Catastrophically Dangerous AI is Possible Before 2030

Alexey Turchin

Science for Life Extension Foundation, Moscow

alexeiturchin@gmail.com

Version 2.0, 2021

[2021 UPDATE: The first version of this article was written at the end of 2018 and was published in 2019 under a different, less sensational name “Assessing the Future Plausibility of Catastrophically Dangerous AI” in [*Futures*](https://www.sciencedirect.com/science/article/abs/pii/S0016328718301319). Reviewers recommended that I should remove concrete timing predictions as it may cause some state agents to pursue AI capabilities. Now, this *informational hazard* seems to vanish as states already made their stakes on AI and other similar predictions have been published.

New data about language models scaling and hardware development could be used to update predictions. We’ve learned in the past 3 years that powerful universal AI is possible, that was GPT-3 moment, but also some earliest predictions about 2021-2022 became clearly not true. We also got *AI scaling laws* for predictions and *chip production crises* which affected the price and availability of compute. Based on all that the *earliest time* of Dangerous AI moved closer to 2025, but I remind the reader that the earliest time is not a *prediction of AI arrival timing*, but it is the first few per cents of the probability distribution. Given the intrinsic uncertainty of AI predictions we assume that the earliest arrival time is before 2030.]

**Abstract**:

In AI safety research, the median timing of AGI arrival is often taken as a reference point, which various polls predict to happen in the middle of 21 century, but for maximum safety, we should determine the *earliest* possible time of Dangerous AI arrival. Such Dangerous AI could be either AGI, capable of acting completely independently in the real world and of winning in most real-world conflicts with humans, or an AI helping humans to build weapons of mass destruction, or a national state coupled with AI-based government system. In this article, I demonstrate that the earliest timing of Dangerous AI, corresponding to 10 per cent of its arrival probability, is before 2030. Several partly independent sources of information are in accordance:

1. The growth of the *hardware available for AI research* makes human-brain-equivalents of compute available for AI research in the 2020s. It is fueled by specialized AI-chips, the use of many chips in one processing unit, and the larger research budgets, among other things.

2. The neural network performance and other characteristics, like the number of parameters, is quickly increasing every year, and extrapolating this tendency suggests that roughly human-level performance in a few years, around 2025.

3. Expert polls show around 10 per cent of the probability of an early appearance of artificial general intelligence (AGI) in the next decade, that is, before 2030.

4. Hyperbolic growth in different big history models converges around 2025-2030 (the technological singularity).

5. Anthropic arguments (similar to the Doomsday argument) suggest that *qualified observers* are more likely to appear near the end of the AI research epoch, as the number of such observers grew exponentially. This number doubles every 5-10 years, and thus we are likely to find ourselves around a decade before the end of AI research, which will happen consequently around 2030.

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

**Highlights**:

* The median timing of the AGI arrival is a wrong measure to use in the AI risk assessment; we should estimate the earliest plausible time of AGI’s arrival, like the first 10 per cent of distribution.
* Dangerous AI level is defined through the AI’s ability to facilitate a global catastrophe; it will be either AGI able to self-improve or AI which help us create dangerous weapons, like new biological viruses.
* The growth rate of hardware performance for AI has accelerated in the 2010s and can provide enough computational power for the Dangerous AI in the 2020s.
* The main measures of neural nets performance will reach a near-human level in the middle 2020s. This includes the number of parameters and perplexity of language models.
* Several trend-extrapolation methods predict some form of discontinuity around 2030.

[1. Introduction 3](#_Toc86616491)

[2. Earliest and median time of AGI arrival from the point of view of risk analysis 4](#_Toc86616492)

[3. Exponential growth of AI capabilities and declining computational complexity of human omnicide 5](#_Toc86616493)

[3.1. Dangerous AI 5](#_Toc86616494)

[3.2. Dangerous AI as independently acting agent capable to model the world and win over humans 5](#_Toc86616495)

[3.3. Dangerous AI as a designer of weapons of mass destruction 6](#_Toc86616496)

[3.4. Declining computational complexity of omnicide 6](#_Toc86616497)

[4. Different ways to estimate the earliest possible arrival of dangerous AI 7](#_Toc86616498)

[5. AI and hardware projection for next 5–15 years 7](#_Toc86616499)

[5.1. Hardware as a proxy of AI performance 7](#_Toc86616500)

[5.2. Effects of hardware on AI progress 8](#_Toc86616501)

[5.3. “Dangerous” level of AI Hardware 8](#_Toc86616502)

[5.4. Future of Moore’s law and the new ways to overcome its limitations via multi-chiplets processors and lower per-computation price 8](#_Toc86616503)

[5.5. AI-related “Moore’s law”: growing budget for AI research helps outperform Moore’s law 10](#_Toc86616504)

[5.6. Specialized hardware outperforms current trends in the growth of computing power 11](#_Toc86616505)

[5.7. Why dollar-per-flops of a GPU is a wrong metric of hardware performance 12](#_Toc86616506)

[5.8. Total computational power of the Internet grows quicker than the power of individual devices 12](#_Toc86616507)

[5.9. Conclusions about hardware 13](#_Toc86616508)

[6. Progress in neural net performance 13](#_Toc86616509)

[6.1. Neural nets revolution dramatically increases the speed of AI progress 13](#_Toc86616510)

[6.2. Neural nets’ performance metrics have doubled every year since 2012 14](#_Toc86616511)

[6.2.1. GPT-3 and the growth of language models 14](#_Toc86616512)

[6.3. Neural nets’ size 15](#_Toc86616513)

[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 15](#_Toc86616514)

[6.5. Signs of self-improving and knowledge transfer in ML 15](#_Toc86616515)

[6.6. There are other fields in AI research that could get a boost in the future 16](#_Toc86616516)

[6.7. Evolution of the “intellect stack” 16](#_Toc86616517)

[6.8. Jumps in algorithmic efficiency in neural networks 16](#_Toc86616518)

[7. Surveys of experts 16](#_Toc86616519)

[8. Historical analogy with the nuclear arms race 17](#_Toc86616520)

[9. All hyperbolic future growth projections converge around 2030 17](#_Toc86616521)

[10. AI timing predicted based on the randomness of the moment of AI creation 19](#_Toc86616522)

[10.1. The growth of the probability to find ground-breaking idea 19](#_Toc86616523)

[10.2. AI timing prediction based on my random location between the beginning of the research and its end 20](#_Toc86616524)

[11. Uncertainty in AI predictions 20](#_Toc86616525)

[12. Infra-human AI level, and its difficulties 21](#_Toc86616526)

[12.1. “Robotic brain” as next milestone 21](#_Toc86616527)

[12.1.1. What is the threshold of independent AI self-improvement and how far is it from the basic robotic brain? 21](#_Toc86616528)

[12.1.2 Could AI get a strategic decisive advantage without SI? 21](#_Toc86616529)

[12.1.3. What is the threshold of dangerously powerful AI relative to the robotic brain? 21](#_Toc86616530)

