9 The Advent of Machine Knowledge beyond our Human Comprehension

Matthias GELBMANN
Q-Success, Austria

9.1 Abstract

Artificial intelligence has made significant progress in recent years in the form of advanced machine learning thait is able to produce knowledge that not only exceeds human knowledge in certain areas but is also based on complex computations that humans are no longer able to understand. While considerably increasing the knowledge accessible to us, this new situation also has severe implications and creates new risks: should we trust systems that nobody understands? What kind of decisions can we delegate to machines when we cannot comprehend why the decisions were made? The paper discusses options to respond to these challenges, as well as likely further developments in this area.

Keywords: artificial intelligence, machine learning

9.2 Introduction

Machine knowledge is the knowledge acquired by machines by means of machine learning. According to a common definition, machine learning is a range of AI methods to automatically learn from data without being explicitly programmed.
If we compare today's machine knowledge and human knowledge, then the amount of human knowledge exceeds machine knowledge by many orders of magnitude. Nevertheless, recent developments in AI have brought us specific, comparatively tiny areas of machine knowledge that is not included in human knowledge. Humans have long ago started to develop tools that are more capable in some specific aspects than their own body. Now we have created tools that are, in some specific aspects, more knowledgeable than our brains.

9.3 Stages of AI evolution

In order to see why we are entering a new era now, we need to have a look at how AI evolved in the last 60 years or so and compare earlier stages of AI to this new quality of the last years.
For AI to work we need two components: we need knowledge and we need the ability to apply knowledge. Knowledge, as represented for example in an encyclopedia, is a passive component, and is not sufficient to build an AI system. We need an active component in addition that uses the knowledge to achieve something useful. Wikipedia by itself is not AI, but a system that uses Wikipedia to answer questions is.
The evolution of AI can be shown by examining the evolution of these two components.

9.3.1 Stage 1 AI - basic AI

AI naturally started with low levels of knowledge and simple methods to apply it. An example of such early-stage AI systems are expert systems, the AI hype of the 1980s. Expert systems represent knowledge in sets of *if-then* rules. A so-called inference engine then matches the input data with the conditions in the if-part and, in case of a match, executes the then-part of the rules. The output of all the then-parts of an inference cycle represents the answer of the system to its inputs.
All the rules of an expert system are written by human experts, therefore the knowledge represented in these systems never exceeds human knowledge. However, an extremely small

part of human knowledge is accessible to these systems. It is very easy for expert systems to "explain" how they come to results, just by listing the rules that were used to come to a result. "A computer can never be smarter than the people who programmed it" - this was true for Stage 1 AI, but not for later stages.

9.3.2 Stage 2 AI - exploitation of processing power

Stage 2 AI still uses a relatively low level of knowledge but increases the ability to use the knowledge by using algorithms that can better exploit the processing power of a computer.
A typical case of Stage 2 AI is a classical chess program. Chess programs can evaluate millions of chess positions per second and use algorithms such as minimax and alpha-beta pruning to search the game tree and to find a good move. Chess programs get better the more processing power is used, and since 1997 they play better than the best humans. It is important to note that the chess knowledge of these programs does not exceed the human chess knowledge, but the ability to apply that knowledge millions of times per second enable them to beat human players [1].
Stage 2 AI can exceed human intelligence in their area of expertise. It is also possible to acquire explanations for the results. In the example of the chess program, we can compare the evaluation of the relevant chess positions and see why one is evaluated better than all the others.

