THE BRIGHT FUTURE OF ARTIFICIAL INTELLIGENCE IS NOT WHERE EXPECTED

Mykola Rabchevskiy

At the beginning of my professional career, I participated in a story so epic as to be comparable to that of the bible. The research group that I became a member of focused on optimizing mechanical systems' performance — improving their performance. Our chief has agreed to cooperate with a large electric generator designed for a nuclear power plant. The aim was to reduce the risk of an accident due to vibration in the generator when starting and stopping.

Arriving at the design bureau, we asked the engineers to discuss the generator design and say which design parameters and limits could be changed. Speaking about each structure element, the engineers explained why these parameter values were chosen and why other options were unacceptable. When the story ended, it turned out that nothing could be changed. Our leader turned to the chief designer: "but how can the design be improved if nothing can be changed?" - and received an immediate answer: "***well, you are optimization specialists - you must know how to do it***." The comic nature of the situation was immediately appreciated by both sides; we re-discussed each of the structural elements and agreed, agreeing on an acceptable range of parameters.

This story illustrates truths that would appear obvious but are not always taken into account: - you cannot improve something without changing something in it - to change something that exists, you need to abandon the decisions made when creating it

I remember this story every time I read another article about how we will finally get better AI systems that will produce explainable and (if necessary) easily corrected decisions.

When an object unknown to it appears, the AGI system will understand that it is something new and not take it for something known. It would also be able to detect the causes of events based on accumulated experience. I remember the story told so look for information in these articles about ***what will be changed*** in the versions of AI currently being used to finally implement old AGI ideas - and, accordingly, ***which of the decisions underlying the current versions of AI will be discarded***. But, alas, no one writes about this. It looks as if the leaders in AI development intend to accomplish the impossible: radically improve artificial intelligence without changing anything in essence.

There is a predictable rejoinder to this stance: there is a competition to increase the number of parameters of the neural network of their version of deep learning; millions of parameters are replaced by billions, billions – by trillions; these are changes! And here is just the right moment to discuss the ***third*** fundamental principle of optimization in addition to the two above. It is about the ***sensitivity of the result to changes*** in decisions.

When the task is to improve something existing, and there is a list of things that can be changed by abandoning some of the previous decisions, it turns out that possible changes can affect the result in very different ways. For example, in the design of a car, you can replace the engine, change the color, and change the body's shape. But suppose the goal is to reduce carbon dioxide emissions. In that case, replacing the engine has a significant effect, a change in form has a weaker effect, and a color change has no effect. Suppose the task is to make the car more noticeable to reduce the accident rate. Then, the paint affects the result significantly, and the engine does not. The sensitivity of the quality criterion to various parameters is different, and the ***rational way of improvement is to change those decisions that give the most significant effect***. This, of course, is essentially also obvious but is likewise not always taken into account; a variant of the well-known anecdote about finding a key under a lamp, and not where it was lost, in real life is found at every step.

Let's look at the sensitivity of some criteria for evaluating AI (explainability, understanding the meaning of phrases, permanent self-learning, the ability to detect cause-and-effect relationships, and so on). These criteria are no more sensitive to increasing the number of neural network parameters than the volume of carbon dioxide emissions from a car to its color. An increase in the number of parameters is not useless, of course — the amount of information that the AI system is capable of storing grows, as does its usefulness – but it does ***not lead to the appearance of the desired qualities***.

Does this mean that the human level of intelligence is unattainable? Of course not – this level of intelligence is achievable, but not by increasing the neural network size. Rather, we must try other solutions underlying AI, those to which the demanded criteria are sensitive.

The neural network approach to building AI is based on the idea that if the brain is a natural network of neuron elements, then artificial intelligence similar to a human can be made based on a computer neural network. This conclusion is undoubtedly correct, but the devil is always in the details. Not any neural network of any neurons will demonstrate human-level intelligence. And those versions of neural networks that are now developed and used are obviously ***radically different from the natural brain***. This is clear since of all the abilities of the human brain, these neural nets implement only one: ***recognition of an object from a set, on which the system is trained in advance, before the start of operation***; the reverse process of forming a generalized image for the specified object-concept is also realizable (again, from the ***set studied before using*** the system in practice). The rest of the required features remains unrealized.

