Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-020-0494-4 OPEN

Why general artificial intelligence will not be realized

Ragnar Fjelland[1]✉

The modern project of creating human-like artificial intelligence (AI) started after World War II, when it was discovered that electronic computers are not just number-crunching machines, but can also manipulate symbols. It is possible to pursue this goal without assuming that machine intelligence is identical to human intelligence. This is known as weak AI. However, many AI researcher have pursued the aim of developing artificial intelligence that is in principle identical to human intelligence, called strong AI. Weak AI is less ambitious than strong AI, and therefore less controversial. However, there are important controversies related to weak AI as well. This paper focuses on the distinction between artificial general intelligence (AGI) and artificial narrow intelligence (ANI). Although AGI may be classified as weak AI, it is close to strong AI because one chief characteristics of human intelligence is its generality. Although AGI is less ambitious than strong AI, there were critics almost from the very beginning. One of the leading critics was the philosopher Hubert Dreyfus, who argued that computers, who have no body, no childhood and no cultural practice, could not acquire intelligence at all. One of Dreyfus' main arguments was that human knowledge is partly tacit, and therefore cannot be articulated and incorporated in a computer program. However, today one might argue that new approaches to artificial intelligence research have made his arguments obsolete. Deep learning and Big Data are among the latest approaches, and advocates argue that they will be able to realize AGI. A closer look reveals that although development of artificial intelligence for specific purposes (ANI) has been impressive, we have not come much closer to developing artificial general intelligence (AGI). The article further argues that this is in principle impossible, and it revives Hubert Dreyfus' argument that computers are not in the world.

[1] Centre for the Study of the Sciences and the Humanities, University of Bergen, Bergen, Norway. ✉email: ragnar.fjelland@uib.no

Introduction

The idea of machines that can perform tasks that require intelligence goes at least back to Descartes and Leibniz. However, the project made a major step forward when in the early 1950s it was recognized that electronic computers are not only number-crunching devices, but may be made to manipulate symbols. This was the birth of artificial intelligence (AI) research. It is possible to pursue this goal without assuming that machine intelligence is identical to human intelligence. For example, one of the pioneers in the field, Marvin Minsky, defined AI as: "… the science of making machines do things that would require intelligence if done by men" (quoted from Bolter, 1986, p. 193). This is sometimes called weak AI. However, many AI researcher have pursued the aim of developing AI that is in principle identical to human intelligence, called strong AI. This entails that "…the appropriately programmed computer is a mind, in the sense that computers can be literally said to understand and have other cognitive states" (Searle, 1980, p. 417).

In this paper, I shall use a different terminology, which is better adapted to the issues that I discuss. Because human intelligence is general, human-like AI is therefore often called artificial general intelligence (AGI). Although AGI possesses an essential property of human intelligence, it may still be regarded as weak AI. It is nevertheless different from traditional weak AI, which is restricted to specific tasks or areas. Traditional weak AI is therefore sometimes called artificial narrow intelligence (ANI) (Shane, 2019, p. 41). Although I will sometimes refer to strong AI, the basic distinction in this article is between AGI and ANI. It is important to keep the two apart. Advances in ANI are not advances in AGI.

In 1976 Joseph Weizenbaum, at that time professor of informatics at MIT and the creator of the famous program *Eliza*, published the book *Computer Power and Human Reason* (Weizenbaum, 1976). As the title indicates, he made a distinction between computer power and human reason. Computer power is, in today's terminology, the ability to use algorithms at a tremendous speed, which is ANI. Computer power will never develop into human reason, because the two are fundamentlly different. "Human reason" would comprise Aristotle's prudence and wisdom. Prudence is the ability to make right decisions in concrete situations, and wisdom is the ability to see the whole. These abilities are not algorithmic, and therefore, computer power cannot—and should not—replace human reason. The mathematician Roger Penrose a few years later wrote two major books where he showed that human thinking is basically not algorithmic (Penrose, 1989, 1994).

However, my arguments will be slightly different from Weizenbaum's and Penrose's. I shall pursue a line of arguments that was originally presented by the philosopher Hubert Dreyfus. He got into AI research more or less by accident. He had done work related to the two philosophers Martin Heidegger and Ludwig Wittgenstein. These philosophers represented a break with mainstream Western philosophy, as they emphasized the importance of the human body and practical activity as primary compared to the world of science. For example, Heidegger argued that we can only have a concept of a hammer or a chair because we belong to a culture where we grow up and are able to handle these objects. Dreyfus therefore thought that computers, who have no body, no childhood and no cultural practice, could not acquire intelligence at all (Dreyfus and Dreyfus, 1986, p. 5).

One of the important places for AI research in the 1950s and 1960s was Rand Corporation. Strangely enough, they engaged Dreyfus as a consultant in 1964. The next year he submitted a critical report titled: "Alchemy and Artificial Intelligence". However, the leaders of the AI project at Rand argued that the report was nonsense, and should not be published. When it was finally released, it became the most demanded report in the history of Rand Corporation. Dreyfus later expanded the report to the book *What Computers Can't Do* (Dreyfus, 1972). In the book he argued that an important part of human knowledge is tacit. Therefore, it cannot be articulated and implemented in a computer program.

Although Dreyfus was fiercely attacked by some AI researchers, he no doubt pointed to a serious problem. But during the 1980s another paradigm became dominant in AI research. It was based on the idea of *neural networks*. Instead of taking manipulation of symbols as model, it took the processes in our nervous system and brain as model. A neural network can learn without receiving explicit instructions. Thus it looked as if Dreyfus' arguments for what computers cannot do were obsolete.

