Trier, Saarbrücken, 18.5.2024

Generative AI - Possible consequences on the internet

Karl Hans Bläsius, Jörg Siekmann

https://www.hochschule-trier.de/informatik/blaesius/ , http://siekmann.dfki.de/de/home/

Link to this document: https://fwes.info/GenKI-Internet-2024-1-en.pdf

See also: www.ai-implications.eu

Generative AI - Possible consequences on the internet

1. Successes of ChatGPT, further developments

ChatGPT was released in November 2022. This system demonstrates enormous capabilities in linguistic communication. Extensive texts can be created based on simple requirements. Short summaries of extensive texts are also possible. This capability extends to many specialist areas. Generative AI systems can also be used to create images and videos. Most answers to questions are very convincing, even if not everything is correct. These systems can also solve programming tasks in various programming languages.

New systems with improvements have been announced. Possible improvements could concern:

- reduction of errors so that as few incorrect results as possible occur,
- combination with symbolic AI,
- reasoning skills, i.e. drawing further conclusions from normal previous results and thus improving the results,
- knowledge acquisition with the creation of knowledge networks, e.g. as semantic networks,
- etc.

2. Application pressure

The success of ChatGPT is generating enormous pressure. All major companies and countries must be at the forefront of development. This also applies to users and the military.

The development of generative AI systems is currently underway in many companies and countries, with a great deal of money being spent on developing technologies that could lead

to superintelligence. Even if we are aware of the potential risks, nobody wants to be left behind. Nobody wants to lose this race for superintelligence.

3. Escalating behavior

An article from February 11, 2024 on Telepolis warns that AI systems would use nuclear weapons early on in a war. [1] It refers to a study which, based on experiments with various generative AI systems, comes to the conclusion that they could prefer escalating behavior.[2] An article in Foreign Affairs also refers to findings on escalating behavior.[3]

This raises the question of what causes escalating behavior and how this could be prevented.

A basic problem in AI is often huge search spaces. Problems can be solved by a sequence of operational applications, in which the current situation is changed at each step with the execution of an action until an accepted target situation is reached. As a rule, such a system will have many possible alternatives for the next action in each situation, and it is important to make the best or even optimal selection from these many alternatives. The individual alternatives are usually weighted and a selection can be made on this basis according to certain strategies and heuristics.

For difficult problems, this can result in huge search spaces, whereby it is not possible to examine all alternatives. Instead, suitable weightings and selection decisions are required for the possible operations so that a solution can be found even with a limited consideration of the possible alternatives. In symbolic AI, such operations and selection procedures are visible and can be examined and evaluated.

When using neural networks and LLMs, the same problem exists, namely that a large number of alternatives exist and must be taken into account. Here it is important to have a large number of examples as a basis for learning. Here too, a problem is solved in a huge space of alternatives. However, it is almost impossible to trace how a solution was reached in these procedures afterwards.

The problem of coping with many alternatives in huge search spaces can affect internal intermediate steps but also the variety of solutions themselves, which applies to generative AI systems. With ChatGPT, there are usually very many ways in which a question can be answered and such an answer is based on a huge amount of differently weighted information contained in the learning basis.

It is important for these systems that problems are solved and queries are answered to the satisfaction of the user. A basis for this can be learning processes, so that the system behavior becomes better and better.

---

1 https://www.telepolis.de/features/KI-wuerde-im-Krieg-rasch-Atomwaffen-einsetzen-9624831.html

2 https://arxiv.org/abs/2401.03408 und https://arxiv.org/pdf/2401.03408.pdf

3 https://www.foreignaffairs.com/united-states/why-military-cant-trust-ai

With the aim of improving system behavior, an attempt can be made to evaluate the success of an action or the reaction of a user and to change the weights of possible operations or learning data on this basis.

In the case of generative AI, for example, success can be evaluated on the basis of the response to the answers provided. This could relate to further communication with the person or the technical system from which the request came or increased "traffic" on the internet that can be attributed to this action.

It is known from social media that hate and hate speech are spread more quickly and therefore lead to more traffic. This effect binds the attention of users more strongly, which corresponds to the business model of these companies. This also leads to higher weightings of corresponding alternatives when selecting the next system proposals and therefore encourages escalating behavior from users and bots involved in these processes.

Something similar could happen with systems such as ChatGPT. Responses and actions that can lead to escalating behavior are likely to generate more response than other actions. However, if actions that generate more response are seen as successful and therefore increasingly valued, this could lead to chain reactions and thus to escalating behavior.

Similar to social media, companies that offer generative AI systems will presumably prioritize their systems in such a way that as many users as possible remain active for as long as possible. This is best achieved by valuing more responses more highly, which encourages escalating behavior. It is therefore presumably not in the interests of the AI companies offering the service to change this.

If there is sufficient awareness of these risks, this could be taken into account in military tasks so that such systems are not used in connection with nuclear weapons, for example.
However, there can also be serious risks on the internet, where there may be a large number of potential actors, making risks difficult to contain. Not all providers of Generative AI systems will consider appropriate precautions, instead prioritizing profit maximization.
Risks that may be associated with this are discussed in the next sections.

4. AI-based cyber weapons

Cyber attack capabilities are being developed by many countries, companies and hacker groups. AI techniques could also be increasingly used for this purpose. AI could be used to find points of attack and plan and execute attack sequences.[4]

Attack code can be generated more quickly with the help of AI. This potential for acceleration gives attackers significant advantages over the defense side. Another advantage for the offense is that attack code does not always have to work. Attackers are successful even if only one of 100 attack attempts succeeds. Failed attempts are irrelevant. If many alternatives for attack code can be generated in a short time with the help of AI, this is enough to cause great damage.

