Article

SOLVING THE AI RACE

Addressing the potential pitfalls of competition towards Artificial General Intelligence

Dipl. Phys. Thomas Schmidt

Correspondence: tssch@gmx.de; Tel.: +49-151-42348382

Abstract: AGI could arise within the next decades, promising a decisive strategic advantage. This paper discusses risks, associated with development of AGI: destabilizing effects on strategic balance, underestimating risks in the interest of victory in the race, egoistically exploiting the huge benefits by a tiny minority. Further: Developed AGI could be beyond human control. Human goals could not be implemented and an intelligence explosion to super intelligence could take place leading to a total loss of control and power. If competition for AGI is non-transparent, secret, uncontrolled and not regulated, it's possible that risks could not be managed and would lead to catastrophic consequences. The danger corresponds to that of nuclear weapons. It is crucial that the key actors of a possible AI Race agree at an early stage on the prevention and transparent regulation of a possible AI Race - similar to measures to secure strategic stability, on arms control measures, disarmament, and prevention of the proliferation of nuclear weapons. The realization that an uncontrolled AI race can lead to extinction of humanity - this time even independent of human will – requires analogous measures to contain, prevent, regulate and secure an AI race within the framework of AGI development.

Keywords: General Artificial Intelligence, Superintelligence, Mitigating Risks, Nuclear Weapons, Strategic Stability

1. Introduction

Physicist Stephen Hawking during a talk at the Web Summit technology conference in Lissabon 2016: *"Computers can, in theory, emulate human intelligence, and exceed it. … Success in creating effective AI, could be the biggest event in the history of our civilization. Or the worst. We just don't know. … Unless we learn how to prepare for, and avoid, the potential risks, AI could be the worst event in the history of our civilization. It brings dangers, like powerful autonomous weapons, or new ways for the few to oppress the many. It could bring great disruption to our economy. … I am an optimist and I believe that we can create AI for the good of the world. That it can work in harmony with us. We simply*

need to be aware of the dangers, identify them, employ the best possible practice and management, and prepare for its consequences well in advance." [1]

The history of evolution and humanity offers four lessons that I regard as fundamental for the consideration of possible risks on the path to transformative AI or Artificial General Intelligence:

1. **Intelligence is the source of power and prosperity:** The history of evolution and the history of human development provide the undoubted, clear experimental proof that intelligence leads to a lasting, unassailable strategic advantage and thus to dominance, power and, consequently, economic strength and prosperity. This may not always apply on an individual level; this is certainly true for human groups, ethnic groups and nations, as well as for corporations and organisations.

2. **The Dual-Use-Problem - after all, it's all about political decisions:** In principle, the tools and technologies developed by people have always been dualistic: they can be used both for the benefit of people and for harm. Every invention produced by the human spirit possesses this dualism. Even the first fist wedge of a man of the Stone Age could be an enormous help in the procurement of food, but could also serve to eliminate an unwanted competitor. Whether a technology is used to benefit or harm is never a question of technology per se, but always a political decision, i.e. a question of interests and values and the ways and means of identifying and forming these interests and values of participating and/or affected stakeholders.

3. **Correlation of technology progress and level of risks:** With a few exceptions, each technological revolution increased the tension between the potential benefits and potential dangers of a particular technology. While some of the risks and harmful effects of the first industrial revolution are only partly visible today, e.g. in the form of imminent climate change, nuclear technology has been under the signet right from the start in order to create such a powerful tool - or even a weapon - whose harmful effects can no longer be limited locally and temporarily in the worst case, but whose use under predictable circumstances could lead to the complete extinction of the entire human race. With the development of AGI, a technology could arise whose risks are no longer controllable, because the mere existence of this technology can lead to catastrophic consequences whose occurrence can no longer be controlled by human will.

4. **"The winning move is not to play!" - or the life-supporting effect of strategic stability:** Because of the catastrophic, existential consequences of the use of nuclear weapons, the major powers have paid a great deal of attention to achieving and maintaining strategic stability. Cornerstone is the threat of

> mutual assured destruction, accompanied by measures to curb second strike defence, to limit and reduce offensive resources and an effective system to prevent the proliferation of nuclear weapons. So far, anyway. Securing strategic stability has been the central prerequisite over the last seven decades for preventing the catastrophic consequences of the use of nuclear weapons. Every attempt to shake this stability led to dangerous situations in which humanity slipped past destruction by a hair's breadth.

These four insights are fundamental to understanding the risks associated with the development of artificial general intelligence and will help to form the cornerstones of a strategy to minimize and mitigate such risks.