The latest off-spring is Big Data. Big Data is the application of mathematical methods to huge amounts of data to find correlations and infer probabilities (Najafabadi et al., 2015). Big Data poses an interesting challenge: I mentioned previously that AGI is not part of strong AI. However, although Big Data does not represent the ambition of developing strong AI, advocates argued that this is not necessary. We do not have to develop computers with human-like intelligence. On the contrary, we may change our thinking to be like the computers. Implicitly this is the message of Viktor Mayer-Schönberger and Kenneth Cukier's book: *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Mayer-Schönberger and Cukier, 2014). The book is optimistic about what Big Data can accomplish and its positive effects on our personal lives and society as a whole.

Some even argue that the traditional scientific method of using hypotheses, causal models, and tests is obsolete. Causality is an important part of human thinking, particularly in science, but according to this view we do not need causality. Correlations are enough. For example, based on criminal data we can infer where crimes will occur, and use it to allocate police resources. We may even be able to predict crimes before they are committed, and thus prevent them.

If we look at some of the literature on AI research it looks as if there are no limits to what the research can accomplish within a few decades. One example is Mayer-Schönberger and Cukier's book that I referred to above. Here is one quotation:

> In the future—and sooner than we may think – many aspects of our world will be augmented or replaced by computer systems that today are the sole purview of human judgment (Mayer-Schönberger and Cukier, 2014, p. 12).

An example that supports this view is the Obama Administration, which in 2012 announced a "Big Data Research and Development Initiative" to "help solve some of the Nations's most pressing challenges" (quoted from Chen and Lin, 2014, p. 521).

However, when one looks at what has actually been accomplished compared to what is promised, the discrepancy is striking. I shall later give some examples. One explanation for this discrepancy may be that profit is the main driving force, and, therefore, many of the promises should be regarded as marketing. However, although commercial interests no doubt play a part, I think that this explanation is insufficient. I will add two factors: First, one of the few dissidents in Silicon Valley, Jerone Lanier, has argued that the belief in scientific immortality, the development of computers with super-intelligence, etc., are expressions of a new religion, "expressed through an engineering culture" (Lanier, 2013, p. 186). Second, when it is argued that computers are able to duplicate a human activity, it often turns out that the claim presuppose an account of that activity that is seriously simplified and distorted. To put it simply: The overestimation of

technology is closely connected with the underestimation of humans.

I shall start with Dreyfus' main argument that AGI cannot be realized. Then I shall give a short account of the development of AI research after his book was published. Some spectacular breakthroughs have been used to support the claim that AGI is realizable within the next few decades, but I will show that very little has been achieved in the realization of AGI. I will then argue that it is not just a question of time, that what has not been realized sooner, will be realized later. On the contrary, I argue that the goal cannot in principle be realized, and that the project is a dead end. In the second part of the paper I restrict myself to arguing that causal knowledge is an important part of humanlike intelligence, and that computers cannot handle causality because they cannot intervene in the world. More generally, AGI cannot be realized because computers are not in the world. As long as computers do not grow up, belong to a culture, and act in the world, they will never acquire human-like intelligence.

Finally, I will argue that the *belief* that AGI can be realized is harmful. If the power of technology is overestimated and human skills are underestimated, the result will in many cases be that we replace something that works well with something that is inferior.

Tacit knowledge

Dreyfus placed AI into a philosophical tradition going back to Plato. Plato's theory of knowledge was constructed on the ideal of mathematics, in particular geometry. Geometry is not about material bodies, but ideal bodies. We can only acquire real knowledge, episteme, by turning the attention away from the material world, and direct it "upwards", to the world of ideal objects. Plato even criticized the geometers for not understanding their own trade, because they thought they were "… doing something and their reasoning had a practical end, and the subject were not, in fact, pursued for the sake of knowledge" (Plato, 1955, p. 517). Skills are merely opinion, doxa, and are relegated to the bottom of his knowledge hierarchy.

According to this view, a minimum requirement for something to be regarded as knowledge is that it can be formulated explicitly. Western philosophy has by and large followed Plato and only accepted propositional knowledge as real knowledge. An exception is what Dreyfus called the "anti-philosophers" Merleau-Ponty, Heidegger, and Wittgenstein. He also referred to the scientist and philosopher Michael Polanyi. In his book, *Personal Knowledge* Polanyi introduced the expression *tacit knowledge*[1]. Most of the knowledge we apply in everyday life is tacit. In fact, we do not know which rules we apply when we perform a task. Polanyi used swimming and bicycle riding as examples. Very few swimmers know that what keeps them afloat is how they regulate their respiration: When they breathe out, they do not empty their lungs, and when they breathe in, they inflate their lungs more than normal.

Something similar applies to bicycle riding. The bicycle rider keeps his balance by turning the handlebar of the bicycle. To avoid falling to the left, he moves the handlebar to the left, and to avoid falling to the right he turns the handlebar to the right. Thus he keeps his balance by moving along a series of small curvatures. According to Polanyi a simple analysis shows that for a given angle of unbalance, the curvature of each winding is inversely proportional to the square of the speed of the bicycle. But the bicycle rider does not know this, and it would not help him become a better bicycle rider (Polanyi, 1958, p. 50). Later Polanyi formulated this insight as "…*we can know more than we can tell*" (Polanyi, 2009, p. 4, italics in original).

However, the important thing in Polanyi's contribution is that he argued that skills are a precondition for articulate knowledge in general, and scientific knowledge in particular. For example, to carry out physical experiments requires a high degree of skills. These skills cannot just be learned from textbooks. They are acquired by instruction from someone who knows the trade.

Similarly, Hubert Dreyfus, in cooperation with his brother Stuart, developed a model for acquisition of skills. At the lowest level the performer follows explicit rules. The highest level, expert performance, is similar to Polanyi's account of scientific practice. An important part of expertise is tacit. The problem facing the development of expert systems, that is, systems that enable a computer to simulate expert performance (for example medical diagnostics) is that an important part of the expert knowledge is tacit. If experts try to articulate the knowledge they apply in their performance, they normally regress to a lower level. Therefore, according to Hubert and Stuart Dreyfus, expert systems are not able to capture the skills of an expert performer (Dreyfus and Dreyfus, 1986, p. 36). We know this phenomenon from everyday life. Most of us are experts on walking. However, if we try to articulate how we walk, we certainly give a description that does not capture the skills involved in walking.