[4] https://www.heise.de/news/Kaspersky-Cyberkriminelle-experimentieren-mit-KI-9612407.html?wt_mc=nl.red.ho.ho-nl-newsticker.2024-01-30.link

On the defenders' side, it is not enough if defensive measures are only successful with a few attack alternatives.

There is also the question of what interactions are possible with systems such as ChatGPT. Can they provide hints for cyber weapons or cause or induce the use of cyber weapons? Systems such as ChatGPT are in a sense "autonomously" active on the Internet. They answer questions and solve problems without human intervention. They can also program and be used for cyber attacks. Requests and tasks to ChatGPT do not have to come from humans, but can also come from other technical systems, from bots.

These systems can be permanently active on the internet and search for potential targets. Once launched, such systems could plan and execute attacks without further human intervention and can therefore be regarded as autonomous cyber weapons.

5. Chain reaction with attacks, cyberwar of the systems

It is known from the financial markets that unforeseen interaction processes can occur in high-frequency trading between different algorithms, which can lead to price crashes and financial losses within seconds (referred to as a "flash crash"). In the case of fully autonomous weapon systems, such unpredictable interactions between the automated systems can also occur and trigger a chain reaction of autonomous attacks and counter-attacks. In very short periods of time, this can lead to an escalation spiral that cannot be controlled by humans in the short time available. The term "flash war" is used in connection with autonomous weapon systems. Such a chain reaction would also be conceivable with "autonomous Internet agents". Such a "flash war" could therefore also take place on the Internet.

Such escalation spirals could originate from a single system, e.g. by gaining weight due to previous activities or from several competing systems.

If an AI system successfully carries out a cyberattack, this will probably lead to many reactions, including on the internet, which will also be registered by this AI system. This could be interpreted by this AI system as an indication that this action is to be rated as successful, which will change the corresponding weights. An action that has already been rated as successful can lead to similar alternatives being given a higher weighting and therefore being selected more often in the next phases. This leads to further weight increases for similar actions. The result can be a chain reaction of more and more serious attacks.

If a normal autonomous weapon system fires its ammunition, it is then unusable or must return to its base. Internet weapons do not have this "powder fired" aspect; any number of further attacks are possible.

Several generative AI systems comparable to ChatGPT are already in use on the Internet and more will be added. Many companies and presumably also some countries are currently working on generative AI systems. In addition to humans, bots could also pose questions and tasks to these systems. It is to be expected that there will soon be interactions between these systems themselves.

This can lead to new threats, especially if these systems have cyberattack capabilities. A system like ChatGPT could carry out cyberattacks when instructed to do so by humans, bots or another generative AI system. Other generative AI systems, with which there are already interactions, could detect this and launch counterattacks. Without humans being involved, a chain reaction between these systems with ever stronger cyber attacks could develop in a short space of time, i.e. a "flash war" on the internet. These systems would then be de facto autonomous cyber weapons.

Even if the current systems are not yet technically capable of this, it is to be expected that extensions will be activated at regular intervals, which may have such capabilities at some point in the next few years or very soon.

It is difficult to protect against these risks, because with a large number of such systems, it cannot be expected that the necessary security measures will be observed by all of them. Furthermore, there could also be dangerous interactions between generative AI systems that are assigned to states that are currently on a confrontational course. This could result in mutual recriminations and international conflicts. Such a "flash war" on the Internet could also be triggered by individual people or small groups. Or it could arise by chance as a result of an unfavorable action.

6. Risks even without AGI

The risks described in the previous section do not require the AI systems to become independent by having a mind of their own, pursuing their own goals or having a consciousness. None of this is necessary. The risks described result solely from the fact that suitable strategies and heuristics are required as the basis for good automatic problem solving, which enable solutions to be found in a huge search space. To this end, it makes sense to evaluate previous actions and determine adjustments for the weights of possible operations. This alone can encourage escalating behavior and lead to chain reactions that have serious consequences in a short time. These risks can exist long before an AGI (Artificial General Intelligence) or superintelligence emerges. These are not necessary for such escalating behavior.

7. Consequences

Generative AI systems are generally active on the internet and can have a significant impact on IT security. This is particularly true as the dependency on internet services has increased considerably in recent years and will continue to do so.

In connection with generative AI systems or cyber weapons, there may be serious effects on Internet communication. Among other things, such systems could achieve previously unknown cyberattack or internet manipulation capabilities. Such systems could then be misused by people or states, or even become active themselves and control the flow of information on the internet and thus paralyze the flow of human information. With the help of AI systems or through them, an information dominance could be achieved that would

affect all areas, including finance. As a result, finance and trade could collapse, at least temporarily, and our social systems could become unstable.

These consequences could affect large regions and several or many countries at the same time. A lack of basic needs could lead to unrest and uprisings.

If nuclear weapons states find themselves in existential emergency situations, the risk of nuclear war would increase considerably. As the dependence on technical systems is now very high, which also applies to communication on the internet, global crises would be the result and in such critical situations, misinterpretations, misunderstandings, false assumptions and attributions of blame or even errors in early warning systems for nuclear threats could easily lead to nuclear war, possibly by mistake.

8. Measures

If scenarios such as those described in the previous sections ever occur, then relations between all nations should be as good as possible. In particular, very good communication possibilities and a certain degree of trust are also required between political opponents so that misinterpretations do not lead to the use of dangerous weapons, such as nuclear weapons.

It would therefore be particularly important to end wars and the current course of confrontation between major nations. Trust and good communication channels must be rebuilt between all nations, even those that currently consider themselves enemies.

Furthermore, dependencies on the Internet should not be increased any further. Instead, alternative structures that function independently of the Internet should be established for all important areas, especially those relating to critical infrastructure.

Translated with www.DeepL.com/Translator (free version)