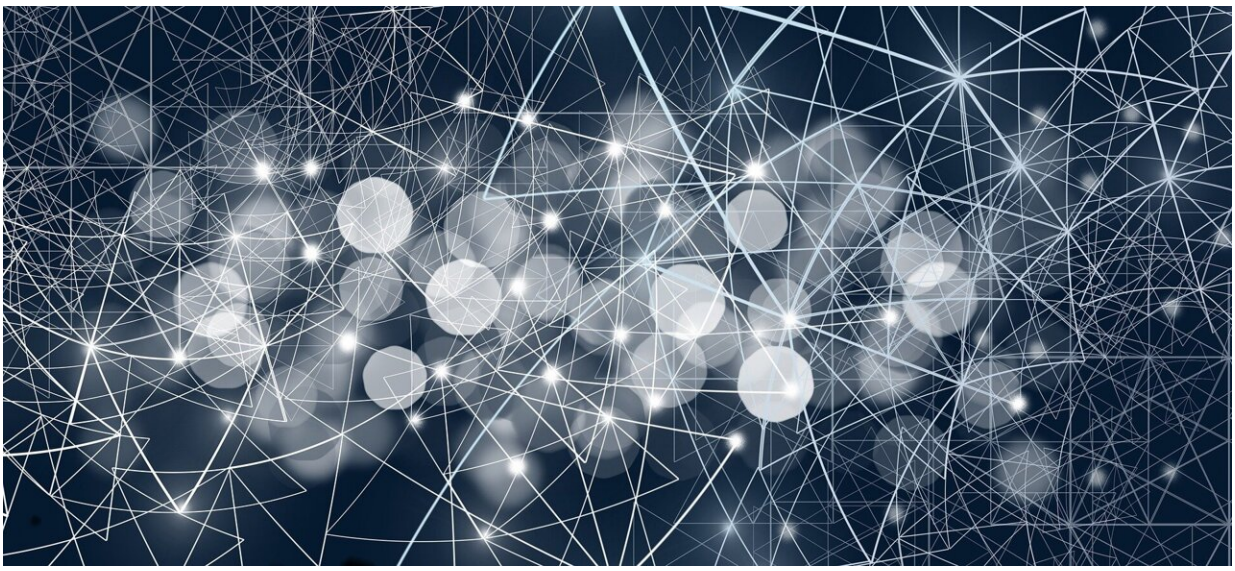
A freeze in training artificial intelligence won't help, says professor

April 3 2023



Credit: Pixabay/CC0 Public Domain

The development of artificial intelligence (AI) is out of control, in the opinion of approximately 3,000 signatories of an open letter published by business leaders and scientists.

The signatories call for a temporary halt to training especially high-performance AI systems. Prof. Urs Gasser, expert on the governance of digital technologies, examines the important questions from which the letter deflects attention, talks about why an "AI technical inspection

agency" would make good sense and looks at how far the EU has come compared to the U.S. in terms of regulation.

Artificial intelligence systems capable of competing with human intelligence may entail grave risks for society and humanity, say the authors of the open letter. Therefore, they continue, for at least six months no further development should be conducted on technologies which are more powerful than the recently introduced GPT-4, successor to the language model ChatGPT.

The authors call for the introduction of safety rules in collaboration with independent experts. If AI laboratories fail to implement a development pause voluntarily, governments should legally mandate the pause, says the signatories.

Professor Gasser, do you support the emergency measures called for in the letter?

Unfortunately the open letter absorbs a lot of attention which would be better devoted to other questions in the AI debate. It is correct to say that today probably nobody knows how to train extremely powerful AI systems in such a way that they will always be reliable, helpful, honest and harmless.

Nonetheless, a pause in AI training will not help achieve this, primarily because it would be impossible to assert such a moratorium on a global level, and because it would not be possible to implement the regulations called for within period of only six months. I'm convinced that what's necessary is a stepwise further development of technologies in parallel to the application and adaptation of control mechanisms.