[12.1.4. Self-driving car as a pathway to the “robotic brain” 22](#_Toc86616531)

[12.1.5 Inference could be made cheap via special chips 22](#_Toc86616532)

[12.2. The non-human nature of human-level AI 22](#_Toc86616533)

[13. “Self-improving organizations” as a step to self-improving AI with humans in the loop 22](#_Toc86616534)

[14. Dangers of AI in near-term AI scenario 24](#_Toc86616535)

[14.1. Risks of Narrow AI 24](#_Toc86616536)

[14.2. Mildly superintelligent, non-self-improving AI gets strategic advantage – or start AI war 24](#_Toc86616537)

[15. What could be done to mitigate the risks of AI in the near future? 24](#_Toc86616538)

[Conclusion 25](#_Toc86616539)

# 1. Introduction

In 2017 the Machine Intelligence Research Institute (MIRI) updated its strategic view and assigned a higher probability to the idea that artificial general intelligence (AGI) will appear before 2035 (MIRI, 2017), without providing any numerical estimates[[1]](#footnote-1). This is a change from its previous assessments that creation of AI is very difficult and will not happen for 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.

But Shakirov (2016) estimated that AGI could appear even earlier based on his extrapolation of the artificial neural networks (ANNs) progress; he estimated that it could happen between 2021 and 2026. He made his extrapolation based on 2011-2015 period. Now, in the fall of 2021, it’s time to revisit the trends he predicted. In section 3 of his article he showed that ImageNet top-5 error was halving every year, from 25 % to 3.8 % in the period he explored. If we look at [current leader board](https://paperswithcode.com/sota/image-classification-on-imagenet?metric=Top%205%20Accuracy) in 2021, we could see that the error was 7.8 % in 2015 but 1.2 in 2020. (2 in 2019, 2.4 in 2018, 3.8 in 2017, 4.3 in 2016.) However, there was no records in 2021 better than beginning of 2020. Thus, the trend of halving the error generally continued, but last 1.5 years there is a slowdown. But superhuman level is achieved. He later updated his prediction to 70 % chance in the next 10 years, that is, before 2027 (Shakirov, 2017).

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). This clearly demonstrates that we are living in a period of explosive growth, which can’t be soon exhausted soon by a lack of available hardware, as AI-hardware grows quicker than Moore’s law, according to the newly coined “[Huang’s law](https://en.wikipedia.org/wiki/Huang%27s_law)”.

Huang stated in 2018 that AI-related hardware grew 25 times in 5 years, while Moore’s law predicted only 10 times growth for that period. The biggest part of that growth is not from flops per dollar, but from the improvement of the whole stack of hardware and software. Chip shortage during lockdowns in 2020-2021 and growing demand for computation probably screwed the price of computation per dollar as a measure of innovation in the field. But cloud computing lowered the price of computations of end users.

Another important metric is the amount of AI-training hardware available for a county, and China is [betting](http://www.caict.ac.cn/kxyj/qwfb/bps/202109/t20210918_390058.htm) on having advantage in it.

The observed acceleration of AI development and large uncertainty about AGI timing suggest that it is important to explore the probability of the appearance of AGI in the near future, but even here we can distinguish two time periods with important practical differences between them: 1) near-term: 0–5 years, and 2) medium-term: 5–15 years.

According to the MIRI strategy, an earlier horizon for the appearance of AI 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 AI safety theory, and some other solutions should be explored, as we have previously discussed (Turchin & Denkenberger, 2017a).

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

# 2. Earliest and median time of AGI arrival 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 arrival of human-level AI are distributed almost linearly in the beginning of the probability graph between 5 and 25 years from 2018, 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 mean (or median) time of predicted AGI arrival. However, from the risk analysis point of view, we need not the mean time, but the earliest time of AI arrival. It can be understood by analogy: we need to know not the mean time when 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 arrival, 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 arrival, which may be the reason for its unpopularity.

Thus, we need not median timing AGI arrival, 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 appearance (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 et al., 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 appearing is 2022. (We should update this estimate by the fact that AGI didn’t appear in 2021.) We assume that human-level AI will pose an existential risk as soon as it appears, but it may actually take years for AI to evolve into dangerous superintelligence.

# 3. Exponential growth of AI capabilities and declining computational complexity of human omnicide

## 3.1. Dangerous AI

We don’t know much about the probability distribution for AI, but one thing seems obvious: the probability of creation of AGI is growing with time. If we estimate the probability of AGI creation during some time period, it means that AGI will be more likely to be created by the end of this period. For example, if AI is expected in the 5 years period, it is more likely that it would appear during 5th year than in the first year.

Most probably, the probability of AGI creation is distributed exponentially in time, with the probability at any one time corresponding to the very steep logistic curve of the accumulated probability.

To escape theoretical discussions of AI completeness (having general intelligence), we introduce here the notion of *powerful 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.

## 3.2. Dangerous AI as independently acting agent capable to model the world and win over humans

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 dozen 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. We will call it a “robotic mind,” as most likely such AI will be developed to effectively control robots, but could also self-replicate in the Internet similar to computer viruses.

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 arrival time of such systems.

The danger of near-human level AI is that it could either effectively kill humans directly, or start an intelligence explosion via self-improving. For example, such a robotic mind could self-replicate in its billions and create a completely non-human economy, which outcompetes our economy (Alexander, 2016), or would require military action to stop it. 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).

## 3.3. Dangerous AI as a designer of weapons of mass destruction

We will define Dangerous AI as AI capable to accelerate creation of the weapons mass destruction by several orders of magnitude and thus making global catastrophe almost inevitable.

Advance AI is dangerous because it could create advanced weapons. Often it was suggested that it will be nanorobotics. Imagine that you get an Oracle AI, completely benign, but which is capable to tell you how to build a nuke from easy available components. Obviously, such AI will be dangerous if it appears in hands of any reckless person.

The idea of dangerous AI helps to escape the questions will it be AGI, conscious, human-level, agential, value-aligned. The main feature of Dangerous AI is its capability to accelerate the creation of complex real-world machines and writing software for them.

There are several types of weapons which could be accelerated by Dangerous AI:

* Biological weapon, mostly artificial viruses
* Nanotech: self-replicating robots and microscopic drones
* Cheaper and more powerful nukes
* Cheap manufacturing of very lethal autonomous drones
* Cyberweapons: computer viruses
* Social manipulation: human reward hacking, power grabbing.

A similar conception is *Transformative AI*: an AI which is capable to change the fate of humanity for better or worse (Gruetzemacher & Whittlestone, 2019). Dangerous AI is only about bad outcomes, so it could be simpler.

## 3.4. Dangerous AI a combination of state-level surveillance, nuclear weapons control and mild superintelligence

The danger of AI is measured by its capability to cause global effect, and older literature assume that pure intelligence is enough. However, neural-net-based-AI is mostly limited in its intelligence by human intelligence level which it “sucks” from training databases.

