AI: Will the development of superintelligence be humanity's last mistake?
November 17, 2024 - Andreas von Westphalen

Translation with www.DeepL.com/Translator from:

https://www.telepolis.de/features/KI-Wird-die-Entwicklung-von-Superintelligenz-zum-letzten-Fehler-der-Menschheit-10032002.html

Note on quotations: Some quotations are taken from German translations of English-language books, which have been translated from German into English using deepl and may therefore differ from the original.

AI can do more and more. Optimists promise a digital paradise, while renowned experts warn of existential dangers. What if we lose control? (Part 1)

Interview with Prof. Karl Hans Bläsius, Professor of Artificial Intelligence at Trier University of Applied Sciences and operator of the websites "Nuclear war by mistake" and "AI consequences".

First, let's talk briefly about the topic of the future. Ray Kurzweil, technical director at Google, warns in his book "The Singularity is Near" that people make two fundamental mistakes: Firstly, we have a linear idea of how the future will develop, and secondly, we only see one change without taking into account that it has an impact on many other aspects.

Kurzweil is convinced that the development of information technology is exponential. Do you share Kurzweil's assessment of our naive vision of the future?

Karl Hans Bläsius: The term exponential is often used carelessly. Surprisingly rapid progress would also be possible with linear improvements in technical fundamentals. What exponential growth means is particularly evident in AI when we try to develop systems that automatically solve problems.

This often involves dealing with exponentially growing search spaces, i.e. a rapidly increasing number of alternatives to be considered. AI is therefore more about the fact that the effort required grows exponentially as the problems become more difficult, not the progress made in development.

However, it happens time and again that a new methodological approach leads to sudden improvements in the search for solutions. In the case of processing by means of logic, in 1965 this was the so-called resolution based on unification, a new set of rules for logical reasoning, which led to a leap in quality for automatic problem solving on the basis of logical calculus.

Generative AI systems also led to a surprising leap in quality at the end of 2022 with the release of ChatGPT. These systems will have a significant impact in many areas, and the consequences cannot yet be accurately assessed. If, as Ray Kurzweil expects, a singularity is reached soon, there could be an explosion of intelligence with perhaps exponentially growing improvements.

However, there have always been setbacks in AI, and the high expectations of new methodological approaches have not always been fulfilled, as was the case with expert systems, for example. Further developments can therefore hardly be predicted.

A narrowed view

▶ Let's turn to the central topic of artificial intelligence. Mo Gawdat, former Chief Business Officer of Google X, explains in his book “Scary Smart”:

"To become an expert in artificial intelligence you need a specialized, narrow view of it. That specialized view of AI completely misses the existential aspects that go beyond the technology: issues of morality, ethics, emotions, compassion and a whole suite of ideas that concern philosophers, spiritual seekers, humanitarians, environmentalists and, more broadly, the common human being (that is to say, each and every one of us).”

Do you agree with him?

Karl Hans Bläsius: In conventional computer science, tasks can often be precisely specified and then converted into solutions that fulfil the tasks using a defined procedure. AI is often about automatically solving problems, whereby not all specific problems are known during programming.

Possible solutions for such problems are also not known but must be found by the system. The quality of a solution is also relevant. It is therefore not important to find a solution at all, but to find an optimal or best possible solution.

When programming, however, it is not possible to estimate what quality can be achieved with what effort. For example, systems for automatically translating a text into another language have been around for some time. Initially, however, these were hardly usable.

For some years now, systems such as deepl are available that produce translations of very good quality. Problems with unforeseeable effort for solutions in good quality often require numerous experiments.

This is why such development activities have a kind of research character, and you have to deal with them particularly intensively, requiring a specialized, narrow perspective. Other aspects, such as possible consequences for our societies, could be overlooked.

In this sense, I agree with the above statement. However, many AI developments are normal technical advancements, as they occur in many technical fields, and are not usually about existential risks.