Humanity is now at the beginning of a new, the 4th, technological revolution - **the development of artificial intelligence.**

However, the subjects of this paper are not the applications of the so-called "Narrow AI", which are already becoming increasingly visible in people's everyday lives. This is AI, which is called "Siri" or "Alexa", which performs services in Elon Musk's Teslas as an autopilot or determines credit scores as a big-data-algorithms. John McCarthy, who died 2011 at the age of 84, and who was one of the true giants of computer science, said "As soon as it works, no one calls it AI any more." [2]

These specialized AI applications are already having a dual effect in business, science and society and will have a dramatic impact on virtually all areas of human life in the future.

Although the risks associated with this development for the labour market, democracy, growing inequality, the legal system, ethical value creation and many other areas should not be underestimated and therefore also require responsible, forward-looking political regulation, this paper considers these problems to be fundamentally solvable and will not discuss them further.

The subject of this working paper is the development of Artificial Generell Intelligence - AGI, the consideration of the possible development paths, the stakeholders involved and their interests, and in particular the analysis of the possible pitfalls and catastrophic risks that might occur - both on the pathway to AGI, and at the moment of the emergence of an AGI system and subsequent developments.

The presumed benefits, the huge strategic advantages that AGI systems promise, create an unprecedented attraction for various actors to make this advantage and benefit available, to acquire it, to exploit it for their own benefit and profit and to use it to the detriment and dominance of their opponents. This is one reason to fear an upcoming AI race.

The other reason lies in the capabilities of AGI systems themselves: The risk that an AGI system could far exceed human cognitive abilities and then fail human control over such a system is greater than zero.

On this basis, the working paper tries to describe a possible strategy and propose concrete measures for its implementation in order to avoid and contain such risks well in advance.

2. AGI and Super Intelligence: definitions, pathways and framework

2.1. Definition of AGI

By AGI I understand the ability of a technical system to independently form a general comprehensive model of the world and the actors acting in it, to pursue goals or sub-goals given by human supervisors or selected by itself on the basis of this model, to develop strategies for achieving these goals, to identify and analyze problems that arise, and to find autonomous solutions for them, in order to then implement concrete actions and to permanently optimize and adapt the model and the strategies to changes through learning.

Simply put: AGI is defined as a machine, capable of general cognitive performance - at first almost as good as humans, then as good and at some point better, faster and more error-free than humans.

In principle, there is no physical law that would prohibit such a system. AGI is compatible with the laws of nature. It is therefore expected that such a technical system will be developed within a short time (from a historical perspective) that has similar or even better cognitive abilities than the human brain.

At present, it is not yet clear whether at all, and if so when, the development will reach this stage. However, there are a number of significant indications that there is a high probability that at least one such system will emerge eventually.

A majority of AI researchers - 42% of AI experts surveyed - expect AGI to develop by 2030, 67% by 2050, within the next 30 years; only 10% see AGI in a time horizon up to 2100 and only a minority of 2% of all AI experts surveyed believe that AGI will never arise. [3]

The expected cognitive abilities of a technical AGI system are probably very high due to at least theoretically unlimited scalability with regard to data volumes to be processed and processing speed. It is therefore very likely that AGI is a technology that is an order of magnitude more powerful tool than anything that has been invented and developed in the course of human history. It is important to understand, that **AGI's capabilities will give those who control the AGI system a decisive strategic advantage over all others.**

If the outstanding capabilities of AGI are applied to the problem of developing improved AGI itself, the logical consequence is a recursive self-improvement of the AGI system, thus an intelligence explosion. AGI or HLAI (= Human Level AI) is not yet the possible or intended endpoint of development.

The self recursive optimization of the AGI system leads with high probability - with exponentially increasing speed - to Super Intelligence.

Philosopher Nick Bostrom mentioned this process "Take Off" in his remarkable book. [4]

Take-Off - this is a process that can lead extremely quickly (in minutes or hours) or slowly (for years) to super intelligence - is likely to lead unstoppably.

The following analogy can plausibly illustrate the problem we face here: on an imaginary "intelligence staircase" stands a chimpanzee two steps below the human; two steps below the chimpanzee a chicken, four steps below the chicken an ant. What an intellectual gap exists between the chimpanzee and us is conceivable even in the absence of concrete experience with chimpanzees. An AGI, which after a few steps of self recursive improvement stands two steps above the human being, is already beyond our imagination. What cognitive performance it performs, how it does it and why is then no longer experiencable, recognizable and controllable for people. Superintelligence, however, as it arises as the result of an explosion of intelligence, is still far, far removed from the human level on a stair level outside the image. [5]

The consequences of this development to super intelligence will be very likely dramatic.