However, the neural net AI is also powered by the size of:

- data to which it has access. It doesn’t need to guess things, if you have access to sensors.

- money and other economic resources. AI doesn’t need to run away from its creators, if they are feeding it with all needed improvements and hardware.

- already exiting weapons in its control. AI doesn’t need to build weapons if it has access to them.

A national state with a large surveillance system and nuclear weapons is a powerful amplificatory of AI system. Such system may slowly drift into direction of dis-alignment and autonomy. A combination of military planner, economy planner and surveillance data analyzer is a step in the direction of this *automated system of government*. Given recent interest of national states to AI domination, this seems a plausible way to Dangerous AI, which doesn’t use less probable events as “self-improvement”, “treacherous turn” and “superintelligence”. I discussed more of this in “Levels of self-improvement”(Turchin & Denkenberger, 2017b) and “Military AI as convergent goal of self-improving AI” (Turchin & Denkenberger, 2018c).

## 3.5. Declining computational complexity of omnicide

To measure the risk of the AI we also introduce the notion of the *computational complexity of human extinction*. It is widely assumed that only superintelligence (Bostrom, 2014; Yudkowsky, 2008) will be able to create a global catastrophe which kills all humans (omnicide), for example, by creating dangerous nanotechnology. However, killing everybody is in some sense a *computational task*—to find the simplest way to do so—and this may be used for blackmail or for any other agential reasons which we ignore here; see (Torres, 2016) for more information.

The complexity of this task depends on the means available. For example, provoking global nuclear war may be relatively simple. There are several other possible ways to commit omnicide in computationally cheap ways, like creating a special biological virus, or some others we will not discuss here. However, these other ways seem technically feasible and do not depend on very advanced and computationally complex technology like molecular manufacturing or solving the protein-folding problem, contrary to the assumptions of Yudkowsky (2008). Instead, an AI may only need to find a way to rearrange several codons in the flu genome to create a superefficient pandemic virus, or several viruses (Turchin et al., 2017).

The *computational complexity of human extinction* as a task for AI is declining as we create more powerful tools. For example, the development of new systems to read and synthesize DNA simplifies the task of creating dangerous flu for any bad agent, human or AI.

Thus, the lower is the computational complexity of human extinction, the simpler is the powerful AI that could cause a global catastrophe directly because of a wrong programming or indirectly by helping a “bad agent”. This means that a smaller future complexity of omnicide could be reached with lower and lower intelligence. Thus, superintelligence may not be required to cause a global catastrophe, but perhaps narrow AI or universal AI just above human level. (If extinction complexity is below human level, humans may kill themselves even without AI, by mistake in the use of already exiting weapons like in the case of accidental nuclear war). This trend, combined with AI capabilities growing, means that the threat of omnicide is fast approaching.

To estimate the timing of the Dangerous AI, we need to find the *earliest point* in time where *AI capabilities will reach the computational complexity of omnicide* (it doesn’t mean that the catastrophe will happen at this moment, but it will become technically possible after that moment).

# 4. Different ways to estimate the earliest possible arrival 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 5–15 years

## 5.1. Hardware as a 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.

## 5.2. Effects of hardware on AI progress

The hardware available for AI depends on multiplication of the effects of the three trends:

1. Continuous progress in chips manufacturing, that is, Moore’s law and its possible successors.
2. Growth of the total budget on AI research, allowing researchers to buy or rent more hardware.
3. Advances in AI-specialized hardware, which includes semi-specialized (GPU) and AI-only hardware (TPU and neuromorphic chips).

OpenAI has published an article, “AI and compute” (OpenAI, 2018), where they showed that amount of computations used for the largest training runs of neural nets have been increasing with a 3.5-month doubling time and have increased 300 000 times from 2012 to the beginning of 2018. Such growth can’t be explained simply by improvements in hardware, but accounts for both the effects of the growth in funding and appearance of specialized hardware. Thus, it is a combined measure of all three trends. OpenAI conclude that the trend is likely to continue.

## 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. However, we are already above this level, as supercomputer Summit has 200 petaflops, and 3 exaflops in deep learning operations. However, 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.

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

Despite computer lithography is close to its limits, the progress in it is expected to continue in 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.

Actually, computer chips become cheap when they become obsolete; one of the first processor 8080 price fall from hundreds of dollars to cents in a decade [ref]. So, looking on the newest processors for cheapest computations is misleading. In 2020s we don’t observe this as we are in the chip shortage period; however, if no war in Taiwan will happen, this shortage will be overcome by market forces, as new fabs are building.

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.

A lot of work has been done already in assessing the expected growth in hardware. Grace estimated that “hardware prices (for single precision flops/$) appear to be falling by around an order of magnitude every 10-16 years” (Grace, 2017a), which seems to be a rather slow prediction, indicating that AGI will appear in the long term, that is, after 2035, however, appearance of GPU and other hardware acceleration methods allowed more *AI-related computation power* (matrix multiplication) from the same number of transistors on a chip.

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). Cerebras [built](https://venturebeat.com/2021/04/20/cerebras-systems-launches-new-ai-supercomputing-processor-with-2-6-trillion-transistors/) an AI-chip with 2.6 trillion transistors in 2021.

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 like Graphic Processor Units (GPU) started to use several chips ([chiplets](https://en.wikichip.org/wiki/chiplet)) 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).

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, while manufacturing costs and energy consumption could diminish.

While achievement of the smaller transistor’s size seems to slow down as new lithographic technologies becomes more expensive and implemented slowly, growth of cheap available computing power continues, because of:

1. Mass production and economy of scale.
2. Graphical processors are much easily paralleled, and single processing unit could now consist of several chips.
3. For AI applications, specialized hardware may be used which provide higher performance in AI related operations.

All this fuel rapid reduction of prices of hardware which is related to AI.

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.

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:

* *GPU growth*. NVIDIA head Huang predicted that GPUs will outperform CPUs 1000× from 2010 to 2025, with their performance increasing 1.5 times every year (Patterson, 2017) and this trend mostly holds until at least 2018.
* *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). The startup Graphcore promises a 100× increase over current performance with its “intelligent processing units” (Graphcore, 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 et al., 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 et al., 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 (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).

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

## 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 Researсh, 2017). To support such rate, largest competitors need to invest in R&D even more. 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*.

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. Smaller computers require less space for datacenters.

## 5.6. Specialized hardware outperforms current trends in the growth of computing power

The revolution in artificial neural nets came largely because of the growth of performance of GPUs, which was at one time specialized hardware for graphics-heavy applications such as computer games. In the 2010s, this formerly specialized hardware started driving computational performance and AI research. There are two types of AI-specialized hardware: semi-specialized, that is GPUs, which can also be used for other proposes; and fully specialized, i.e. ASICs (application specific integrated circuits). ASICs can provide performance several orders of magnitude higher if the exact type of needed calculation is fully known. In the current field of AI many different architectures are constantly tested, thus time of fully hardwired ASICs has not come. However, they could provide very large computational gains in inference of the already trained neural networks if they will be realized as physical neural network.