Three "milestones" in AI research

However, after Hubert Dreyfus published *What Computers Can't Do*, AI has made tremendous progress. I will mention three "milestones" that have received public attention and contributed to the impression that AGI is just "around the corner".

The first "milestone" is IBM's chess-playing computer *Deep Blue*, which is often regarded as a breakthrough when it in 1997 defeated the world champion of chess, Garri Kasparov. However, Deep Blue was an example of ANI; it was made for a specific purpose. Although it did extremely well in an activity that requires intelligence when performed by humans, no one would claim that Deep Blue had acquired general intelligence.

The second is IBM's computer *Watson*. It was developed with the explicit goal of joining the quiz show *Jeopardy!*. This is a competition where the participants are given the answers, and are then supposed to find the right questions. They may for example be presented the answer: "This 'Father of Our Country' didn't really chop down a cherry tree". The correct question the participants are supposed to find is: "Who was George Washington?"[2]

Jeopardy! requires a much larger repertoir of knowledge and skills than chess. The tasks cover a variety of areas, such as science, history, culture, geography, and sports, and may contain analogies and puns. It has three participants, competing to answer first. If you answer incorrectly, you will be drawn and another of the participants will have the opportunity to answer. Therefore, the competition requires both knowledge, speed, but also the ability to limit oneself. The program has enjoyed tremendous popularity in the United States since it began in 1964, and is viewed by an average of seven million people (Brynjolfson and McAfee, 2014, p. 24).

Watson communicates using natural language. When it participated in *Jeopardy!* it was not connected to the Internet, but had access to 200 million pages of information (Susskind and Susskind, 2015, p. 165; Ford, 2015, p. 98ff). In 2011 it beat the two best participants in *Jeopardy!*, Ken Jennings and Brad Rutter. Jennings had won 74 times in a row in 2004, and had received over $3 million in total. Rutter had won over Jennings in 2005, and he too had won over $3 million. In the 2-day competition, Watson won more than three times as much as each of its human competitors.

Although Watson was constructed to participate in *Jeopardy!*, IBM had further plans. Shortly after Watson had won *Jeopardy!* the company announced that they would apply the power of the computer to medicine: It should become an AI medical

super-doctor, and revolutionize medicine. The basic idea was that if Watson had access to all medical literature (patients' health records, textbooks, journal articles, lists of drugs, etc.) it should be able to offer a better diagnosis and treatment than any human doctor. In the following years IBM engaged in several projects, but the success has been rather limited. Some have just been closed down, and some have failed spectacularly. It has been much more difficult than originally assumed to construct an AI doctor. Instead of super-doctors IBM's Watson Health has turned out AI assistants that can perform in routine tasks (Strickland, 2019).

The third "milestone" is Alphabet's *AlphaGo*. Go is a board game invented more than 2000 years ago in China. The complexity of the game is regarded as even larger than chess, and it is played by millions of people, in particular in East Asia. In 2016, AlphaGo defeated the world champion Le Sedol in five highly publicized matches in Seoul, South Korea. The event was documented in the award-winning film *AlphaGo* (2017, directed by Greg Kohs).

AlphaGo is regarded as a milestone in AI research because it was an example of the application of a strategy called *deep reinforcement learning*. This is reflected in the name of the company, which is DeepMind. (After a reconstruction of Google, Google and DeepMind are subsidiaries of Alphabet.) It is an example of an approach to AI research that is based on the paradigm of artificial neural networks. An artificial neural network is modeled on neural networks. Our brain contains approximately one hundred billion neurons. Each neuron is connected to approximately 1000 neurons via synapses. This gives around a hundred trillion connections in the brain. An artificial neural network consists of artificial neurons, which are much simpler than natural neurons. However, it has been demonstrated that when many neurons are connected in a network, a large enough network can in theory carry out any computation. What is practically possible, is of course a different question (Minsky, 1972, p. 55; Tegmark, 2017, p. 74).

Neural networks are particularly good at pattern recognition. For example, to teach a neural network to identify a cat in a picture we do not have to program the criteria we use to identify a cat. Humans have normally no problems distinguishing between, say, cats and dogs. To some degree we can explain the differences, but very few, probably no one, will be able to give a complete list of all criteria used. It is for the most part tacit knowledge, learned by examples and counter-examples. The same applies to neural networks.

A deep learning neural network consists of different layers of artificial neurons. For example, a network may have four different layers. In analyzing a picture the first layer may identify pixels as light and dark. The second layer may identify edges and simple shapes. The third layer may identify more complex shapes and objects, and the fourth layer may learn which shapes can be used to identify an object (Jones, 2014, p. 148).

The advantage is that one must not formulate explicitly the criteria used, for example, to identify a face. This is the crucial difference between the chess program Deep Blue and AlphaGo. Although a human chess player uses a mixture of calculation and intuition to evaluate a particular board position, Deep Blue was programmed to evaluate numerous possible board positions, and decide the best possible in a given situation. Go is different. In many cases expert players relied on intuition only, and were only able to describe a board position as having "good shape" (Nielsen, 2016). I have mentioned earlier that one of Hubert Dreyfus' main arguments against AGI was that human expertise is partly tacit, and cannot be articulated. AlphaGo showed that computers can handle tacit knowledge, and it therefore looks as if Dreyfus' argument is obsolete. However, I will later show that this "tacit knowledge" is restricted to the idealized "world of science", which is fundamentally different from the human world that Dreyfus had in mind.