For the first time in human history, the human species is creating a technology that, under certain circumstances, has the ability to make and implement decisions that affect the existence of humanity on a global scale intentionally, deliberately and completely autonomously - that is, without any human control - and, in the worst case, can lead to its total extinction.

In the "Risks" section of this working paper, I will go into more detail on the problem of "superintelligence".

2.2. Pathways to AGI

The development of Artificial Intelligence is still in its infancy, despite all the spectacular advances of recent years that have significantly influenced the public image of AI, such as Siri and Alexa, self-propelled cars, Alpha Go or, most recently, Google's phone AI Duplex.

However, the development of AGI by the groups/projects known so far is based on this progress. Improved software modules, new experiences and economic successes of specialized AI applications are both a platform for AGI development and a framework under which AGI development takes place.

The hype surrounding AI applications, which are already attracting investors, research resources and talent as digital business models, will also stimulate the intensity and pace of AGI development.

In December 2017, the Global Catastrophic Risk Institute CGRI published a study detailing 45 AGI R&D projects dedicated to developing an AGI, including big players such as Google's Deep Mind, Open AI or The Human Brain Project. [6]

The R&D projects mentioned in this study for the development of an AGI system are those that have become publicly known through the publication of information on the projects. However, it cannot be completely ruled out that there are other projects that are kept secret.

In the sense of the definitions above, AGI must at least reach the "computing power" of the human brain. Ray Kurzweil has estimated a necessary size of about 10^{16}, or 10 trillion cps (calculations per second), as a target value for the computing capacity of a computer that is supposed to reach the performance of a human brain.

The fastest supercomputer in the world today, the Chinese Tianhe-2, has even surpassed this number and reached around 34 trillion cps. However, Tianhe-2 is a huge system that occupies an area of 720 square meters and consumes 24 MW of power (in comparison the human brain needs 20 W).

Moore's law, so far very reliable in predicting, says that the performance of the hardware doubles every 2 years (some calculations even determine only 18 months). Today you get a computing power of about 10 trillion cps for about 1000 USD. As a result, the hardware base for AGI will very likely reach about 10 trillion cps for 1000 USD in a few years according to Moore's law.

However, hardware alone does not make a generally intelligent machine.

The most difficult part of AGI's development is the creation of the appropriate software. In the AGI projects, 3 strategies are currently being pursued:

1. Re-engineering of the human brain: On the one hand, this involves the reproduction of neural networks in connection with the learning of "correct" correlations from input to output. This strategy is already being successfully pursued for numerous narrow AI applications. More radical and ambitious is the approach to rebuild the entire human brain itself by examining actual human brain structures and then technically emulating them on powerful hardware as software. So far, such a brain could be emulated from 302 neurons. A human brain consists of approximately 86 billion neurons connected by billions of synapses. Whether AGI can actually be achieved in this way is currently not certain, but also not excluded.

2. Emulation of Evolution: Here genetic algorithms are used, where groups of computers perform tasks, where the successful programs are half merged with successful programs of other computers, while the unsuccessful program codes are eliminated. Similar to biological evolution, this artificial evolution leads to ever better algorithms over billions of iterations. Advantage over biological evolution: Computing power, processing speed and parallelism of a theoretically unlimited number of iterations reduce the time required by several orders of magnitude.

3. AI creates AI: This means building a computer whose main skills are the exploration of artificial intelligence and the improvement of its own architecture and program code. This concept builds on developments and progress at Narrow AI and is probably the most likely and fastest path to AGI development.

2.3. Conditions under which AGI is developed

Which of the strategies mentioned - or not yet invented - will ever lead to the successful development of an AGI system: This process does not simply take place in a controlled, clean environment of a scientific laboratory whose AI researchers are above all human weaknesses and compete only according to pure teaching for the first and best AGI.

It is very important to understand that any AGI development takes place under the real economic, political and social conditions of this real world today.

Two central conditions are of main importance here, which have a considerable influence on development and help to decide whether AGI development leads to controlled, regulated and above all safe AI development for the benefit of all humanity, or whether an uncontrolled AI race to achieve a decisive strategic advantage between economic and political major powers arises, whereby even careful and forward-looking attention to AI safety in favour of profit could tend to be underestimated and which could undermine the foundations of security and stability in the world by its very existence.

First condition: Globalization.

Globalization is one of the megatrends that are among the central challenges of the 21st century. Due to the internationalization of markets and companies, emerging and developing countries are increasingly participating in world trade. Huge transnational corporations are emerging, which are becoming major economic and political powers due to their financial strength and global purchasing and market power. In particular in the area of high-tech companies, scientific and technical innovations are being driven forward. Products of the digital communication world and digital business models in the area of services enlarge the gradient to monopolistic structures or function permanently only as a monopoly.

