Why AI won't take over the world

Peter Gärdenfors[1]

Review of Jobst Landgrebe and Barry Smith, *Why Machines Will Never Rule the World*, Routledge, 2023

Artificial intelligence has recently had spectacular successes. The capabilities of language programs such as ChatGPT, Bing AI and Bard and of imaging programs such as Midjourney and DALL-E 2 have surprised many. There have also been breakthroughs in other areas, for example when it comes to describing the three-dimensional structure of proteins, which is a difficult problem for researchers in biomedicine.

The rapid success of AI has led to overconfidence in what is possible for AI systems to achieve. Many AI researchers, among them the Swedes Nick Boström, Max Tegmark and Olle Häggström, claim that AI will soon develop into AGI – which is to say: artificial *general* intelligence. Such a system is described as having all the intellectual abilities that humans have and more. Some researchers claim that there is a danger that AGI will take over the world. They perceive the development of AGI as an engineering problem and see no principled obstacles. The question is whether there are sufficient arguments for this opinion.

Central to this question is how one could determine whether an AI system really has general intelligence. The fact that a computer program is better than a human in some specialized area – such as playing chess or recognizing faces – says very little about general intelligence. When it comes to people, IQ is often used as a measure of intelligence – mostly because there is nothing better. But this metric doesn't work for machines. It would be relatively easy to construct a program that achieves top score on the intelligence tests used – not because the program is particularly intelligent, but because the tests follow limited mathematical, linguistic, and visual patterns, and the vocabulary the program needs can be easily obtained from the Internet.

---

[1] Published in Swedish as "Varför AI inte kommer att ta över världen", SANS 2, 2024

Curiously, discussions about how to measure AGI among AI researchers are quite shallow. Boström gives three proposals for criteria in his book *Superintelligence* (Oxford University Press, 2014). The strongest claim is what he calls 'quality superintelligence' (p. 56), which he defines as 'a system that is at least as fast as a human mind and qualitatively far smarter'. Unfortunately, this doesn't say very much, because you have to know what a human mind is capable of, and the word 'smarter' makes it almost a circular definition.

Häggström is more precise. In the book *Tänkande maskiner* (*Thinking Machines*, (*Fri Tanke*, 2021) he defines AGI as a system that has 'all the abilities that underlie human intelligence: short- and long-term memory, logical thinking, mathematical ability, geometric and spatial visualization, pattern recognition, induction, planning , creativity, social manipulation and many others'. If the system's intelligence exceeds human intelligence, the system is said to be superintelligent. Such a description provides a better tool for assessing the thinking ability of machines.

A definition that can be applied to humans and animals as well as machines comes from the German psychologist William Stern: 'Intelligence is the general ability of an individual to consciously adapt his thinking to new requirements; it is a general mental adaptability to new problems and life conditions.' Humans (and animals) can respond *immediately* to new situations, although the reactions are not always the best. AI systems cannot handle cases that go outside the domain for which they have been trained, and their intelligence is thus limited according to Stern's definition.

In addition to being able to determine whether an AI program is superintelligent, a central problem is how to *construct* such a program. None of the researchers in the field have any constructive ideas about how this should be done. They often imply that once a program reaches a certain level, it will evolve itself into increasingly advanced general intelligence.

However, there are good reasons to believe that AGI will not be that revolutionary. A huge collection of arguments for this is presented in the new book *Why machines will never rule the world,* written by AI researcher Jobst Landgrebe and philosopher Barry Smith. Their main argument can be summarized as follows: human intelligence arises in a very complex system consisting of the brain's interaction with the body and of the body's interaction with other individuals and the surrounding world. Systems with this degree of complexity cannot be captured in mathematical models. Therefore, they will

never be able to be reproduced in AI systems. This argument is substantiated in the book's three parts.

In the first part, the authors review some of the characteristics of human thinking. Above all, they highlight the complexity of language. ChatGPT and similar systems respond to texts typed on computer screens. Human language, however, is mainly a matter of dialogues, where the interaction is highly dependent on the context: the theme of the discussion (which may change along the way), the speakers' intentions in taking part in the dialogue, their expectations of the other participants in the dialogue, their memory of previous interactions, the environment, and so on. Existing language programs cannot handle such factors. For example, they have no intentions underlying their language production. Landgrebe and Smith argue that human dialogues are so varied and situational that it is impossible to collect enough data for an AI system to learn how to deal with them.