The power of ASICs could be illustrated by bitcoin mining networks as an example of how specialized hardware could be created in time to solve complex computational task when such a task is specified. (We remind the reader that we mention the bitcoin network only as an illustration, and do not claim that such a network could become AI or be used for other types of distributed calculations, as its data exchange rate is slow.)

In January 2018, the Bitcoin network reached a computational power of 8 exahashs that is almost three times higher than at the beginning of 2017. It reaches 25 exahashss in March 2018 (Redman, 2018), 37 in June 2018, 61 in September 2018, i.e., it has grown almost eight times during the writing first version of this article. In 2021, it was 200 exahashs, so its growth has slowdown.

Each hash is a special mathematical function that requires many calculations (Connell, 2017). The net computational power doubling time has been around eight months for the last several years. If one hash were calculated on an ordinary computer, it would represent 12 000 FLOPS (Siluk, 2013). Thus, the total computational power equivalent of the bitcoin net would require 1023 FLOPS in the beginning of 2018, 60 times above the total upper limit of the calculation power of entire Internet as estimated by Grace in 2016. The total cryptocurrency network is even larger, as Bitcoin is only one of many “coins”.

While bitcoin network is specialized for some type of calculation, the deep learning AI also relies on another special type of calculations and thus approximately the same size computer infrastructure could be built in a few years if similar financial initiatives will appear. The important difference is the need for high-speed information exchange between nodes, which could be partly solved by implementing 5G networks in the 2020s, with data speeds around 10 GB/sec. It is more likely to be addressed by implementing deep learning algorithms, which will enable more parallel training with less data transfer between nodes, like the evolution strategies suggested by OpenAI (Salimans et al., 2017).

Google’s *Tensorflow* TPU chips are one example of specialized ASICs for ANN. Some bitcoin hardware manufacturers have already turned to building specialized neural net chips (Businesswire, 2017).

## 5.7. Why dollar-per-flops of a GPU is a wrong metric of hardware performance

If we compare GPU from 2010 and 2020 we could be surprised that the number of flops per dollar didn’t grow that dramatically. GPU grew from around hundreds of billions to hundreds of trillions of operations per second, but their price also grew from 100USD to 10 000, so total growth of the amount of computation a dollar can buy is not that impressive.

One reason why it is wrong metric is that for any GPU, one need a computer, a space where this computer is located, and electricity. A whole computer price is around 1000 USD, so the total price of the whole system with circa 2010 GPU will be 1100 dollars, and for 2020 it will be 6000 dollars.

We need to take the *whole costs* in account to calculate total effectiveness of a GPU. Another important cost comes from the space where the computer is located. For a home computer of 2010s it is free, but as soon as we want something larger, it is a small data center. Small datacenters are not very cost-effective. For example, if sq. meter of a building costs 5000 USD, and the small datacenter with 10 old style GPU requires a full room of computers, the price of such room itself will be around 100.000 USD, and it is just for 10 GPU, each of which costs 100 USD.

Such system will be also very energy-inefficient as there will be 10 computers, a commutation system between them and air-conditioning, which will be minimum 10kW of electric power. The price of electricity [varies](https://www.datacenterknowledge.com/archives/2016/08/23/what-is-the-data-center-cost-of-1kw-of-it-capacity) with data centres’ size: “The average annual data center cost per kW ranges from $5,467 for data centers larger than 50,000 square feet to $26,495 for facilities that are between 500 and 5,000 square feet in size.” Datacenters also could be located in places with lower price of electricity.

A problem with owning a small data center is that you pay all cost *upfront*, but most time you will not use the system or it will be not enough at peaks of usage, and it will obsolete quickly. Larger GPU in data-centers allow a person to pay as you go.

Thus, you need either just one very powerful GPU at home assuming that home is free, or it should be installed in a very large data center to be cost-effective. This explains explosion of cloud computing.

Largest consumers like Google, Nvidia and Tesla create their own hardware for their need and thus don’t pay for marketing.

There is another advantage of larger GPU: they have larger RAM which limits the size of models, and they have better communication between different modules, which helps to create even larger models.

All this means that even if the price for *installed flops* is the same, larger and more expensive GPUs produce significant economy as they require less space and energy and could train larger models.

## 5.8. Total computational power of the Internet grows quicker than the power of individual devices

For AI creation, not only is the price of computation important, but also the *total available computational power*. Even if Moore’s law is dead, the total available computer power may grow *linearly* as new hardware is built; similarly, the total computational power of the Internet is growing because the number of connected devices is growing as well as their *connectivity*. Grace calculated that the current speed of this growth is rather slow, based on the price of installed hardware, estimated at 2 × 1020–1.5 × 1021 flops, with grow of around 25 percent a year (Grace, 2016).

Another way to calculate the total computing power of the Internet is to calculate the actual number of connected devices and multiply it by the mean level of their computational power. However, most new connected devices are smartphones, which until recently had much less computing power. The iPhone X (2017), though, has a capacity of around 300–600 gigaflops, and is estimated to be more powerful than the newest MacBook Pro (Smith, 2017). The 2021 model with Bionic A15 chip [has](https://3dnews.ru/1049106/mobilnie-protsessori-apple-stali-medlennee-pribavlyat-v-proizvoditelnosti-i-dalshe-budet-tolko-huge) 15 trillion flops neural engine on board. The first iPhone, released ten years ago (2007), had a performance of only one 1 × 109 flops (Ojala, 2015). The number of smartphones in use is projected to grow around an order of magnitude between 2010 and 2020 (Ericsson, 2014).

 This means that the total computational power of mobile networks grew approximately 3 orders of magnitude in 10 years, doubling every year. In 2020, billions of smartphones capable of performance similar to that of the iPhone X will be in use, meaning smartphones only may represent computing power of around 1021 flops. While smartphone markets will soon reach saturation, there are other markets of connected devices ready to explode, that is IoT and self-driving cars.

## 5.9. 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. Not only Moore’s law didn’t die, but the price of computations started to decline in the much steeper rate after 2015 because of advances of GPU and TPU. This was overshadowed by high demand for computation and chip shortage in early 2020s; in other words, higher prices are driven by the demand from crypto and growing economy.

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.

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.

However, training very large neural nets by conventional means could become prohibitly expensive because of scaling laws [ref gwern?].

From this follows that, the main limit for AI becomes algorithm’s performance, not hardware.

# 6. Progress in neural net performance

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

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

After the 2012, the performance of many metrics of neural networks 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 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 computer understanding of text, but it still below human performance.

The quality of machine translation [in English-French pair](https://paperswithcode.com/sota/machine-translation-on-wmt2014-english-french) increased, in BLEU score, from 37 (2014) to 41 (2017), while the professional human level is above 50. It grew to 46 in 2021. At this rate of progress, it would require six years (2023) to reach a point equal to human-level performance.