The Internet as a global medium promotes global communication in virtual space. The IoT Internet of Things - a sub-trend that is becoming increasingly important - ultimately leads to the total networking of all technical mobile and immobile devices with each other and with central elements of the infrastructure.

The conditional framework in which AGI development takes place also includes cybercrime, i.e. large-scale criminal activities committed through the exploitation of global communication, such as digital industrial espionage, identity abuse, copyright infringements, digital counterfeiting and fraud, but also cyberattacks by criminal hackers, activists, states and intelligence services, as well as corresponding defence and surveillance measures.

Second condition: Strategic Stability and multipolar world order.

Technological research and development, in particular the development of AGI, does not take place in a vacuum, but in a world that, despite the megatrend of globalization, continues to be shaped by geostrategically, economically and in part also ideologically-religiously rival actors.

The discussion of possible pitfalls and the assessment of possible risks in the context of the development of AGIs must recognize geopolitical realities, the prevailing actors and the changes taking place as serious framework conditions and include them in the formation of strategies.

Two aspects of geopolitical realities are of crucial importance here:

1. Strategic Stability: The strategic stability between the major nuclear powers of the USA, Russia and China, which has existed for about 70 years and is essentially based on the existence of nuclear weapons and deterrence through mutual assured destruction.

2. Multipolar World Order: The temporal coincidence of rapidly increasing, exponentially growing technological capabilities with the emergence of a new multipolar world order in which the USA is trying to defend by all means the global dominance, briefly achieved after the collapse of the USSR and the Eastern Bloc, while this dominance is increasingly called into question not only by Russia and China.

What does this have to do with the assessment of risks in the development of AGI?

Most publications from the field of AI research on AI/AGI safety deal primarily with the question of how an omnipotent AGI or super intelligence can be controlled and its effect on human goals committed.

There is also broad consensus in the academic community and in leading high-tech companies that the use of AI in weapons technology for the creation of lethal autonomous weapons (LAWS) can lead to dangerous consequences for security and freedom.[7] That is why a ban on autonomous weapons systems is rightly being called for from there, but also increasingly from politicians and some military leads.

Military publications and security policy think tanks are increasingly discussing how AI will influence and change future warfare, mostly addressing the impact of AI and automation on weapons technology, intelligence and decision-making mechanisms, with emerging new technology powers such as China being recognized as a threat by US experts.[8]

A recent report by RAND Corporation has investigated the effects of AI on existing strategic stability and identified a possible threat to stability as probable.[9]

So far, however, in my opinion, not enough attention has been paid to the question of whether and if so what effects a race of different actors - state and non-state actors - has on the creation of AGIs for security, stability and so-called AI safety per se. However, this is a central point in the evaluation of risks and possible pitfalls in the development of AGI systems.

Therefore, the following section first takes a closer look at the main actors in the AI Race, examining their interests and goals as well as their means and skills in the development of AGI.

3. Actors and stakeholders in regards to an AI race and their interests and capabilities

In order to investigate the risks associated with the development of AGI and to develop strategies and measures to minimize such risks, it is necessary to identify and evaluate the key players in the possible AI race.

The starting point for this is the above-mentioned study by the Global Risk Institute of December 2017.[10]

Author of the study Seth D. Baum describes 45 projects that have set themselves the goal of developing AGI.

These projects from 30 countries on six continents are mainly located in major corporations and academic institutes, some of which are large and massively financed. Many of AGI's development projects are linked through joint personnel, joint parent companies or through cooperations.

- Most of the projects are located in companies and scientific institutes.
- Predominantly - at least until now - open source code is published.
- Some projects have military links and/or are financed by the military.
- The most frequent AGI projects exist in the USA; almost all AGI projects come from the USA or close allies of the USA.
- The only projects that are not currently based in the US or with its allies come from Russia and China, all with strong academic or Western ties.
- Most projects claim that their goals are oriented towards the benefit of humanity as a whole or that they simply want to expand the boundaries of knowledge.

- The vast majority of R&D projects do not actively address AI safety problems in their research.

- Almost all currently active - known - projects are small to medium-sized projects.

- The three largest AGI projects are:
  - Deep Mind (a London based project of Google)
  - The Human Brain Project (an academic project based in Lausanne, Switzerland)
  - OpenAI (a non-profit project in San Francisco (co-founded and supported by Elon Musk)

The investigation of Seth D. Baum identifies four overarching trends (regardless of size, location, financing):

1. A **group of academic AGI projects**, according to their own statements, pursue the scientific goal of expanding knowledge and findings. These projects are not **actively involved in AI Safety**.