Heaven or hell

▶ There is hardly any other topic about which there are such different basic attitudes. On the one hand, for example, software developer and investor Marc Andreesen announced a year ago:

“Fortunately, I am here to bring the good news: AI will not destroy the world, and in fact may save it.” In “The Next Stage of Evolution”, Ray Kurzweil even says that we “have a moral obligation to fulfill the promises of these new technologies.”

On the other hand, the “1 sentence statement” was published more than a year ago. It states:

“Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.”

It was signed by Demis Hassabis (Deep Mind), Sam Altman (OpenAI) and Bill Gates, among others. What was the response and reaction to this call? A few days ago, Israeli historian Yuval Harari explained in an interview with Der Spiegel:

„Artificial intelligence will radically change our lives on this planet. We must realize the enormity of this invention. If we as a species are not careful, it could end in disaster. (...) It's as if aliens had come from orbit. We wouldn't just hand over control to them if we didn't know how they make decisions. AI has the potential to enslave humanity."

What is your assessment of the weal and woe of AI?

Karl Hans Bläsius: Most applications of AI are positive, make our work easier and improve our quality of life. This applies to many areas, such as medicine. As in other technical fields, there are also dual-use aspects in AI, so there can be positive and negative applications. Autonomous driving can have many positive effects, while autonomous weapon systems are dangerous.

In my view, generative AI systems, which could pose existential risks to humanity, must be assessed differently. Systems such as ChatGPT have enormous capabilities in linguistic communication, can answer difficult questions and can do so for almost all specialist areas. Although not everything is always correct, the answers are often quite good, and this has once again spurred euphoric predictions.

Marc Andreesen's optimistic assessment may be correct. It could be that AI can save the world, because it is questionable whether humanity can do it. It is questionable whether we humans are capable of finding and implementing solutions to climate change, nuclear weapons, autonomous weapons and other dangers.

Nevertheless, we should not rely on an AI to do this for us and save us. On the contrary, advanced AI systems also pose considerable risks.

I do not share Ray Kurzweil's optimistic view. If AI development continues to make enormous progress and there are no serious negative consequences, the positive effects will probably initially only affect a few people who will benefit enormously. If it does indeed come to pass that a large part of our work will soon be done by machines, the machine owners will benefit, but what about the people who did the work before?

Warnings of significant risks from leading AI scientists and heads of major AI companies came very soon after the publication of ChatGPT. First, the Future of Life Institute called for a six-month pause in March 2023 and at the end of May 2023, the "1 sentence statement" was published with the warning that AI could lead to the extinction of humanity.

These warnings are also linked to the fact that AGI (artificial general intelligence) is expected soon, i.e. systems whose level of intelligence is comparable to ours in many areas. It is feared that such systems could then get out of control.

These warnings have been widely criticized, with the following arguments being used, among others: These systems merely calculate probabilities on the basis of which results are delivered. Others argue that the systems only execute predefined command sequences. Further arguments

are that the systems do not understand anything and have no will or similar. The systems are a long way from an AGI or superintelligence and pose no danger.

However, such arguments must be contradicted. In AI, there is a wide variety of methods that can be combined with each other, and it is not always just about probabilities. A central concern of AI is the automatic solving of problems, whereby at the time of programming it is completely unclear which problems will occur, and which solutions are possible, which the systems must find themselves.

It is therefore not appropriate to regard this as simply the execution of a sequence of commands. One can argue about when one can speak of understanding, consciousness and similar characteristics. However, the risks that can emanate from such systems are independent of whether or not one assigns properties such as understanding or consciousness to these systems.

Theoretical limits in terms of computability and decision-making also apply to AI systems, of course, but even humans cannot overcome these limits. However, there are also no known limits that could prevent an AGI or superintelligence.

Generative AI systems are active on the Internet and humans could lose control over their behavior even before an AGI is created. Harari's warnings are also justified. Mustafa Suleyman, co-founder of DeepMind, also warns in his book “The Coming Wave” of huge changes that are not necessarily all positive.

However, many authors, such as Suleyman, remain vague in their statements about possible consequences, which is mainly due to the fact that it is almost impossible to predict what the effects will be.