The advantage of not having to formulate explicit rules comes at a price, though. In a traditional computer program all the parameters are explicit. This guarantees full transparency. In a neural network this transparency is lost. One often does not know what parameters are used. Some years ago a team at University of Washington developed a system that was trained to distinguish between huskies and wolves. This is a task that requires considerable skill, because there is not much difference between them. In spite of this the system had an astonishing 90% accuracy. However, the team discovered that the system recognized wolves because there was snow on most of the wolf pictures. The team had invented a snow detector! (Dingli, 2018).

AlphaGo was developed by the researchers of DeepMind, and is regarded as a big success. DeepMind's approach was also applied successfully to the Atari games Breakout and Space Invaders, and the computer game Starcraft. However, it turned out that the system lacks flexibility, and is not able to adapt to changes in the environment. It has even turned out to be vulnerable to tiny changes. Because real world problems take place in a changing world, deep reinforcement learning has so far found few commercial applications. Research and development is costly, but DeepMind's losses of 154 million dollars in 2016, 341 million in 2017, and 572 million in 2018 are hardly a sign of success (Marcus, 2019).

The latest hype: Big Data

The challenge of neural networks is that they must be able to handle huge amounts of data. For example AlphaGo was first trained on 150,000 games played by competent Go players. Then it was improved by repeatedly playing against earlier versions of itself.

Computers' increasing ability to process and store huge amounts of data has led to what is called the "data explosion", or even "data deluge". Already in 2012 it was estimated that Google processed around 24 petabytes (24×10^{15}) of data every day. This is thousands of times the amount of printed material in the US Library of Congress (Mayer-Schönberger and Cukier, 2014, p. 8). At the same time it was estimated that 2.5 exabytes (2.5×10^{18} bytes) were created in the world per day. This is estimated to be approximately half of all the words ever spoken by humans. This amount of data is beyond human imagination, and it is the background for the Big Data approach.

Although Big Data analysis may be regarded as a supplemental method for data analysis for large amounts of data, typically terabytes and petabytes, it is sometimes presented as a new epistemological approach. Viktor Mayer-Schönberger and Kenneth Cukier start their book *Big Data* with the example of a flu that was discovered in 2009. It combined elements from viruses that caused bird flu and swine flu, and was given the name H1N1. It spread quickly, and within a week public health agencies around the world feared a pandemic. Some even feared a pandemic of the same size as the 1918 Spanish flu that killed millions. There was no vaccine against the virus, and the only thing the health authorities could do was to try to slow it down. But to be able to do that, they had to know where it had already spread. Although doctors were requested to inform about new cases, this information would take 1–2 weeks to reach the authorities, primarily because most patients do not consult a doctor immediately after the appearance of the symptoms of the disease.

However, researchers at Google had just before this outbreak invented a method that could much better predict the spread of the flu. Google receives more than three billion search queries

every day, and save them all. People who have symptoms of flu tend to search the internet for information on flu. Therefore, by looking at search items that are highly correlated with flu, the researchers could map the spread of flu much quicker than the health authorities (Mayer-Schönberger and Cukier, 2014, p. 2).

Mayer-Schönberger and Cukier regard this a a success story. But this may be an example of what is sometimes called "the fallacy of initial success". In 2013 the model reported twice as many doctor visits for influenza-like illnesses as the Centers for Disease Control and Prevention, which is regarded as a reliable source of information. The initial version of the model had probably included seasonal data that were correlated with the flu, but were causally unrelated. Therefore, the model was part a flu detector and part a winter detector. Although the model has been updated, its performance has been far below the initial promises (Lazer et al., 2014; Shane, 2019, p. 171).

Correlations and causes

The previous examples just involved correlations. However, in the sciences and also in everyday life, we want to have causal relations. For example, one of the big questions of our time involves causal knowledge: Is the global warming that we observe caused by human activity (the release of greenhouse gases into the atmosphere), or is it just natural variations?

The nature of causal relationships has been discussed for centuries, in particular after David Hume criticized the old idea of a necessary relationship between cause and effect. According to Hume we have to be satisfied with the observation of regularities. His contemporary Immanuel Kant, on the contrary, argued that causal relationships are a prerequisite for the acquisition of knowledge. It is necessary that every effect has a cause.

However, instead of going into the philosophical discussion about causal relationships, which has continued until this day, it is more fruitful to see how we identify a causal relationship. The philosopher John Stuart Mill formulated some rules (he called them "canons") that enable us to identify causal relationships. His "second canon" which he also called "the method of difference" is the following:

> If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon (Mill, 1882, p. 483).

From this quotation we see that the distinguishing mark of a causal relationship is a 100% correlation between cause and effect. But most correlations are not causal. For example, there is a high positive correlation between gasoline prices and my age, but there is obviously no causal relationship between the two. A correlation may therefore be an indication of a causal link, but it need not be.

Therefore, in the quotation above, Mill requires that the two cases be equal in all circumstances. But still we can only decide that the difference between the two is either the cause or the effect, because correlation is a symmetrical mathematical relationship: If A is correlated with B, B is correlated with A. In contrast, if C is the cause of E, E is not the cause of C. Therefore, correlations cannot distinguish between cause and effect. To make this distinction we need something more: The cause produces, or at least brings about, the effect. Therefore, we may remove the assumed cause, and see if the effect disappears.

We have a famous example of this procedure from the history of medicine (more specifically epidemiology). Around 1850 there was a cholera epidemic in London. John Snow was a practicing physician. He noted that there was a connection between what company people got the water from and the frequency of cholera. The company Southwark and Vauxhall, which had water intake at a polluted site in the Thames, had a high frequency of cholera cases. Another company, the Lambeth Company, had significantly lower numbers. Although this was before the theory of bacteria as the cause of disease, he assumed that the cause of the disease was found in the water. Here are Snow's numbers:

Company	Deaths per 10,000 households
Southwark and Vauxhall	315
Lambeth Company	37
The rest of London	59

After Snow had sealed a water pump that he believed contained infectious water, the cholera epidemic ended (Sagan, 1996, p. 76).

