francois-chollet/

Salimans, T., Ho, J., Chen, X., Sidor, S., & Sutskever, I. (2017). Evolution Strategies as a Scalable Alternative to Reinforcement Learning. *ArXiv:1703.03864 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1703.03864

Sato, K. (2017). An in-depth look at Google’s first Tensor Processing Unit (TPU) | Google Cloud Big Data and Machine Learning Blog. Retrieved January 24, 2018, from https://cloud.google.com/blog/big-data/2017/05/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

Schmidhuber, J. (2006). HISTORY CONVERGING - AGAIN? Omega Point - Historic Singularity. Retrieved November 2, 2017, from http://people.idsia.ch/~juergen/history.html

Schwab, K. (2015, December 12). The Fourth Industrial Revolution. *Foreign Affairs*. Retrieved from https://www.foreignaffairs.com/articles/2015-12-12/fourth-industrial-revolution

Shakirov, V. (2016). Review of state-of-the-arts in artificial intelligence with application to AI safety problem. *ArXiv Preprint ArXiv:1605.04232*. Retrieved from https://arxiv.org/abs/1605.04232

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *ArXiv Preprint ArXiv:1701.06538*.

Shi, Z., Siva, P., & Xiang, T. (2017). Transfer learning by ranking for weakly supervised object annotation. *ArXiv Preprint ArXiv:1705.00873*.

Shoham, Y., Perrault, R., Brynjolfsson, E., & Clark, J. (2017). AI INDEX 2017. Retrieved from https://aiindex.org

Siluk, S. (2013, May 13). Bitcoin network out-muscles top 500 supercomputers. Retrieved January 23, 2018, from https://www.coindesk.com/bitcoin-network-out-muscles-top-500-supercomputers/

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., … Hassabis, D. (2017). Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *ArXiv:1712.01815 [Cs]*. Retrieved from http://arxiv.org/abs/1712.01815

Smith, R. (2014). NVIDIA Launches Tesla K80, GK210 GPU. Retrieved November 12, 2018, from https://www.anandtech.com/show/8729/nvidia-launches-tesla-k80-gk210-gpu

Smith, R. (2017). The NVIDIA GPU Tech Conference 2017 Keynote Live Blog. Retrieved November 10, 2018, from https://www.anandtech.com/show/11360/the-nvidia-gpu-tech-conference-2017-keynote-live-blog

Solca, B. (2018). Nvidia DGX-2 is the “world\’s largest GPU.” Retrieved April 2, 2018, from https://www.notebookcheck.net/Nvidia-DGX-2-is-the-world-s-largest-GPU.292930.0.html

Sophon. (2018). New edge AI chip Released by Bitmain for Cloud-Edge Integration. Retrieved November 12, 2018, from https://www.sophon.ai/post/36.html

Sotala, K. (2017). How feasible is the rapid development of artificial superintelligence? *Physica Scripta*, *92*(11). https://doi.org/10.1088/1402-4896/aa90e8

Sotala, K. (2018). Disjunctive scenarios of catastrophic AI risk. *Artificial Intelligence Safety And Security, (Roman Yampolskiy, Ed.), CRC Press*. Retrieved from http://kajsotala.fi/assets/2017/11/Disjunctive-scenarios.pdf

SQuAD. (2018). The Stanford Question Answering Dataset. Retrieved November 11, 2018, from https://rajpurkar.github.io/SQuAD-explorer/

Stanford DAWN. (2018). Stanford DAWN Deep Learning Benchmark (DAWNBench) · CIFAR10 Training. Retrieved November 12, 2018, from https://dawn.cs.stanford.edu/benchmark/CIFAR10/train.html

Storrs Hall, J. (2016). Where is my Flying Car?: The coming AI phase change. Retrieved November 11, 2018, from http://wimflyc.blogspot.com/2016/08/the-coming-ai-phase-change.html

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *ArXiv:1707.02968 [Cs]*. Retrieved from http://arxiv.org/abs/1707.02968

Sverdlik, Y. (2016, August 23). What is the Data Center Cost of 1kW of IT Capacity? *Data Center Knowledge*. Retrieved from https://www.datacenterknowledge.com/archives/2016/08/23/what-is-the-data-center-cost-of-1kw-of-it-capacity

Sze, K. (2009, December 21). FASTRA II – the Most Powerful Desktop Supercomputer. Retrieved November 12, 2018, from http://www.itechnews.net/2009/12/20/fastra-ii-the-most-powerful-desktop-supercomputer/

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable* (Vol. 2). Random house.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193.