Stuart Russell and Peter Norvig also address such risks in their book “Artificial Intelligence - A Modern Approach” (german version of 3rd edition, page 1194) and write: “Almost any technology has the potential to do harm in the wrong hands, but for artificial intelligence and robotics we have the new problem that the wrong hands may belong to the technology itself.”

Superintelligence

► We keep hearing the term “superintelligence”, which you have just used. What is superintelligence?

Karl Hans Bläsius: Back in 1965, Irving John Good described an “ultra-intelligent machine” that can far surpass all the intellectual abilities of a human being. A machine that matches our intellectual abilities can itself develop a new machine with better abilities, and so on. This leads to an explosion of intelligence, far surpassing the intelligence of humans. The first ultra-intelligent machine would be the last invention that humans have to make.

The realization of such a system, which far exceeds the human intelligence level in almost all areas, is also referred to as superintelligence.

The achievement of human intelligence by a machine as a starting point for an intelligence explosion was described by Vernor Vinge, math professor and science fiction author, as a technological singularity. In 1993, Vinge predicted that superhuman intelligence could be created within 30 years and that the human era would end shortly afterwards.

Sudden change in behavior

▶ In your latest article you write:

Nick Bostrom fears, however, that this increasing intelligence may lead to a superintelligence and that there will be a tipping point. As long as an AI is powerless, it will behave cooperatively. As soon as it is strong enough and eventually forms a superintelligent singleton, it will change strategy without warning and optimize the world according to its own goals.

Can you please explain this in more detail?

Karl Hans Bläsius: In his book "Superintelligence", Nick Bostrum looks at many possibilities of how a superintelligence could emerge, what steps are necessary to achieve this and what the effects will be in each case. It is also important to note which assumptions and preconditions underlie individual possible steps. Bostrum assumes that these systems can achieve a certain degree of consciousness and will of their own.

Aspects such as the consciousness, feelings and will of machines are highly controversial. However, there are no known limits to such aspects. It is also not a problem in principle to realize systems that simulate something like consciousness or feelings. The only question is in what quality is this possible, i.e. how effectively can such aspects be realized.

If a good quality can be achieved, it is irrelevant for the effectiveness and possible consequences whether one speaks of these systems having such properties or whether they only simulate them. This is completely irrelevant for possible dangers.

Whether, when and how a superintelligence can emerge and what consequences this will have is completely open and incalculable. In addition to the development of a hidden superintelligence described by Bostrum, there could also be many other paths to such systems.

Possible control

▶ Everyone agrees on the need for control over artificial intelligence. But opinions differ widely as to what this looks like and how promising it can be. Mo Gawdat asks:

How do we ensure that, in addition to its intelligence, the machine also has the values and compassion to know that it is not necessary to destroy the fly we are becoming? How can we protect humanity?

Some say we should control the machines: Build firewalls, enact government regulations, lock them in a box, or limit the power supply to the machines. These are all well-intentioned, if vigorous, efforts, but anyone who knows anything about technology knows that the smartest hacker in the room will always find a way through any of these barriers. That smartest hacker will soon be a machine.

Is that so?

Karl Hans Bläsius: Yes, that is to be expected. Currently, the emergence of an AGI could come from generative AI. These systems are active on the Internet, possibly distributed across many data

centers on different continents. How are firewalls or energy restrictions supposed to work as long as there is no agreement between all nations?

Furthermore, systems such as ChatGPT are also very good at programming and can also be used for cyber attacks. It is therefore to be expected that the best hackers will soon be machines.

▶ You also write:

An AI system needs a proper utility function that gives the system a goal to achieve or optimize. Specifying such a utility function can be difficult because it is not possible to predict what conclusions an AI system can draw from given situations and specified utility functions.

Stuart Russell and Peter Norvig write in “Artificial Intelligence - A Modern Approach” (german version of 3rd edition, page 1195): “Let's hope that a robot intelligent enough to figure out how to wipe out the human race is also intelligent enough to figure out that this was not the intended utility function.”

How realistic do you think it is to set up this necessary utility function?