If the effect always follows the cause, everything else equal, we have deterministic causality. However, many people smoke cigarettes without contracting cancer. The problem is that in practice some uncertainty is involved. Therefore, we need a definition of a causal relationship when we have <100% correlation between cause and effect. According to this definition a probabilistic cause is not always followed by the effect, but the frequency of the effect is higher than when the cause is not present. This can be written as $P(E|C) > P(E|\text{not-}C)$. $P(E|C)$ is a conditional probability, and can be read as "the probability of E, given C".

However, although this looks straightforward, it is not. An example will show this. After World War II there were many indications that cigarette smoking might cause lung cancer. It looks as if this question might be decided in a straightforward way: One selects two groups of people that are similar in all relevant aspects. One group starts smoking cigarettes and another does not. This is a simple randomized, clinical trial. Then one checks, after 10 years, 20 years, 30 years, and so on, and see if there is a difference in the frequency of lung cancer in the two groups.

Of course, if cigarette smoking is as dangerous as alleged, one would not wait decades to find out. Therefore, one had to use the population at hand, and use correlations: One took a sample of people with lung cancer and another sample of the population that did not have cancer and looked at different background factors: Is there a higher frequency of cigarette smokers among the people who have contracted lung cancer than people who have not contracted lung cancer. The main criterium is "ceteris paribus", everything else equal.

One thing is to acknowledge that we sometimes have to use correlations to find causal relations. It is quite another thing to argue that we do not need causes at all. Nevertheless, some argue that we can do without causal relationship. In 2008 the chief editor of *Wired Magazine*, Chris Anderson, wrote an article with the title: "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". In the article he argued that correlations are sufficient. We can use huge amount of data and let statistical algorithms find patterns that science cannot. He went even further, and argued that the traditional scientific method, of using hypotheses, causal models and tests, is becoming obsolete (Anderson, 2008).

According to Mayer-Schönberger and Cukier, Anderson's article unleashed a furious debate, "… even though Anderson quickly backpedaled away from his bolder claims" (Mayer-Schönberger and Cukier, 2014, p. 71). But even if Anderson modified his original claims, Mayer-Schönberger and Cukier

agree that in most cases we can do without knowing causal relations: "Big Data is about what, not why. We don't always need to know the cause of a phenomenon; rather, we can let data speak for itself" (Mayer-Schönberger and Cukier, 2014, p. 14). Later they formulate it in this way: "Causality won't be discarded, but it is being knocked off its pedestal as the primary fountain of meaning. Big data turbocharges non-causal analyses, often replacing causal investigations" (Mayer-Schönberger and Cukier, 2014, p. 68). Pearl and Mackenzie put it this way: "The hope—and at present, it is usually a silent one—is that the data themselves will guide us to the right answers whenever causal questions come up" (Pearl and Mackenzie, 2018, p. 16). I have to add that Pearl and Mackenzie are critical of this view.

The mini Turing test

Anderson was not the first to argue that science can do without causes. At the end of the 19th century one of the pioneers of modern statistics, Karl Pearson, argued that causes have no place in science (Pearl and Mackenzie, 2018, p. 67) and at the beginning of the 20th century one of the most influential philosophers of that century, Bertrand Russell, wrote the article "On the Notion of Cause" where he called "the law of causality" a "relic of a bygone age" (Russell, 1963, p. 132). For example, when bodies move under the mutual attraction of gravity, nothing can be called a cause, and nothing an effect according to Russell. There is "merely a formula" (Russell, 1963, p. 141). He might have added that Newton's mechanics had been reformulated by Joseph-Louis Lagrange and William Hamilton to an abstract theory without the concept of force.

However, Russell looked for causality at the wrong place. He took simply Newton's theory for granted, and had forgotten that Newton himself subscribed to what in his time was called "experimental philosophy". Physics is no doubt an experimental science, and to carry out experiments the physicist must be able to move around, to handle instruments, to read scales, and to communicate with other physicists. As the physicist Roger Newton has pointed out, a physicist "…effectively conducts experiments by jiggling one part of Nature and watching how other parts respond" (Newton 1997, p. 142). To find out if A causes B, it is important for "A to be *under our control*" (Newton, 1997, p. 144, italics in original).

I have already quoted Pearl's and Mackenzie's book *The Book of Why* (2018). The main argument in the book is that to create humanlike intelligence in a computer, the computer must be able to master causality. They ask the question:

> How can machines (and people) represent causal knowledge in a way that would enable them to access the necessary information swiftly, answer questions correctly, and do it with ease, as a three-year-old child can? (Pearl and Mackenzie, 2018, p. 37).

They call this the "mini-Turing test". It has the prefix "mini" because it is not a full Turing test, but is confined to causal relations.

Before I go into the mini-Turing test I will briefly recall the Turing test. In the article "Computing Machinery and Intelligence" (Turing, 1950). Alan Turing asked the question: How can we determine if computers have acquired general intelligence? He starts by saying that the question he tries to answer is: "Can machines think?", but instead of going into the question of what intelligence is, he sets up a kind of game. In the game a questioner can communicate with a computer and a human being. He has to communicate through a key-board, so he does not know who is the computer and who is the human. The point is that the machine pretends to be human, and it is the job of the questioner to decide which of the two is the computer and who is the human. If the questioner is unable to distinguish, we can say that the computer is intelligent. Turing called this the "imitation game", but it is later known as the "Turing test". If the computer passes the test, it has, according to Turing, acquired general intelligence.