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

Teich, P. (2018, May 10). Tearing Apart Google’s TPU 3.0 AI Coprocessor. Retrieved November 12, 2018, from https://www.nextplatform.com/2018/05/10/tearing-apart-googles-tpu-3-0-ai-coprocessor/

Torres, P. (2016, November 14). What a Trump Presidency Means for Human Survival: One Expert’s Take. Retrieved November 4, 2017, from https://medium.com/@philosophytorres/what-a-trump-presidency-means-for-human-survival-one-experts-take-ed26bf9f9a21

Torres, Phil. (2018). Superintelligence and the Future of Governance: On Prioritizing the Control Problem at the End of History.

Trazzi, M., & Yampolskiy, R. V. (2018). Building Safer AGI by introducing Artificial Stupidity. *ArXiv:1808.03644 [Cs]*. Retrieved from http://arxiv.org/abs/1808.03644

Turchin, A. (2015). Using the Copernican mediocrity principle to estimate the timing of AI arrival. Retrieved June 10, 2018, from https://www.lesswrong.com/posts/txYsGGCvSn6MJeWDa/using-the-copernican-mediocrity-principle-to-estimate-the

Turchin, A. (2017). Messaging future AI. *Manuscript*. Retrieved from https://goo.gl/YArqki

Turchin, A. (2018). Narrow AI Nanny: Deceive Strategic Advantage via Narrow AI to Prevent Creation of the Superintelligence. *Manuscript*.

Turchin, A., & Denkenberger, D. (2017a). *Global Solutions of the AI Safety Problem*. manuscript.

Turchin, A., & Denkenberger, D. (2017b). Levels of self-improvement. *Manuscript*.

Turchin, A., & Denkenberger, D. (2018a). *Classification of Global Catastrophic Risks Connected with Artificial intelligence*. Under review in AI&Society.

Turchin, A., & Denkenberger, D. (2018b). Global Catastrophic and Existential Risks Scale. *Futures, in Press*.

Turchin, A., & Denkenberger, D. (2018c). Military AI as convergent goal of the self-improving AI. *Artificial Intelligence Safety And Security, (Roman Yampolskiy, Ed.), CRC Press*.

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*.

Tuwnier, J. (2018). 5 Best Bitcoin Mining Hardware ASICs 2018 (Comparison). Retrieved November 12, 2018, from https://www.buybitcoinworldwide.com/mining/hardware/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 6000–6010).

Vinge, V. (1993). Technological singularity (pp. 30–31). Presented at the VISION-21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute, VISION-21.

Von Foerster, H., Mora, P. M., & Amiot, L. W. (1960). Doomsday: Friday, 13 november, ad 2026. *Science*, *132*(3436), 1291–1295.

Waldrop, M. M. (2016). The chips are down for Moore’s law. *Nature News*, *530*(7589), 144. https://doi.org/10.1038/530144a

Walton, M. (2017, May 11). Nvidia Tesla V100: First Volta GPU is one of the largest silicon chips ever. *Ars Technica*. Retrieved from https://arstechnica.com/gadgets/2017/05/nvidia-tesla-v100-gpu-details/

Wiltz, C. (2018). The Battle of AI Processors Begins in 2018 | Design News. Retrieved November 12, 2018, from https://www.designnews.com/electronics-test/battle-ai-processors-begins-2018/212131505757984/page/0/1

Yudkowsky, E. (2017). There’s No Fire Alarm for Artificial General Intelligence. Retrieved January 22, 2018, from https://intelligence.org/2017/10/13/fire-alarm/

Yudkowsky, E., & Hanson, R. (2008). The Hanson-Yudkowsky AI-foom debate. In *MIRI Technical report*.