Performance on the *Stanford Question Answering Dataset* has increased from 80 (2016) to 85 (2017), and to 90.9 in 2021, while the average human level is 93. It suggests that human level will be achieved in 2020s, despite some diminishing returns.

However, the progress in AI is not smooth: some fields generate most of return. In recent AI research, it was transformers and language models.

### 6.2.1. GPT-3 and the growth of language models

The most dramatic growth in AI area closes to AGI has happened in language models. In 2015 Karpathy wrote famous article “[Unreasonable effectiveness of recurrent neural networks](http://karpathy.github.io/2015/05/21/rnn-effectiveness/).” He shows that RNN with only 3 million parameters is able to generate grammatically correct text (Karpathy, 2015).

The progress in text-generating models was:

Karpathy’s LTSM: 2015: 3.5 mln parameters

GPT-1 June 2018: 110 mln parameters

GPT-2 in Feb 2019: 1.5 billion parameters

GPT-3 in May 2020: 175 billion

Megatron-Turing Oct 2021 530 billion parameters

If we take only GPT family, it grew with speed like 2 order of magnitude a year, but GPT-4 was not yet published, which means that it is either too strong or too weak, or too expensive to train. There were [rumors](https://towardsdatascience.com/gpt-4-will-have-100-trillion-parameters-500x-the-size-of-gpt-3-582b98d82253) that 100 trillions of parameters GPT-4 model is in work, but will be ready in a few years.

If we take other data points, the growth of language models is more modest one order of magnitude a year and GPT-3 was an outliner. Anyway, even if language models grow 1 order of magnitude a year, they will reach 100 trillions parameters to 2024.

Mattew Barnett [wrote](https://www.metaculus.com/notebooks/8329/human-level-language-models/?fbclid=IwAR0LqBcSP2_x7AyZkad3YUlbJ6H2d_nI_T2wP-oDZ4Y3uGkadEk2PjCX5Dg) in 2021: “My result is a remarkably short timeline: Concretely, my model predicts that a human-level language model will be developed some time in the mid 2020s, with substantial uncertainty in that prediction.” He used analysis of text entropy.

Gwern [explored](https://www.gwern.net/newsletter/2020/05#gpt-3) scaling laws for language models and the growth of compute trends. He [shows](https://www.gwern.net/newsletter/2020/05#gpt-3) that GPT-3 is a big deal and that sub-human performance will be reached in 2020s. GPT-3 could be improved in many ways, so its performance is not the end. He expects that human level performance will require million time more compute than current models. He [wrote](https://www.lesswrong.com/posts/SZ3jDHXHb4WF4jmbr/where-is-human-level-on-text-prediction-gpts-task): “If we assume compute follows the current trend of peak AI project compute doubling every 3.4 months, then 2.2e6× more compute would be log2(2.2e6) = 22 doublings away - or 22\*(3.4/12) = 6.3 years, or 2027. (Seems a little unlikely.)” Lanrian also [got](https://www.lesswrong.com/posts/k2SNji3jXaLGhBeYP/extrapolating-gpt-n-performance#Comparisons_and_limits) 1 000 000 times increase of computation is needed to get from GPT-3 to human-level. But such amount of compute would cost more than trillion dollars in current prices and unlikely to be implemented. If we account for expected lowering of prices, it will be achievable only in 2030s.

GPT-n can’t be AGI itself, but it is very close to, as it is universal. There are several obvious ways how it could be augmented: by larger memory, by visual processing, as a robot mind or by a truth predictor.

## 6.3. Neural nets’ size

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, but they [estimated](https://static.googleusercontent.com/media/research.google.com/en//archive/unsupervised_icml2012.pdf) that human visual cortex is has million times more synapses. The Google’s result is outlier, which just demonstrated the power of technology. In ImageNet [leaderboard](https://paperswithcode.com/sota/image-classification-on-imagenet), the size of NN grew from 60 millions for AlexNet in 2013 to 14 billion in 2021 for a mixture of experts net, and the next one is 2.4 billions. So, the growth is relatively modest 50-250 times for 8 years, and at this speed 100 trillions of parameters will be achieved in 10-15 years, or in the middle of 2030s.

But there are reasons to assume that there is a hardware overhang. In 2020 a few companies declared the capability to train neural nets up to 100 trillion parameters in size (Cerebras, Microsoft [ref]).

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.

## 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 et al., 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.

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).

## 6.5. Signs of self-improving and knowledge 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 have demonstrated better ability to learn (Shi et al., 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 intellect stack, 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 promises 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 et al., 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 et al., 2018).

# 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 et al., 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 arrival 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.

*Metaculus* prediction market in 2021 [gives](https://www.metaculus.com/questions/3479/when-will-the-first-artificial-general-intelligence-system-be-devised-tested-and-publicly-known-of/) 25 per cent of AI arrival before 2031 and 50 per cent to 2045.

 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 didn’t appear in the last 10 years in 2010s.

# 8. Historical analogy with the nuclear arms race

A possible historical analogue of the current situation in AI is the nuclear arms race in the 1940s. From the discovery of radioactivity in 1896, nuclear research was rather slow and minimally funded. The idea of the nuclear chain reaction was invented by Szilard in 1933 (Smith, 2007), and it was only after discovering the ability of uranium to fission in 1938 scientists understood that nuclear weapons could be created. This motivated Einstein and several other scientists to write a letter to Roosevelt in 1939 which eventually gave rise to the Manhattan project in 1941. The financial and scientific investments in nuclear research grew by many thousand-fold at that time.

The actual start of Manhattan project was in January 1942; it required only 3.5 years to create the first nuclear bomb. It was around 6 years after the creation of the first nuclear bomb when the cobalt bomb, which could be able to kill all humanity, was conceived by Szilard (Smith, 2007). Teller worked on a 10 gigaton (not mistake) bomb around 1960, until the project was canceled (Wellerstein, 2012). So, from Einstein’s letter it took only 6 years to create the first bomb and 23 years until the Cuban missile crisis and the multimegaton bombs of the 1960s.

The accepted date of the beginning of the deep learning revolution is 2012 (Felsberg, 2017), and its effect could be compared to the discovery in uranium’s ability to fission in 1938. The importance of AI as a weapon has become more and more accepted (De Spiegeleire et al., 2017). Putin said in 2017 that the creator of AI will be the ruler of the world, and China sees an AI as a unique strategic opportunity for global domination (Putin, 2017).

 So, we see that after the idea of a new powerful weapon was conceived and supported by an arms race, it took 4–6 years to weaponize it. It was thought there might be a risk of igniting Earth’s atmosphere during first nuclear test, later proved false. However, in 10–20 years global catastrophic nuclear war became a real possibility.

 2017 can be thought of as the new 1939, when major players started an arms race to nuclear power. China wants to become the main player (Ding, 2018). If the nature of arms races is generalizable, it would mean that military-grade AI—like a complex system of espionage, planning army management and military robots as described in (De Spiegeleire et al., 2017) and universal AIs potentially able to catastrophic “chain reaction” of self-improvement—will appear somewhere in early 2020s. The date coincides with our estimates for neural net development to near-human level.