2. Another group of **companies' projects** have set themselves the goal, according to their own publications, of creating AGIs for the benefit of humanity. These projects are also **actively involved in safety aspects** in their research and development work.

3. The group of projects with military links are academic US projects that are wholly or partly funded from military sources.

4. All six Chinese projects are relatively small, but some are part of large companies/organisations and therefore have resources to scale quickly when needed.

5. All of the AGI R&D projects identified and investigated have provided more or less intensive and detailed information about their goals and plans (publications, websites, conferences).

6. However, it cannot be ruled out that there could be other AGI projects that are kept secret, although the probability is currently rather low simply because of the limited HR in the field of AI research and the attraction of directly economically interesting narrow AI projects, with which money and fame can be earned today or in the near future.

For the risk assessment, however, the question arises whether this will remain if - as outlined above - AGI development makes further tangible progress and the necessary hardware power becomes possible at economically acceptable costs.

In particular, other powerful players can then be expected to become participants in a massive AI race, unless this coming AI race is prevented or at least contained and regulated beforehand.

The closer the purely scientific-technical progress of the existing AGI projects comes to the stage of an actually possible developed AGI, the more likely it is that the expected superhuman abilities and thus the hoped-for strategic advantages will become known to other actors.

The will to obtain, secure and exploit the presumed strategic advantages of a developed AGI will be the decisive motive for taking up the race for the first AGI and activate such actors as listed below.

It is irrelevant whether the presumed strategic advantage actually occurs and could actually be monopolized. Only a low but realistic probability, i.e. the assumption of such a strategic advantage, which would be synonymous with an enormous increase in profit and/or power, will bring predictably powerful actors into play:

1. The major global high-tech companies and Internet platform companies (such as Apple, IBM, Google, Facebook, Microsoft, Huawei - to name but a few). These big players have virtually unlimited financial resources and a dominant market power, including the ability to influence public opinion as they see fit. They already have huge resources of know-how and research capacity.

2. The major powers and associations of states, in particular all permanent UNSC members (USA, Russian Federation, China, France, UK), Canada, Australia, the EU, India, Brazil and other emerging countries and regional powers. These actors will most likely attempt to intervene in an AI race that is about to begin, through government funds and organizations such as the military, intelligence services, government agencies, and government or state-funded research institutions or private public partnerships. The importance of AI development (including AGI development) especially for the major powers and for their objective interest in strategic stability and geostrategic rivalry can be seen in the statement of Russian President Putin in September 2017: "Artificial intelligence is the future, not only for Russia, but for all humankind. It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world."[11] In July 2016, China unveiled a plan to become the world leader in artificial intelligence and create an industry worth $150 billion

to its economy by 2030[12] and End of 2016 the US Administration issued "The National Artificial Intelligence Research and Development Strategic Plan".[13] It is therefore very likely that sooner or later the great powers will actively intervene in the AI Race if they do not already do so.

3. Large financial investors and venture capital firms in connection with innovative start-ups, which could try to bypass the large established private-business or state AGI projects as late entrants and side entrants in open or secret projects and thus only enjoy the strategic advantage of AGI shortly before the end of the AI Race.

4. Finally, this cannot be ruled out: The inclusion of AGI Research or the illegal appropriation and exploitation of the results of such research by organised crime and terror organisations. Like all legal actors, attempts to enter and win an AI race are likely to be made from this direction because of the incredible incentives for power gain and profit.

All of the above stakeholders and actors have the resources and power tools needed to successfully enter an AI race.

All share the extremely strong interest in securing and exploiting the supposed strategic advantages of an AGI system for themselves. (Apart from individual interests in glory and fame, which goes hand in hand with scientific success and engineering breakthroughs, or individual pathological "interests" in destructions.)

Likewise, all these actors have a common interest in each case in preventing the potential rival(s) from securing the strategic advantages of AGI and monopolising it.

On the other hand, however, all actors share the objective common goal of not being defeated in the race for AGI, as this may entail a total loss of power and prosperity, or at least a loss of dominance.

Last but not least, all actors share the goal of being able to control a developed AGI system under all circumstances and to ensure that a superintelligence can also be fixed on human goals and values.

This raises a very crucial question: Are these goals realistically achievable, or do steps to achieve these goals lead to unacceptable risks?

If so, are such risks to be contained and is it then possible to achieve the above objectives anyway?

Or does an unbiased risk analysis show that AGI development must lead to completely new strategic approaches if the existence of the entire human race is not to be put at risk?

I try to answer these questions in the following two sections.