According to Pearl and Mackenzie a minimum requirement to pass the Turing test is that the computer is able to handle causal questions. From an evolutionary perspective this makes sense. Why *Homo sapiens* has been so successful in the history of evolution is of course a complex question. Many factors have been involved, and the ability to cooperate is probably one of the most important. However, a decisive step took place between 70,000 and 30,000 years ago, what the historian Yuval Harari calls the Cognitive Revolution (Harari, 2014, p. 23). According to Harari the distinguishing mark of the Cognitive Revolution is the ability to imagine something that does not exist. Harari's example is the ivory figurine "the lion man" (or "the lioness woman") that was found in the Stadel Cave in Germany, and is approximately 32,000 years old. It consists of a human body and the head of a lion.

Pearl and Mackenzie refer to Harari, and add that the creation of the lion man is the precursor of philosophy, scientific discovery, and technological innovation. The fundamental precondition for this creation is the ability to ask and answer questions of the form: "What happens if I do ……?" (Pearl and Mackenzie, 2018, p. 2).

The mini-Turing test is restricted to causal relationships. If computers can handle causal knowledge, they will pass this test. However, the problem is that in this regard computers have not made any progress for decades: "Just as they did 30 years ago, machine-learning programs (including those with deep neural networks) operate almost entirely in an associative mode…" (Pearl and Mackenzie, 2018, p. 30). But this is insufficient. To answer causal questions we must be able to intervene in the world.

According to Pearl and Mackenzie the root of the problem is that computers do not have a model of reality. However, the problem is that nobody can have a model of reality. Any model can only depict simplified aspects of reality. The real problem is that computers are not in the world, because they are not embodied.

The real Turing test

Pearl and Mackenzie are right in arguing that computers cannot pass the mini-Turing test because they cannot answer causal question. And I shall argue that they cannot pass the full Turing test because they are not in the world, and, therefore, they have no understanding. A few examples from the Turing test will show this.

There is an annual competition where the award (Loebner Prize) is given to the AI program most similar to a human being. The competition is the Turing test, but gold or silver medals have never been awarded. The program Mitsuku won in 2013, 2016, 2017, 2018, and 2019.

The philosopher Luciano Floridi recounts how he joined the judging panel when the competition was first held in England in 2008. One of the judges started with the question: "If we take each other's hand, whose hand am I holding then?" The computer replied by talking about something completely different: "We live in eternity. So, yeah, no. We don't believe." Another question was: "The four capital cities of England are three, Manchester and Liverpool. What's wrong with this sentence?" The computer had no meaningful answer (Floridi, 2014, p. 133).

The conversation that Floridi refers to, took place more than 10 years ago. I was curious to see if the program had improved.

I have tried it out several times, and it has not improved much. Below is an example. When Mitsuku in 2008 was asked: "If we take each other's hand, whose hand am I holding then?" the computer failed. This time it came out with the right answer. However, it did not take long before it failed. In the printout below "Millie" can either be the Mitsuku computer, or a human. At the end of the conversation one is supposed to guess if one talked to a computer or to a human being. It was not difficult to determine that I had talked to a computer.

Here is the "conversation":

> Mitsuku chatbot Welcome Ragnar, now talking. * Topic is 'Turing Test' * Set by admin (1 day ago)
>
> <Ragnar> If I shake your hand, whose hand am I holding?
>
> <Millie> I will be holding yours adn you will be holding mine. What's with the dumb questions?
>
> <Ragnar> Why is the question dumb?
>
> <Millie> I want to know everything about you.
>
> <Ragnar> But you did not explain why you regarded the question as dumb.
>
> <Millie> Sorry, I must have forgot. What was your question again?
>
> <Ragnar> If we shake hands, whose hand am I holding?
>
> <Millie> We would be holding each other's hand.
>
> <Ragnar> And why is that a dumb question?
>
> <Millie> Why do I exist? (MitsukuHomepage. I have kept Mitsuku's misprinting)

Computers fail because they are not in the world. Mitsuku characterized the first question as dumb, but could not explain why. Any child would be able to do that.

However, the competition rules of the Loebner Prize have been changed. The aim of getting computers to pass the Turing test has been given up, because "… chatbots are typically so quickly unmasked that the prize was always one for 'best of'" (LoebnerPrize).

Conclusion: computers are not in the world

The main thesis of this paper is that we will not be able to realize AGI because computers are not in the world. However, it is crucial that we clarify what is meant by "world".

As the historian of science Alexandre Koyré has pointed out, the most important achievement of the scientific revolution of the 17th century was the replacement of Aristotelian science by an abstract scientific ideal ("paradigm") (Koyré 1978, pp. 38–39). Koyré argued convincingly that Galileo was basically a Platonist (Koyré, 1968). As in the case of Plato, the key was mathematics. According to Galileo the book of nature is written in the language of mathematics (Galilei, 1970, p. 237). Therefore, Galileo's world is an abstract and idealized world, close to Plato's world of ideas.

The system that comes closest to this ideal world is our solar system, what Isaac Newton called "the system of the world". Newton's mechanics became the model for all science. The best expression of this ideal was given by the French mathematician Pierre Simon de Laplace. He argued that there is in principle no difference between a planet and a molecule. If we had complete knowledge of the state of the universe at one time, we could in principle determine the state at any previous and successive time (Laplace, 1951, p. 6). This means that the universe as a whole can be described by an algorithm. Turing referred to this passage from Laplace in his article "Computing Machinery and Intelligence", and added that the predictions he (Turing) was considering, were nearer to practicability than the predictions considered by Laplace, which comprised the universe as a whole (Turing, 1950, p. 440).

As Russell pointed out, in this world we cannot even speak about causes, only mathematical functions. Because most empirical sciences are causal, they are far from this ideal world. The sciences that come closest, are classical mechanics and theoretical physics.

Although this ideal world is a metaphysical idea that has not been realized anywhere, it has had a tremendous historical impact. Most philosophers and scientists after Galileo and Descartes have taken it to be the real world, which implies that everything that happens, "at the bottom" is governed by mathematical laws, algorithms. This applies to the organic world as well. According to Descartes all organisms, including the human body, are automata. Today we would call them robots or computers. Descartes made an exception for the human soul, which is not a part of the material world, and therefore is not governed by laws of nature. The immaterial soul accounts for man's free will.