# 9. All hyperbolic future growth projections converge around 2030

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 et al., 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. Covid pandemic started in 2020 and it changed the way we live, but there is no other signed of significant acceleration of history at the end of 2021.

 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 than Diakonov-Panov. He also predicted that AI would reach superhuman visual recognition in 2020, which appears to be true.

Predicted mean timings based on historical data are presented in the bottom row of the 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 et al., 2016).

 Acceleration moved from Moore’s law to neural nets’: 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 appears, 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.

# 10. AI timing predicted based on the randomness of the moment of AI creation

## 10.1. The growth of the probability to find ground-breaking idea

Yudkowsky (Yudkowsky & Hanson, 2008) suggested that the appearance 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.

Thus, the number of AI researchers is currently growing exponentially, increasing around an order of magnitude every decade, with a doubling period of around 2.2 years, surprisingly similar to the rate of Moore’s law. 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 10–15 years. If we assume the start of AI hype in 2012 and “normal”, “for sure” predicted timing of AI as up to the end of 21 century, as it appears in some polls, then such an increase of the number of researchers because of hype will move the first appearance of AI to 2022–2027.

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.

Alternative view is “bitter lesson” that not an idea, but raw computational power ensures progress in AI. As we show above, we will have enough computational power to train and run human-level AI in 2020s.

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

There is an even more speculative way to predict the future based on the logic similar to the Doomsday argument, which the author have 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 appearance by Grace’s poll (Grace et al., 2017)). However, this prediction is based on the assumption of a linear distribution of the probability of AI appearance within the entire period and of our observation of that 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.

Alternatively, if we look at the probability that we will ask the question about AI timing using DA, it is distributed not randomly, as there are more people who knows anthropic reasoning to the end of that period. The idea of anthropics appeared only in 1973 when it was suggested by Carter, and becomes popular in 1990s and especially after advent of the internet and LessWrong blog where it is often discussed. Thus, I am exponentially more likely to be located closer to the end of the period when AI is in development, which implies either AI creation in 2020s, or the end of the world by some other global catastrophe soon; doubling period of interest to anthropics is between 5 and 10 years (Turchin, 2018)

This is just a 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 arrival.

# 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 arrival for Dangerous AI in the period around 2025, 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.

When we say “around 2025” there is still uncertainty plus or minus a few years around that date, and thus we reframe as more valid “before 2030”. However, if we want to create effective instruments to keep Dangerous AI at bay, we have only a few years, not decades.

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 appear. As it is not easy to quantify the level of the probability above which we should worry, or the actual probability distribution of AI appearing, 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 appearing before 2030, which corresponds to the “yellow level” in the color-coded scale of risks (Turchin & Denkenberger, 2018b).

# 12. Infra-human AI level, and its difficulties

## 12.1. “Robotic brain” as next milestone

The ultimate milestone of neural nets development is something akin to *home robot brain*. It will be mostly driven by large demand from self-driving cars, military and home robotics. It will combine several features, like the ability to understand commands in human language in correct context, create a model of the world based on the sensory information and support walking. Basically, it is AGI built from known algorithms (see https://arbital.com/p/KANSI/).

The “Turing test” for such robot is its ability to prepare a breakfast. It still may be not conscious, or will be bad on modeling human emotions or philosophy, but it needs human language understanding + human world model understanding. In 2021 we seems to have first ingredient in the form of language models. Google suggested in 2021 “pathways” as an approach to multisense AI, and Open AI also is working on multimodal GPT.

Such system could be built by fine-tuning many already existing algorithms and will probably not require solving any secret problem of the intelligence. The road to such system in next 5 years seems pretty straightforward, but what will happen next is less clear.

### 12.1.1. What is the threshold of independent AI self-improvement and how far is it from the basic robotic brain?

The AI on robotic brain level will not able to independently self-improve, but surely will able to create self-improving loops together with humans. This will give traction to power the self-improving organizations.

Several problems must be solved before unlimited self-improvement, which I discuss in “Levels of Self-improvement” (Turchin & Denkenberger, 2017b), including the problem of correct measurement of improvement and bugs prevention. Both requires growing amount of testing, especially for “messy AI”, which can’t be mathematically proved, and could limits its improvement ability, by either increasing failure rate or exponential growth of internal testing time. One type of errors is that situation when presumably infrahuman AI is failed to replicate expected human behavior, or human level judgments etc.

So, the robotic brain needs to be hand-polished before it will be able mimic humans 99.9 times. The problem seams eventually solvable, but it may require years to go from infra-human level to above-human performance.

### 12.1.2 Could AI get a strategic decisive advantage without SI?

AI may be used as geopolitical game-solver, but correct world-model is needed.

### 12.1.3. What is the threshold of dangerously powerful AI relative to the robotic brain?

It seems that even narrow AI may become very useful and cheap instrument to build dangerous weapons in bio and nanotech. Robotic brains could be safe robot-solders, but some bio-oriented narrow AI may be dangerous instrument in the hands of bioterroists.

Thus, first types of powerful Dangerous AI may appear in 5 years, but independently self-improving AI is like 10 years from now, closer to 2030.

### 12.1.4. Self-driving car as a pathway to the “robotic brain”

Despite the name “robotic brain”, the most direct way to the near-human AI is likely not from human-like robots (recently, 2021, promised by Tesla) and even not from robodogs, but from the self-driving cars.

The reason for that that the car must act in the real time and in real world, and there is a strong market demand for full self-driving. But level 5 self-driving is likely a AGI-equivalent task, as it requires very good understanding of the world, as well as text commands from the driver, police and street signs. Self-driving cars also require ethics from the start, as their decisions could kill people.

### 12.1.5 Inference could be made cheap via special chips

Training AI is the most expensive thing, but inference also costs much. However, if no changes are expected in the trained AI, the resulted neural net could be made lithographically as single chip where each neuron and weight is in hardware. There are other ways for cheap inference including optic chips.

## 12.2. The non-human nature of human-level AI

Often, AI reaches human-level or above human-level performance using completely non-human ways of optimization. For example, the brute force search algorithms could find a winning strategy in some game, like chess, completely not using human style of thinking, like the use of symbols, planning etc. In the most cases, where AI is comparable with humans in result, it is completely non-human in the ways how it reaches them.

The main problem of non-human human-level AI (NH-HLAI) is that it makes mistakes in completely non-human ways, which creates unpredictable risks.

The “good convergence” hypothesis: universal robotic mind converges into human-like AI.

In other words, to be intelligent is the same to be a human, and any non-human intelligence will be less effective in human-important areas, like the actions in human world. It is good, because it will be understandable for humans, and if it fails, it will fail in the ways obvious to human, so it could be predicted, detected and avoided. However, the hypothesis probably not true, and when the opposite is true:

“Bad alternative” hypothesis:

Universal human-level AI is possible by completely non-human ways of optimization and problem-solving, and it is also cheaper and simpler than modeling a human mind.