4. Risks associated with AGI and Super-Intelligence and with a possible AI race towards AGI

The consideration of risks and potential pitfalls in connection with the development and creation of AGI is based on the following assumptions, which have been explained in the previous sections and are summarized here once again:

1. AGI as defined above will be developed within a manageable period of less than 80 years, probably even within the next 30 years.

2. This AGI will have cognitive abilities that will exceed human performance many times over. It follows: Whoever controls AGI - if AGI can be controlled at all - controls superhuman intellectual achievements and abilities and is thus in possession of serious "decisively economic and political power generating, i.e. decisive strategic advantages.

3. Scenarios in which AGI will undergo a recursive improvement and optimization process that creates an intelligence explosion to machine superintelligence are highly likely.

4. The prospect of decisive strategic advantage for actors and stakeholders in AGI development represents an extremely strong incentive to be the first to develop, control, own and use AGI. This "The Winner takes it all" scenario is the trigger for an AI race.

5. The development of AGI is taking place under the conditions of globalization and the emerging multipolar world order. If an AI race occurs under these conditions, its course and results can have devastating effects on the strategic stability of the multipolar world order, as well as further increase or even prevent the already risky, still completely unsolved problems of AGI's control.

These assumptions form the basis for the evaluation of possible risks and pitfalls that could occur during the development and use of AGI.

I differentiate between three main risk groups:

1. Risks related to the progressive development of AGI and a possible AI Race

2. Risks in connection with the application and use of AGI

3. Risks in the area of safety and ensuring human control and supervision of AGI

4.1 Risks during the race to AGI

An uncontrolled, unrestrained race between key stakeholders to be the first to secure the strategic advantages of AGI for themselves can have destabilizing effects on the strategic balance:

- The attempt to gain strategic advantages through AGI inevitably leads to corresponding countermeasures in advance.
- AGI might provoke a new arms race also in regards to autonomous weapon systems
- AGI might challenge the basic rules of nuclear deterrence.
- Success in AGI development would lead to an strategic advantage of an opponent and would trigger counter measurements, might lower the nuclear threshold and would increase the probability to intent a preemptive strike. A recently published study by RAND Corporation has pointed out that the time from today to a fully developed and safe AI is particularly dangerous, because only then - if at all - risk minimizing effects of AI could arise, because better, faster, more reliable information and decision-making aids would be available through AI.[14]

In addition, an AI Race could result in not paying enough attention to the possible risks to take a higher risk, either recklessly or deliberately. Key stakeholders may ignore or underestimate safety procedures in favour of faster utilisation.

The interested public of companies, investors and researchers / engineers who develop so-called Narrow AI and are looking for economic success with AI applications does not focus on the topic of AI safety and risks of AGI. Even in the 45 projects dedicated to AGI research to date, safety, ethics and policies are either not on the agenda at all or are only evaluated as a second- or third-tier problem and dealt with accordingly.

4.2 Risks appearing once AGI exists

As soon as AGI as defined above exists, two new key risks arise, whose dangerous effects must be prevented without ifs and buts in advance.

- The "The Winner Takes it all" scenario

- The "control problem" - with AGI (before the intelligence explosion) and superintelligence

The first risk is that developed AGIs actually offer a decisive strategic advantage to those who are the legal or factual owner, disposer or controller, and who then try to monopolize that strategic advantage and turn it into power gain for themselves and their allies (which also always means profit growth).

Thus a possible great benefit of AGI is not made accessible to the whole of humanity, but is egoistically exploited by a tiny minority.

A scenario that would lead to measures being taken in the run-up to a successful AGI development to secure the actual or suspected strategic advantage, which in turn would change strategic stability, but would also attack and undermine the social fabric and democratic structures. Both would have very dramatic consequences for the security and functionality of civilization.

The second risk is that possibly developed AGIs may escape human control and it proves impossible to firmly implement top human goals.

With the emergence of developed AGIs at human level and before a possible and probable intelligence explosion or in the event that a superintelligence could actually be controlled as part of a solution to the control problem, which is highly doubtful, the following risks are very likely to arise:

- The control problem comes to light openly as AGI finds and implements solutions to problems in pursuit of given goals and completion of set tasks which can no longer be influenced by people because they are neither predictable nor comprehensible.

- It is also possible that an AGI system, as part of its problem solving strategy for a particular task, actively prevents any influence because it has recognized this as the most effective form of problem solving.

- However, this can actually also be concealed if an AGI system identifies human interaction itself as a sub-problem in solving a given problem and chooses hiding its actual intentions as an effective strategy to prevent this interaction.