What issues should preferably be receiving the

attention instead of a moratorium?

First of all, the open letter once again summons up the specter of what is referred to as an artificial general intelligence. That deflects attention from a balanced discussion of the risks and opportunities represented by the kind of technologies currently entering the market. Second, the paper refers to future successor models of GPT-4.

This draws attention away from the fact that GPT-4's predecessor, ChatGPT, already presents us with essential challenges that we urgently need to address—for example misinformation and prejudices which the machines replicate and scale. And third, the spectacular demands made in the letter distract us from the fact that we already have instruments now which we could use to regulate the development and use of AI.

What would such regulations be oriented towards, what instruments do we have?

Recent years have seen the intensive development of ethical principles which should guide the development and application of AI. These have been supplemented in important areas by technical standards and best practices. Specifically, the OECD Principles on Artificial Intelligence link ethical principles with more than 400 concrete tools.

And the US National Institute of Standards and Technology (NIST) has issued a 70-page guideline on how distortions in AI systems can be detected and handled. In the area of security in major AI models, we're seeing new methods like constitutional AI, in which an AI system "learns" principles of good conduct from humans and can then use the results to monitor another AI application. Substantial progress has been made in terms of security, transparency and data protection and there are even specialized inspection companies.

Now the essential question is whether or not to use such instruments, and if so how. Returning to the example of ChatGPT: Will the chat logs of the users be included in the model for iterative training? Are plug-ins allowed which can record user interactions, contacts and other personal data? The interim ban and the initiation of an investigation of the developers of ChatGPT by the Italian data protection authorities are signs that very much is still unclear here.

The open letter demands that no further development of AI systems should take place until one can be confident that the AI systems will have positive effects and their risks are manageable. At what point in development would it be possible to predict the impacts of an AI system so well that this kind of regulation would make sense?

The history of technology has taught us that it is difficult to predict the "good" or "bad" use of technologies, even that technologies often entail both aspects and negative impacts can often be unintentional. Instead of fixating on a certain point in a forecast, we have to do two things: First, we have to ask ourselves which applications we as a society do not want, even if they were possible. We need clear red lines and prohibitions.

Here I'm thinking of autonomous weapons systems as an example. Second, we need comprehensive risk management, spanning the range from development all the way to use. The demands placed here increase as the magnitude of the potential risks to people and the environment posed by a given application grow. European legislature is correct in taking this approach.

According to the proposal, independent experts

should assess the risks of AI.

This kind of independent inspection is a very important instrument, especially when it comes to applications that can have a considerable impact on human beings. And by the way, this is not a new idea: we already see inspection procedures and instances like these at work in the wide variety of aspects of life, ranging from automobile inspections to general technical equipment inspections and financial auditing.

However, the challenge is disproportionally greater with certain AI methods and applications, because certain systems develop themselves as they are used, i.e. they are dynamic in nature. And it's also important to see that experts alone won't be able to make a good assessment of all societal impacts. We also need innovative mechanisms which for example include disadvantaged people and underrepresented groups in the discussion on the consequences of AI. This is no easy job, one I wish was attracting more attention.

The authors also address the political sector. Politics would be responsible for anchoring such an 'AI technical inspection agency' in the system.

We do indeed need clear legal rules for artificial intelligence. At the EU level, an act on AI is currently being finalized which is intended to ensure that AI technologies are safe and comply with fundamental rights. The draft bill provides for the classification of AI technologies according to the threat they pose to these principles, with the possible consequence of prohibition or transparency obligations.

For example, plans include prohibiting evaluation of private individuals in terms of their social behavior, as we are currently seeing in China. In the U.S. the political process in this field is blocked in Congress. It

would be helpful if the prominent figures who wrote the letter would put pressure on US federal legislators to take action instead of calling for a temporary discontinuation of technological development.

Provided by Technical University Munich