This indicates that ***knowledge representation*** by modern neural networks is significantly different from the representation of knowledge in the human brain. This is evidenced not only by the apparent difference in the demonstrated capabilities of the systems. First of all, absolutely all AI developers know that the ***natural network*** of neurons has a ***variable structure***: connections between neurons are created and disappear during the functioning of the brain, making it ***possible to encode stored knowledge by the connections structure***. So the network may reflect ***logical connections*** between the elements of accumulated knowledge. In other words, the natural ***brain's structure*** can reflect the ***knowledge structure***. In modern neural networks, the design of connections is ***unchanged from the beginning*** and therefore cannot be used as a "replica" of any conceptual structures; all stored knowledge is encoded only by the numerical parameters of the connections of computer "neurons." Such a ***distributed representation of knowledge*** results in a chain of consequences that limit the possibilities of such an approach.

The name "***distributed representation***" reflects the fact that when knowledge is represented in the form of a fixed structure of links and the variables of the parameters of these links, the stored ***knowledge is not localized***: there is ***no way to indicate which of their parameters are responsible for which part of the accumulated knowledge***. If you need to adjust a ***small piece*** of knowledge, there is no way to do this by changing the ***corresponding bit of data***; it is necessary to ***change all the quantities***. In turn, a change in each parameter affects potentially every element of stored knowledge; that is, any ***addition of new knowledge potentially changes everything memorized***. This circumstance is a ***fundamental obstacle*** to permanent learning, adjustability, and, indirectly, other required properties.

Humans and animals obviously do not have such restrictions; learning something new does not damage previously accumulated knowledge and skills. While forgetting occurs, it depends only on whether specific knowledge is referred to. So the natural brain is obviously not based exclusively on a distributed representation of knowledge. This makes it possible to permanently replenish and correct knowledge throughout life. The explainability of decisions is based on the fact that the structure of connections reflects logical connections between concepts; the same is required in analyzing cause-and-effect relationships. These claims will seem dubious to many (the argument in their favor is too voluminous to be given here; skeptics may consider this a hypothesis to be confirmed).

Thus, these ***criteria for assessing AI necessary to achieve a higher level of intelligence are sensitive to how knowledge is represented*** and not to the number of parameters involved in implementing the representation. ***Increasing the network size does not lead to new abilities*** (although it is helpful for applications that do not require "full" intelligence).

The capabilities of permanent learning, adaptability, detection of cause-effect relationships, and previously unknown objects required from AGI are operating with ***logical things***. This, in turn, requires ***separate access to concepts and facts***; that is, they need the ability to ***add, delete, and change the details of an individual object without damaging the rest of the accumulated data***. Neural networks of a fixed structure cannot provide this due to distributed information representation. Attempts to implement purely abstract actions (for example, arithmetic operations) are used in one form or another to ***supplement the neural network with addressable, that is, localized memory***. Moreover, ***convolution*** operations, which have significantly expanded the capabilities of neural networks, introduce an ***element of addressing using indexe***s.

Semantic networks (directed graphs that reflect relationships between logical entities) are a natural alternative for representing knowledge. The capabilities of these two alternatives are radically different, despite both being based on a network of similar elements. The structure of the connection of neurons in the brain, the design of links in the artificial neural network Deep Learning, and the structure of connections of the elements of the semantic network are described mathematically in one and the same way as ***directed graphs***. The difference is that the brain and semantic networks implement the principle of information ***localization using network nodes to represent logical entities***. Still, the current neural networks are based on a ***distributed representation*** of information by connection parameters.

The preceding makes it possible to predict significant progress in the field of AI soon, but ***not as a result of increasing the size of existing versions of neural networks, but rather as a result of replacing the way of representing information***. Such a change entails an equally significant difference in the ***algorithms*** and ***methods of training*** used. Awareness of this aspect by the community of specialists involved in the development of AI is close to the critical value that leads to breakthrough solutions. Information storage systems

(databases) based on ***variable structure graphs*** have become widespread in recent years due to the versatility of the approach. This is the proper foundation on which the "true" artificial intelligence, which the inventors of the term had in mind back in 1956, will be built.

The story described at the beginning tells of a situation where an engineer, knowing that it is impossible to improve a product without changing it, ignores this knowledge at some point. This case explains the phenomenon that AI developers, knowing about the innate restrictions of distributed representation, about what a technological leap was the transition from analog computers with distributed memory to digital computers with localized representation, no less hope to build true intelligence without changing the paradigm. As a result - the search "under the lantern" and not where it is possible to find intelligence. For example, experiments on implementing arithmetic by neural networks essentially come down to ***"reducing" a single elementary computer operation to millions of them***.

Subscribe to AGI engineering

By Mykola Rabchevskiy · Launched 2 years ago

AGI: fundamentals, architecture, implementation, source code