- A major risk is that human goals cannot be implemented. First, because lack of effective structures, lack of knowledge base, underdeveloped democratic cultures, it is not or not timely possible to agree within the human race on such universal, universal mega-goals, which would be necessary to make an AGI act safely and permanently for the good and benefit of humanity and in accordance with such universal values.

- Secondly, because it is not possible, or not at all possible, to find technical-functional solutions in the development of AGI, with which every action and action of AGI can be permanently and irrevocably committed to these universal principles and values. It is about "burning" human values and mega-goals into the AGI "DNA" - as far as it would be possible to define such goals and values universally and to legitimize them in a global democratic process.

- A further risk is that no means exist, or if they could at least theoretically exist, cannot be found with which an undesirable action of an AGI can be sanctioned and, if necessary, changed.

However, if there is an explosion of intelligence towards superintelligence, there is a risk of total loss of control and power for humans.

Super Intelligence - It is imperative to initiate further research on take-off and superintelligence. Miscalculations might lead to catastrophic consequences. A Superintelligence could intentionally (self defence) or accidentially wipes the humanity.

4.3 Comparative consideration of the risks of the outbreak of a nuclear was against risks associated with AGI and super-intelligence.

An AGI on a human level which has been given the task - or which, in the context of solving another problem, has set itself the task of recursively becoming active for its own improvement and optimisation - will not regard the Human Level as an important milestone and then stop with the optimisation.

AGI will then perform a fast (minutes or hours) or slow (years or decades) intelligence improvement that will go far beyond human intelligence.

In our world "smart" means an IQ of about 130, while "stupid" is perhaps well described with an IQ of 80.

But what if we're dealing with intelligence with an IQ of 10,000?

An intelligence explosion occurs with exponential growth and taking advantage of speed, storage capacity, connectivity and extremely powerful hardware. It may only take 20 minutes to achieve a significant increase in intelligence performance.

We must recognise that it is quite possible that shortly after the great news about the first machine to reach AGI on a human level, humanity will be confronted with a machine intelligence that is a million times more intelligent than human beings.

With the possible development of AGI into superintelligence, humanity is facing an incalculable existential risk whose global effects will be as profound and disastrous in the worst case as a full scale nuclear.

None of the risks described above need therefore become a real danger or cause actual irreparable damage.

Just as the mere existence of thousands of nuclear weapons in the arsenal of the two nuclear superpowers, the USA and Russia, has not yet led to a devastating use of nuclear weapons, which in all probability could not be limited and limited, but would inevitably end in full-scale nuclear war.

The dangers of an AI Race are similar. On the one hand, because of possible **direct effects on strategic stability**, although there is complete agreement among experts that the risk of nuclear war increases with every shift in the balance. On the other hand, because **failure to solve the so-called control problem and loss of control over AGI** and/or an intentional or accidentally induced intelligence explosion, in the course of which a machine, artificial superintelligence develops, can lead to catastrophic consequences for the whole of humanity.

From all these considerations, there is only one reasonable conclusion:

The risks of an AI race must first be taken and treated similarly to the risks of an outbreak of nuclear war.

5. A strategy to avoid an AI race and to mitigating risks associated with an AI Race and AGI/SI

The risks associated with the development, creation, use and control of AGI and a machine, artificial superintelligence that is likely to arise in succession are, as we have just seen, comparable to the somewhat fortunate risks of a nuclear-armed conflict between the major powers of the USA, Russia and China, which have existed for more than 70 years.

If damage occurs due to these risks, their effects are global and affect civilization on this planet per se.

The strategy to avoid an AI race and to mitigate the corresponding risks should therefore comply with the following basic principles:

1. **International:** Measures to avoid an AI race and to mitigate risks related to AGI and possible super intelligence require an international framework analogous to existing arms control measures.

2. **Constraint:** Such measures must be very restrictive and effective in the face of existential threats and dangers and must have the highest priority, be enshrined in international law and national laws and guaranteed by a strong mandate from the UNSC.

3. **Transparent and democratically legitimized:** Despite and even because of the necessary restrictive measures, these must be prepared, decided and implemented very transparently on a global scale. Ensuring a benefit for all of humanity, preventing the primordial surpation of AGI / SI and the development of a canon of values that is obligatory for AGI require a global democratic legitimation, independent of property and wealth, origin, culture, religion, race, age or gender.

4. **Committed to the highest human values:** In particular, the debate about and the development of a universal human canon of values, which must be burnt into every AGI or into the fundamental "DNA" - if this is possible at all - must be oriented towards the highest and best human values.

Risks in an AI race and risks of AGI itself create a global problem with potential global damage and existential threats to the entire human race. Therefore, the solution to the problem must also be global and international.