However, most advocates of AGI (and advocates of strong AI) will today exclude Descartes' immaterial soul, and follow the arguments of Yuval Harari. In his latest book *21 Lessons for the 21st Century* he refers to neuroscience and behavioral economics, which have allegedly shown that our decisions are not the result of "some mysterious free will", but the result of "millions of neurons calculating probabilities within a split second" (Harari, 2018, p. 20). Therefore, AI can do many things better than humans. He gives as examples driving a vehicle in a street full of pedestrians, lending money to strangers, and negotiating business deals. These jobs require the ability "to correctly assess the emotions and desires of other people." The justification is this:

> Yet if these emotions and desires are in fact no more than biochemical algorithms, there is no reason why computers cannot decipher these algorithms—and do so far better than any *Homo sapiens* (Harari, 2018, p. 21).

This quotation echoes the words used by Francis Crick. In *The Astonishing Hypothesis* he explains the title of the book in the following way:

> The Astonishing Hypothesis is that "You", your joys and your sorrows, your memories and your ambitions, your sense of personal identity and free will, are in fact no more than the behavior of a vast assembly of nerve cells and their associated molecules (Crick, 1994, p. 3).

However, there is a problem with both these quotations. If Harari and Crick are right, then the quotations are "nothing but" the result of chemical algorithms and "no more than" the behavior of a vast assembly of nerve cells. How can they then be true?

If we disregard the problem of self-reference, and take the ideal world of science that I have described above to be the (only) real world, then Harari's argument makes sense. But the replacement of our everyday world by the world of science is based on a fundamental misunderstanding. Edmund Husserl was one of the first who pointed this out, and attributed this misunderstanding to Galileo. According to Husserl, Galileo was "…at once a discoverer and a concealing genius" (Husserl, 1970, p. 52). Husserl called this misunderstanding "objectivism". Today a more common name is "scientism".

Contrary to this, Husserl insisted that the sciences are fundamentally a human endeavor. Even the most abstract theories are grounded in our everyday world, Husserl's "lifeworld". Husserl mentions Einstein's theory of relativity, and argues that it is dependent on "Michelson's experiments[3] and the corroborations of them by other researchers" (Husserl, 1970, p. 125). To carry out this kind of experiments, the scientists must be able to move around, to handle instruments, to read scales and to communicate with other scientists.

There is a much more credible account of how we are able to understand other people than the one given by Harari. As Hubert Dreyfus pointed out, we are bodily and social beings, living in a material and social world. To understand another person is not to look into the chemistry of that person's brain, not even into that person's "soul", but is rather to be in that person's "shoes". It is to understand the person's lifeworld.

The American author Theodore Roszak has constructed a thought example to illustrate this point: Let us imagine that we are watching a psychiatrist at work. He is a hard working and skilled psychiatrist and obviously has a very good practice. The waiting room is full of patients with a variety of emotional and mental disorders. Some are almost hysterical, some have strong suicidal thoughts, some hallucinations, some have the cruelest nightmares and some are driven to madness by the thought that they are being watched by people who will hurt them. The psychiatrist listens attentively to each patient and does his best to help them, but without much success. On the contrary, they all seem to be getting worse, despite the psychiatrist's heroic efforts.

Now Roszak asks us to put this into a larger context. The psychiatrist's office is in a building, and the building is in a place. This place is Buchenwald and the patients are prisoners in the concentration camp (Roszak, 1992, p. 221). Biochemical algorithms would not help us to understand the patients. What does help, in fact, what is imperative, is to know the larger *context*. The example simply does not make sense if we do not know that the psychiatrist's office is in a concentration camp.

Only few of us are able to put ourselves in the shoes of a prisoner of a concentration camp. Therefore, we cannot fully understand people in situations that are very different from what we have ourselves experienced. But to some degree we *can* understand, and we can understand because we are also in the world.

Computers are not in our world. I have earlier said that neural networks need not be programmed, and therefore can handle tacit knowledge. However, it is simply not true, as some of the advocates of Big Data argue, that the data "speak for themselves". Normally, the data used are related to one or more models, they are selected by humans, and in the end they consist of numbers.

If we think, for example like Harari, that the world is "at the bottom" governed by algorithms, then we will have a tendency to overestimate the power of AI and underestimate human accomplishments. The expression "nothing but" that appears in the quotation from Harari may lead to a serious oversimplification in the description of human and social phenomena. I think this is at least a part of the explanation of the failure of both IBM Watson Health and Alphabet's DeepMind. "IBM has encountered a fundamental mismatch between the way machines learn and the way doctors work" (Strickland, 2019) and DeepMind has discovered that "what works for Go may not work for the challenging problems that DeepMind aspires to solve with AI, like cancer and clean energy" (Marcus, 2019).

The overestimation of the power of AI may also have detrimental effects on science. In their frequently quoted book *The Second Machine Age* Erik Brynjolfson and Andrew McAfee argue that digitization can help us to understand the past. They refer to a project that analyzed more than five million books published in English since 1800. Some of the results from the project was that "the number of words in English has increased by more than 70% between 1950 and 2000, that fame now comes to people more quickly than in the past but also fades faster, and that in the 20th century interest in evolution was declining until Watson and Crick discovered the structure of DNA." This allegedly leads to "better understanding and prediction—in other words, of better science—via digitization" (Brynjolfson and McAfee, 2014, p. 69). In my opinion it is rather an illustration of Karl Popper's insight: "Too many dollars may chase too few ideas" (Popper, 1981, p. 96).

My conclusion is very simple: Hubert Dreyfus' arguments against general AI are still valid.