Another area where AI systems fail is human empathy. We can, almost automatically, interpret other people's feelings, intentions, values and knowledge. For example, understanding that someone is being ironic means that you understand that the person who is speaking ironically does not mean what she says. It is extremely difficult for an AI system to pick up on the subtle cues that lead to interpreting an utterance as ironic.

In the area of *affective computing*, some researchers try to make AI systems understand people's emotions. People's language, facial expressions and body language are used as input. So far, researchers have not progressed very far in this area. They also want the AI systems' values to match those of humans. This too is troublesome because it is not clear what it means for a system to evaluate.

In the second, more technical part of the book, Landgrebe and Smith argue that the vast majority of natural systems are so complex that they cannot be modeled. Thus, they cannot be handled by any computer system. Even the simplest forms of life are so complex that they cannot be simulated by a computer. A single biological cell contains around one hundred billion atoms that form one hundred thousand different RNA molecules. Living systems are also self-organizing, and they maintain themselves by drawing energy from their surroundings. The nervous systems of animals, perhaps above all that of humans, are the most advanced biological systems in existence. The models of neurons used in AI systems are radical simplifications of real biological cells.

When comparing human intelligence to AI, a common argument is to compare the number of neurons in the human brain to the number of artificial

neurons that AI systems use. As support for the systems' intelligence, it has been pointed out, for example, that the large language model GPT-4 has the equivalent of one hundred billion neurons, while the human brain has eighty to one hundred billion neurons. This comparison does not hold, however, because the brain is made up of so much more than just neurons. Neurotransmitters such as dopamine, adrenaline and oxytocin play a large role in brain processes, and these have no counterparts in AI systems. A new theory also claims that the magnetic fields created by the electrical currents in the neurons also affect processes throughout the brain. Such a phenomenon cannot be captured in the artificial neural networks that AI systems use.

The brain is not a machine. Even if we could measure the brain's molecular properties precisely enough, this data would not allow the creation of AGI because there is no model that can describe how these properties relate to each other. In short, the assumption that the brain's activities can be captured in a computer system does not hold. All claims that one could "upload" a human brain to a computer therefore fall flat.

A technical concept central to Landgrebe and Smith's argument is *ergodicity*. Slightly simplified, a system is ergodic if the data you can collect about the system in the long run become representative of the system's behavior. Landgrebe and Smith argue that computers (and any other system that performs calculations) can only model ergodic systems. They also argue that most natural systems are not ergodic. That is, no matter how long we study such a system, we will never be able to predict its behavior. Tomas Tranströmer has an apt metaphor for this: 'An abstract picture of the world is as impossible as a blueprint for a storm.' The artificial neural networks used in so-called *deep learning* are based on statistical patterns, and thus they cannot handle situations that fall outside the framework given by their training data.

Landgrebe and Smith are probably right that most natural systems are not ergodic, but they cannot prove this. It is conceivable that the behavior of simple biological systems, for example insects, can be described ergodically, even if the individual cells are non-ergodic. The behavior at the macro level can perhaps be described exhaustively even if at the micro level it is still incalculable.

In the third part of the book, they describe the limitations of AI systems in several areas and argue that there is nothing that comes close to AGI. In particular, this applies to human language and human behavior. ChatGPT and other language models find complex patterns in sequences of words, but they

do not understand the meaning of the words. The systems mimic the patterns of human language, but they do not *interpret* them. In addition, the answers become increasingly flat if you continue to chat with them, because the system cannot keep track of the context of the dialogue.

Another limitation is that the language models are text-based, where human communication is based on so much more than text. The systems cannot see the person they are talking to. A dialogue is also influenced by tone of voice, non-linguistic sounds, glances, facial expressions, gestures and so on.

It will be equally difficult to construct systems that exhibit something similar to human empathy. We cannot build machines that have intentions and will, because we know too little about how they arise in humans. Several AI researchers postulate that this is possible, without, however, providing any arguments.

Although Landgrebe and Smith believe that AGI will never be achieved, they are not opposed to AI. Indeed, they provide examples in several areas of how AI can develop. One area they highlight is disease diagnoses. However, since the human body is not an ergodic system, the responses of the AI systems must always be interpreted by doctors to cover the cases that the algorithms cannot catch. A more unpleasant area, which is sure to grow, is military systems. Such applications are particularly dangerous, as they do not exhibit AGI and they, too, cannot adapt to new situations.