In that case, its performance will be not always equal to human but will be very superhuman in some areas, and inferior to humans in other. For example, better in math, but worse in common sense logic. Such AI will be dangerous.

# 13. “Self-improving organizations” as a step to self-improving AI with humans in the loop

Biggest AI progress is now happening inside "self-improving organizations" where they create software that helps them to design another software and the main example of them is Google and Elon Musk’s cluster of startups (Tesla, Neurolink and SpaceX), which are very conscious about constant self-improving.

 Google [created](https://research.googleblog.com/2017/05/using-machine-learning-to-explore.) neural nets able to design other neural nets, but humans are still needed in the loop in many steps of such proto self-improving process. "In our approach (which we call "AutoML"), a controller neural net can propose a “child” model architecture, which can then be trained and evaluated for quality on a particular task.” html

The advantages of SI-organizations in AI creation:

• Use of additive nature of human intelligence.

• Use of money and market forces to attract best minds.

• Large existing computer net and database.

• Will of the goal-oriented leader.

• Pressure from the market forces.

• If AI ideas appear unexpectedly (like black swan), organisations still have advantage, as they already hired best minds or could buy startups.

• Easy money for marginal costs for the realization of the projects.

• Able to design their own hardware

• Able to change internal structure to get maximum effect (social engineering).

• Could test many ideas simultaneously.

• Could use corporate culture to affect minds of people and make them more effective.

• Could use large non-AI related revenues to fund long-term research projects.

• Highly prestigious employers are able to attract best minds.

• Use of PR and promises to attract resources (Musk).

Rivals of the self-improving organization in AI creation:

• Individuals

• Startups

• Universities

• Governments

Self-improving organization and AI risks:

• AI is distributed, so most boxing solutions will not work.

• Internal theft or value function hacking.

• In the case of China, largest IT companies may have larger level of integration with government and defense. The same is true for Russia, and much less for US. Baidu is de facto owned by government, which may compensate their lag to US in creation of the military AI.

• SI-organizations are probably effective in acceleration of AI research.

Benefits from the point of AI risk:

• SI organization could get the decisive advantage without eliminating humans from their jobs inside the system.

• Public opinion pressure makes them create AI safety advisory boards.

• They can’t pursue dangerous (and non-economic, that is not about money) goals legally (“don’t be evil”). But military AI and small secret groups could have global non-economic goals. However, large pressure of companies exists now to promote important social topics, like climate change.

The problem of integration with government defense and military:

• Most likely Baidu is directly and indirectly owned by the government of the People's Republic of China

# 14. Dangers of AI in near-term AI scenario

## 14.1. Risks of Narrow AI

I explored such risks in details in the article “Classification of the global risks of AI” (Turchin & Denkenberger, 2018a). In short, there are several main risks:

1) Narrow AI viruses affecting hardware globally (Example: Self-driving cars start to hit humans; planes crashes into nuclear power stations.)

2) Failures of military AI, including accidental nuclear war, or war between side in AI arms race (and losing party may use nuclear weapons again wining party, as it will soon lose its excellence)

3) Narrow AI helps to create other powerful weapons, which could be used as instruments of war or terrorism. (Example: narrow AI helps bioterrorist to construct deadly virus).

As we discussed above, computational complexity of the task of human extinction is diminishing, and the power of AI problem-solving ability is increasing. It means that in some moment AI will be able to solve such tasks.

However, the ability does not mean necessity, as “doomsday agent” (Torres, 2016) or chain of events is required, which would result in dangerous application of AI or it’s misuse. For example, global nuclear war was possible for decades, but didn’t happen.

Anyway, if powerful AI system will be widely available, there will be many agents, and the law of big numbers will result in all possible applications.

## 14.2. Mildly superintelligent, non-self-improving AI gets strategic advantage – or start AI war

Basically, it will be what Paul Christiano called “prosaic AI”. Calling it *mild superintelligence* we mean that it is 100s or 1000s of times above human level (but not trillions) and that its abilities are not uniformly above human level, so it may need assistance form human in many tasks (from manufacturing to programming).

Such system may get strategic advantage in military terms by constructing to their owners powerful weapons (robots, bio and cyberwepons as well as effective strategic planning and game winning ability.) Strategic advantage is a *pivotal act* which gives one power over the world or result in global war for AI domination, of there are 2 rival AIs.

The *war between two superhuman AIs* is a global risks situation (even if they remain aligned with goals of their creators), because it would produce very quick, unstable arms race, which would produce weapons, enormously exceeding nuclear weapons and because both sides could blackmail each other using Doomsday weapons. If there are more than two sides, the situation will be even more unstable, though coalitions are likely.

A combination of even mild superintelligence with resources of a national state with nuclear weapons, like China, is major shift of global power.

Even if such prosaic superintelligence is under control of humans, it will drift away from it as war and arms race will accelerate. It may have to take quicker and complex decisions which it will not have time to explain to humans.

# 15. What could be done to mitigate the risks of AI in the near future?

First, what we can’t do: there is no time to create perfect AI from scratch or perfect AI safety theory for that AI and persuade all players to implement it. Maybe extreme acceleration of such efforts will be able to produce perfect benevolent and rally superintelligent AI which will be able to unite humanity peacefully.

There is also impossible to ban AI, because it is seen as a weapon.

One idea which has merit is to concentrate research of AI in UN, like creating CERN for AI.

# Conclusion

Our analysis of current computer hardware and neural net development implies that Dangerous AI could appear in 2020s. 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 and Jonathan Yan for thoughtful discussion of the topic. No external funding sources for this study.

Literature:

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

Alexander, S. (2016). *Ascended economy?* Star Slate Codex. http://slatestarcodex.com/2016/05/30/ascended-economy/

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

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., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., … Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. *ArXiv:1806.01261 [Cs, Stat]*. 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.

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

Bostrom, N. (2014). *Superintelligence*. Oxford University Press.