Karl Hans Bläsius: Most authors who deal with superintelligence also emphasize the need to provide a starting AI with appropriate values and to define meaningful utility functions. The question is what effects such given settings can have.

When an intelligence explosion occurs, a development process that would take humanity many years or several generations takes place in a matter of days or months. Machines are always developing better machines. It is questionable whether any of the originally intended utility functions and values will remain after a few cycles. The machines could create their own values in unpredictable ways with a completely open outcome.

Trying to pass on values to an initial AI that still apply after many cycles of an intelligence explosion is perhaps comparable to teaching your own children values with the aim of ensuring that their descendants will still have them in 10 or 20 generations. The prospects of achieving something like this are likely to be extremely slim.

When designing an AI system, it will be difficult or almost impossible to take into account all possible further developments in the creation of a superintelligence in such a way that the utility function always remains friendly towards humans. Such a utility function cannot be fixed statically for all time, but must be able to change over time and adapt to new needs.

Russell and Norvig explain the disadvantage of a fixed static utility function using the following example: If it had been possible to realize a superintelligence around 1800, with a static utility function it would still introduce slavery today and abolish the right to vote for women, as this corresponded to the moral concepts of the time.

Necessary regulations

▶ To what extent could state regulation, as we have long known from other areas such as medicine, provide protection against undesirable AI behavior?

Karl Hans Bläsius: Especially since the warning of May 30, 2023, which stated in one sentence that AI could lead to the destruction of humanity, there have been increasing calls for AI to be regulated. The AI scientists and development company executives who signed this "one-sentence statement" also consider effective regulation to be extremely important.

Authors of books on superintelligence had also previously called for such regulations. For example, Stuart Russell writes in his book "Human Compatible - Artificial Intelligence and the Problem of Control" that due to potential risks for humans and perhaps for humanity as a whole, regulations for software development, especially in the field of AI, are extremely important and compares this with corresponding rules in the field of medicine.

Incidentally, Stuart Russell is no longer so optimistic that humans will retain control over superintelligent machines, and he now also expects a superintelligence to emerge much sooner than depicted in his book.

▶ To what extent does the EU's planned AI regulation create the necessary regulations?

Karl Hans Bläsius: The AI Act adopted by the EU is an important measure to reduce certain risks associated with advances in AI. In particular, some especially critical applications can be made more difficult or prevented. These include determining personal characteristics, attitudes and health hypotheses from texts, photos or sound recordings.

However, generative AI systems also pose other risks that are not covered by this AI regulation. This relates in particular to the risk of possible superintelligence. Research and development in this direction are not affected by the AI Act.

▶ To what extent do we then need global AI legislation? In 2023, almost thirty nations - including China, the USA and Germany - signed the Bletchley Declaration. It critically states:

"There is potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of these models. Given the rapid and uncertain rate of change of, and in the context of the acceleration of investment in technology, we affirm that deepening our understanding of these potential risks and of actions to address them is especially urgent."

And they promise:

"Many risks arising from are inherently international in nature, and so are best addressed through international cooperation. We resolve to work together in an inclusive manner to ensure human-centric, trustworthy and responsible that is safe, and supports the good of all through existing international fora and other relevant initiatives, to promote cooperation to address the broad range of risks posed by."

How do you assess the chances of global AI legislation, given the race in the AI industry, particularly between the USA and China?

Karl Hans Bläsius: The goals described in this declaration are very important and urgent. However, I would rather fear that the current course of confrontation between the West and Russia and the

threat of confrontation with China will lead to some other countries, such as China, working at full speed on projects similar to ChatGPT.

None of the major nations wants to be left behind and lag behind the capabilities of competing powers. The result could be that competing super-intelligent systems soon emerge that act against each other and also against humanity.

Protection against the risks posed by a possible superintelligence is only possible by working together. All states must see this as a common human task. This is the only way to create effective AI regulations and reduce the risks.

There may be very little time left for this. As a superintelligence with incalculable consequences threatens in the next few years or decades, appropriate measures, such as ending wars and the current confrontational course between major nations, would be necessary immediately in order to create the conditions for effective agreements.