The key stakeholders - in particular the permanent members represented in the UNSC, but also the globally operating, most important and largest high-tech companies - are called upon to initiate, actively operate and secure an international process in analogy to existing arms control and disarmament agreements and to the functioning regime for the prevention of the proliferation of nuclear weapons with appropriate guarantees.

6. Recommendations

Possible solutions / measures to reduce the risks of an AI race or to mitigate the risks associated with an AI race:

1. Creation of an international agreement among the most important state actors (e.g. within the UNSC) with the involvement of the Big Tech Companies = a so-called AI RACE CONVENTION, e.g. within the framework of the Geneva Disarmament Conference. Such a convention was proposed in regards to cyber security by Brad Smith, President of Microsoft.[15] He called it Digital Geneva Convention - such principles and frameworks might also helpful in regards to Solving an AI Race.

2. With this agreement, the key stakeholders agree on the following principles:
    a. Renouncement of an uncontrolled AI Race
    b. Commitment to the general principles and regulations for AI Safety
    c. Enforcement of these rules in national and international law
    d. Agreement and enforcement of mitigation measures against all non-signatories to the Convention

3. Stigmatization and restrictive prosecution of any violation of the rules of the Convention by states, companies and individuals.

4. Establishment of an IAEA-like organization that globally monitors the proliferation of AGI capabilities, compliance with the rules of the convention, transparency obligations, etc., identifies and locates violations and proposes sanctions, a International AI Control Agency IAICA

5. Establishment of a Taskforce, which is equipped with skills and competences to prevent an AGI project from getting out of control in an emergency or to switch off and/or contain an out-of-control AGI/SI and/or to make an argumentative commitment to the observance of human goals and interests. /violence/technical/negotiation solutions

6. Creation of something like a Guardian AI under international control by IAICA, which searches for undeclared AGI projects based on Big Data monitoring and pattern recognition, identifies, locates and infiltrates them.

7. Establishment of a policy document similar to the Hippocratic Oath - a kind of AGI research ethic to which all AI researchers subscribe and submit, and which threatens global stigma and economic/social ostracism if violated.

8. Creation of international regulations, in particular appropriate protection for whistle blowers from the fields of AI/AGI research.

9. AGI research and development must be organised under international control and similarly as is the case with nuclear installations that are monitored and controlled internationally by the IAEA. AGI research should take place in a strict regulatory corset, similar to research/development/production of sensitive military goods and technologies.

- These measures are designed to ensure that
- several stakeholders do not start an AI race
- potential rivals in other fields (military, economic, ideological) do not compete against each other
- no researching AGI will be done undercover
- several AGI projects do not reach the "Ready for Take Off" status simultaneously
- not a "winner takes it all" scenario can take effect

In case of doubt, the restrictions must extend to such an extent that the development of AGI is completely stopped or has to be interrupted until questions of the control problem, the goal orientation - in particular the global understanding of human values and goals - are completed.

References

[1] CBNC Special Report, Stephen Hawking at Web Summit Technology Conference, Lissabon, Portugal 06/11/2016

[2] https://cacm.acm.org/blogs/blog-cacm/138907-john-mccarthy/fulltext

[3] Study, conducted recently by author James Barrat at Ben Goertzel's annual AGI Conference, in Our Final Invention: Artificial Intelligence and the End of the Human Era, St. Martin's Press, 2013

[4] Nick Bostrom, Superintelligence: Paths, Dangers, Strategies, Oxford 2014

[5] https://medium.com/ai-revolution/ai-revolution-101-8dce1d9cb62d

[6] Seth D. Baum, A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy, GCRI 2017

[7] Open Letter of Ai Researchers and Company Executives, https://futureoflife.org/autonomous-weapons-open-letter-2017

[8] Elisa B. Kania, Battlefield Singularity, Artificial Intelligence, Military Revolution, and China's Future Military Power, CNAS, Nov 2017

[9] E. Geist, A. J. Lohn, How might AI affect the risk of nuklear war, RANDCorp., April 2018

[10] Seth D. Baum, A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy, GCRI 2017

[11] Vladimir Putin during an open lesson to students on 1st of September 2017, https://www.rt.com/news/401731-ai-rule-world-putin/

[12] https://www.technologyreview.com/s/610546/china-wants-to-shape-the-global-future-of-artificial-intelligence/

[13] https://www.nitrd.gov/news/national_ai_rd_strategic_plan.aspx

[14] https://www.rand.org/blog/articles/2018/04/how-artificial-intelligence-could-increase-the-risk.html

[15] https://youtu.be/EMG4ZukkClw