Received: 19 January 2020; Accepted: 7 May 2020;
Published online: 17 June 2020

Notes

1 Polanyi normally uses "knowing" instead of "knowledge" to emphasize the personal dimension. However, I will use the more traditional "knowledge".
2 The example is taken from the Wikipedia article on Jeopardy! (Wikipedia: Jeopardy).
3 I have given a detailed description of Michelson's instruments in Fjelland (1991).

References

Anderson C (2008) The end of theory: the data deluge makes the scientific method obsolete. Wired Magazine
Bolter D (1986) Turing's man. Western culture in the computer age. Penguin Books
Brynjolfson E, McAfee A (2014) The second machine age. Norton & Company
Chen X-W, Lin X (2014) Big Data deep learning: challenges and perspectives. IEEE Access 2:514–525
Crick F (1994). The astonishing hypothesis. The scientific search for the soul. Macmillan Publishing Company
Dingli A (2018) "Its Magic….I Owe You No Explanation!" 2018. https://becominghuman.ai/its-magic-i-owe-you-no-explanation-explainableai-43e798273a08. Accessed 5 May 2020
Dreyfus HL (1972) What computers can't do. Harper & Row, New York, NY
Dreyfus HL, Dreyfus SE (1986). Mind over machine. Basil Blackwell
Fjelland R (1991) The Theory-Ladenness of observations, the role of scientific instruments, and the Kantian a priori. Int Stud Philos Sci 5(3):269–80
Floridi L (2014) The 4th revolution. How the infosphere is reshaping human reality. Oxford University Press, Oxford
Ford M (2015) The rise of the robots. technology and the threat of mass unemployment. Oneworld Publications, London
Galilei G (1970) Dialogue concerning the two chief world systems (1630). University of California Press, Berkeley, Los Angeles
Harari YN (2014). Sapiens. A brief history of humankind. Vintage
Harari YN (2018) 21 Lessons for the 21st century. Jonathan Cape, London
Husserl E (1970) The crisis of european sciences and transcendental phenomenology. Northwestern University Press, Evanston
Jones N (2014) The learning machines. Nature 505:146–48
Koyré A (1968) Galileo and Plato. In: Metaphysics and measurement (1943). John Hopkins Press, Baltimore, London.
Koyré A (1978) Galileo studies (1939). Harvester, London
Lanier J (2013) Who owns the future? Allen Lane, London
Laplace PS (1951) Philosophical essay on probabilities (1814). Dover Publications, New York
Lazer D, Kennnedy R, King G, Vespignani A (2014) The parable of google flu: traps in big data analysis. Science 343:1203–1205
LoebnerPrize. https://artistdetective.wordpress.com/2019/09/21/loebner-prize-2019. Accessed 5 May 2020
Marcus G (2019) DeepMind's losses and the future of artificial intelligence. Wired 14.8.2019. https://www.wired.com/story/deepminds-losses-future-artificial-intelligence/. Accessed 6 Jan 2020
Mayer-Schönberger V, Cukier K (2014) Big data: a revolution that will transform how we live, work, and think. Eamon Dolan/Mariner Books
Mill JS (1882) A system of logic, ratiocinative and inductive, being a connected view of the principles of evidence, and the methods of scientific investigation, 8th edn. Harper & Brothers, New York
Minsky M (1972) Computation: finite and infinite machines. Prentice-Hall International

MitsukuHomepage. http://www.square-bear.co.uk/mitsuku/home.htm. Accessed 12 Sept 2017
Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E (2015) Deep learning applications and challenges in big data analytics. J Big Data 2(1). https://doi.org/10.1186/s40537-014-0007-7
Newton RG (1997) The truth of science. Physical theories and reality. Harvard University Press, Cambridge
Nielsen M (2016). Is alphago really such a big deal? Quanta Magazine, March 29, 2016. https://www.quantamagazine.org/is-alphago-really-such-a-big-deal-20160329/. Accessed 7 Jan 2020
Pearl J, Mackenzie D (2018) The book of why. The new science of cause and effect. Basic Books, New York
Penrose R (1989) The emperor's new mind. Concerning computers, minds, and the laws of physics. Oxford University Press, Oxford
Penrose R (1994) Shadows of the mind. A search for the missing science of consciousness. Oxford University Press, Oxford
Plato (1955) The republic. Penguin Books, Harmondsworth, p. 1955
Polanyi M (1958) Personal knowledge. Routledge & Kegan Paul, London
Polanyi M (2009) The tacit dimension (1966). The University of Chicago Press, Chicago
Popper KR (1981) The rationality of scientific revolutions (1975). In: Hacking Ian ed Scientific revolutions. Oxford University Press, Oxford, pp. 80–106
Roszak T (1992) The voice of the earth. Simon & Shuster, New York
Russell B (1963). On the notion of a cause (1912). In: Mysticism and logic. Unwin Books, London
Sagan L (1996) Electric and magnetic fields: invisible risks? Gordon and Breach Publishers, Amsterdam
Searle J (1980) Minds, brains, and programs. Behav Brain Sci 3(3):417–57
Shane J (2019) You look like a thing and I love you. Wildfire, London
Strickland E (2019) How Ibm Watson overpromised and underdelivered on Ai Health Care. IEEE Spectrum, 2 April 2019. https://sprectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care. Accessed 5 Jan 2020
Susskind R, Susskind D (2015) The future of the professions. Oxford University Press, Oxford
Tegmark M (2017) Life 3.0. Being human in the age of artificial intelligence. Alfred A. Knopf, New York
Turing A (1950) Computing machinery and intelligence. Mind LIX 236:433–60
Weizenbaum J (1976) Computer power and human reason. Freeman & Company, San Francisco
Wikipedia: Jeopardy!, https://en.wikipedia.org/wiki/Jeopardy! Accessed 2 Feb 2017

Acknowledgements

I want to thank the participants of the workshop *Ethics of Quantification*, University of Bergen 5.12.2012, and Adam Standring and Rune Vabø, for useful comments.

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.F.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.