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*. http://www.pbs.org/wgbh/nova/next/tech/brain-mapping-supercomputer/

Businesswire. (2017, October 25). *Bitmain Introduces Its First Hardware for Accelerating Artificial Intelligence (AI) Applications*. https://www.businesswire.com/news/home/20171025005379/en/Bitmain-Introduces-Hardware-Accelerating-Artificial-Intelligence-AI

Cannell, J. (2015). The Brain as a Universal Learning Machine. *LessWrong*. 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., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). Universal Sentence Encoder. *ArXiv:1803.11175 [Cs]*. http://arxiv.org/abs/1803.11175

Chklovski, T. (2017, May 22). Artificial Intelligence—What should we tell our children. *Huffington Post*. https://www.huffingtonpost.com/entry/artificial-intelligence-what-should-we-tell-our-children\_us\_59235f2de4b0b28a33f62ecc

Connell, J. (2017, January 16). Why Bitcoin Is Close to a Record-Breaking 3 Exahashes Of Processing Power. *Bitcoin News*. https://news.bitcoin.com/why-bitcoin-is-close-to-a-record-breaking-3-exahashes-processing-power/

Constantin, S. (2017, January 28). Performance Trends in AI. *Otium*. https://srconstantin.wordpress.com/2017/01/28/performance-trends-in-ai/

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)*. https://www.darpa.mil/attachments/3DSoCProposersDay20170915.pdf

De Spiegeleire, S., Maas, M., & Sweijs, T. (2017). *Artificial intelligence and the future of defence*. The Hague Centre for Strategic Studies. http://www.hcss.nl/sites/default/files/files/reports/Artificial%20Intelligence%20and%20the%20Future%20of%20Defense.pdf

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

Ding, J. (2018). *Deciphering China’s AI Dream*.

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*. 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*. 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*. https://www.top500.org/news/thomas-sterling-talks-exascale-chinese-hpc-machine-learning-and-non-von-neumann-architectures/

Felsberg, M. (2017). *Five years after the Deep Learning revolution of computer vision: State of the art methods for online image and video analysis*. Linköping University Electronic Press. http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-143676

Fortunelords. (2016, March 19). Youtube Statistics—2018. *Digital Marketing Education*. https://fortunelords.com/youtube-statistics/

Fujitsu. (2016). Fujitsu Laboratories Develops New Architecture that Rivals Quantum Computers in Utility—Fujitsu Global. *Fujitsu Laboratories Ltd.* 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., Lou, X., Meng, Z., Liu, Y., & 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. *Research Blog*. 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*. https://aiimpacts.org/brain-performance-in-flops/

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

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

Grace, K., 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 Researсh. (2017). *Artificial Intelligence Market Size Analysis | Industry Growth Report 2025*. https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market

Graphcore. (2017). *Graphcore’s technology for accelerating machine learning and AI*. https://www.graphcore.ai/technology

Grey, A. (2017). *These charts will change how you see the rise of AI*. World Economic Forum. https://www.weforum.org/agenda/2017/12/charts-artificial-intelligence-ai-index/

Gruetzemacher, R., & Whittlestone, J. (2019). Defining and unpacking transformative AI. *ArXiv191200747 [Cs]*.

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*. https://spectrum.ieee.org/computing/hardware/how-ibm-got-brainlike-efficiency-from-the-truenorth-chip

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

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

Karpathy, A. (2015). The unreasonable effectiveness of recurrent neural networks. *Andrej Karpathy Blog*. 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*. Home. //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), 157–168.

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

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.

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? *Ailev*. https://ailev.livejournal.com/1382464.html

Mannes, J. (2017). *Intel to ship new Nervana Neural Network Processor by end of 2017*. 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].

MIRI. (2016, May 4). A new MIRI research program with a machine learning focus. *Machine Intelligence Research Institute*. 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

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.

Nichol, A., Achiam, J., & Schulman, J. (2018). On First-Order Meta-Learning Algorithms. *ArXiv:1803.02999 [Cs]*. 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*. https://phys.org/news/2017-12-gallium-nitride-processornext-generation-technology-space.html

OpenAI. (2018, May 16). AI and Compute. *OpenAI Blog*. https://blog.openai.com/ai-and-compute/

Ormerod, P. (2007). Cascades of failure and extinction in dynamically evolving complex systems. *Noise and Stochastics in Complex Systems and Finance*, *6601*, 66010Q.

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*. https://www.eetimes.com/document.asp?doc\_id=1331836

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]*. http://arxiv.org/abs/1804.03867

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

Putin, V. (2017). *Open lesson “Russia looking to the future.”* http://kremlin.ru/events/president/news/55493

Redman, J. (2018, March 18). BTC Prices Below $8K Could Spell Disaster for Bitcoin Miners. *Bitcoin News*. https://news.bitcoin.com/btc-prices-below-8k-could-spell-disaster-for-bitcoin-minersbtc-prices-below-8k-could-spell-disaster-for-bitcoin-miners/

Rosebrock, A. (2018, July 2). An interview with Francois Chollet. *PyImageSearch*. 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]*. 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*. Google Cloud Platform. 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*. http://people.idsia.ch/~juergen/history.html

Schwab, K. (2015, December 12). The Fourth Industrial Revolution. *Foreign Affairs*. 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*. https://arxiv.org/abs/1605.04232

Shakirov, V. (2017). *AI progress metrics*. Google Docs. https://docs.google.com/spreadsheets/d/11C1CT3G-aAgKILpI9kHzcgg\_jQ6oF3gcaGoFTdAqwF8/edit?usp=embed\_facebook

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*.

Siluk, S. (2013, May 13). *Bitcoin network out-muscles top 500 supercomputers*. CoinDesk. https://www.coindesk.com/bitcoin-network-out-muscles-top-500-supercomputers/

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2017). Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *ArXiv:1712.01815 [Cs]*. http://arxiv.org/abs/1712.01815

Smith, P. D. (2007). *Doomsday Men: The Real Dr. Strangelove and the Dream of the Superweapon*. St. Martin’s Press;

Solca, B. (2018). Nvidia DGX-2 is the “world\’s largest GPU.” *Notebookcheck*. https://www.notebookcheck.net/Nvidia-DGX-2-is-the-world-s-largest-GPU.292930.0.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*. http://kajsotala.fi/assets/2017/11/Disjunctive-scenarios.pdf

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

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

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

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

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

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

Turchin, A. (2018). *A Meta-Doomsday Argument: Uncertainty About the Validity of the Probabilistic Prediction of the End of the World*. https://philpapers.org/rec/TURAMA-4

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 Communication Scale. *Futures*, *102*, 27–38.

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*.

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

Vinge, V. (1993). *Technological singularity*. 30–31.

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*. https://arstechnica.com/gadgets/2017/05/nvidia-tesla-v100-gpu-details/

Wellerstein, A. (2012). In Search of a Bigger Boom. *Restricted Data: The Nuclear Secrecy Blog*. http://blog.nuclearsecrecy.com/2012/09/12/in-search-of-a-bigger-boom/

Yudkowsky, E. (2008). *Artificial Intelligence as a Positive and Negative Factor in Global Risk, in Global Catastrophic Risks* (N. B. and M.M. Cirkovic, Ed.). Oxford University Press: Oxford, UK.

Yudkowsky, E. (2017). There’s No Fire Alarm for Artificial General Intelligence. *Machine Intelligence Research Institute*. https://intelligence.org/2017/10/13/fire-alarm/

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

1. “There’s no consensus among MIRI researchers on how long timelines are, and our aggregated estimate puts medium-to-high probability on scenarios in which the research community hasn’t developed AGI by, e.g., 2035. On average, however, research staff now assign moderately higher probability to AGI’s being developed before 2035 than we did a year or two ago.” [↑](#footnote-ref-1)