9.3.3 Stage 3 AI - automatic creation of new knowledge

Stage 3 AI increases the level of the available knowledge, typically by using machine learning. Machine learning is not new, but a remarkable breakthrough has been made in the last years by developing methods to successfully implement deep learning.
Deep learning uses the concept of Artificial Neural Networks. Neural networks mimic the way neurons work in the human brain. They are good at pattern recognition, be it visual patterns or patterns in more abstract data sets. They have been used for decades, but their use was limited to relatively simple applications such as character recognition.
Deep learning uses large neural networks: hundreds of layers with hundreds of thousands of neurons and millions of connections between the neurons. Much processing power is needed to train such networks, but just as important are specialized algorithms that were only developed in the last decade, and first introduced by Alex Krizhevsky et al in 2012 [2].
Deep learning networks are still tiny compared to the estimated 90 billion neurons in a human brain, and also even the small part of the inner workings of a human brain that we understand is far more complex that the simple calculations used in a neural network. Nevertheless, deep learning can lead to results that are comparable to, and in some cases exceed, human capabilities in specific problem areas.
Although the mathematical foundation of neural networks is relatively simple, the sheer amount of data and calculations in a deep learning network makes it impossible for humans to understand exactly why they come to conclusions from any given input. It is easy to see how each individual neuron works, but impossible to grasp the "big picture" in such a large network. There is also no way for deep learning networks to explain to a human how it works, because the only explanation would be the data flow through its network.
While Stage 2 AI can be better than humans due to sheer processing power, Stage 3 AI systems are the first man-made systems that have knowledge beyond our human comprehension. We can expect that development to continue to a point where machine knowledge not only covers a large part of human knowledge, but also exceeds the amount of human knowledge, see figure 1.

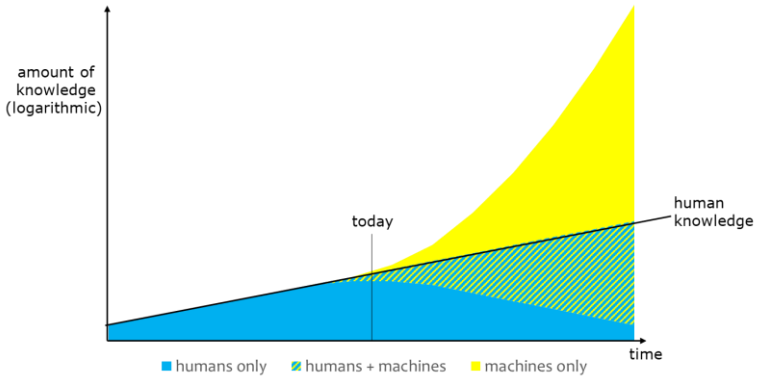

***Figure 1**: human knowledge vs machine knowledge*

New applications of deep learning networks are published daily. Some examples are

- Image recognition, object identification (identifying for example a cat or a dog in a picture), face recognition, mood recognition in faces.
- Speech recognition, including speaker recognition, language translation.
- Stock market trading, where stocks picked by AI often perform better than stocks picked by human traders.
- Game playing systems are among the most published applications. AlfaGo won against the best human Go player in 2016 using deep learning to evaluate Go positions [3]. A similar program, AlfaZero won against the best chess program in 2017 after training its deep learning network for four hours only by playing against itself and without feeding in any human chess knowledge [4]. This is an example of a Stage 3 AI outperforming the best Stage 2 AI due to its superior knowledge.
- Perhaps the most impressive examples are in the medical diagnostic field. AI systems can detect Alzheimer's earlier than doctors, which is very helpful to start treatment at a very early state [5]. AI systems can identify skin cancer from pictures as good or better than the best dermatologists [6]. Another impressive example is a system called Deep Patient. The Mount Sinai Hospital in New York trained its deep network using medical records from about 700,000 individuals, and it was better at predicting disease from a patient's data than doctors. It surprised its makers by also predicting psychiatric disorders, which was not in the original scope of the project, and which doctors do not understand how it can be derived from the available data [7].

9.3.4 Stage 4 AI - automatic creation of new abilities to use knowledge

Stage 3 AI is state-of-the-art today. The further stages described here are assumptions on how AI will evolve in the future.

If we look at the example of the chess-playing AlfaZero, saying that it achieved its performance by applying deep learning is only part of the story. AlfaZero used deep learning essentially to evaluate positions, but the rest of the code implementing the Monte Carlo Tree Search algorithm was hand-coded and hand-optimized code. The deep learning network was a central part, but by itself it could not win a single chess game. Similarly to how machine learning already finds patterns in data without being explicitly programmed, we can expect future AI to find algorithms without being explicitly programmed. This is what Stage 4 AI will provide, thus primarily improving the ability to use knowledge.

Genetic Programming (GP) is one of the techniques available today that are able to automatically generate algorithms. GP is not a new technique but, just as Artificial Neural Networks were used for decades before deep learning brought the impressive results we see today, it would take a similar breakthrough for GP or some similar techniques to advance AI to Stage 4. We will then have machine-generated algorithms that consist of thousands of steps, so that humans are no longer able to understand what they do. We will not be able to see the

“big picture” due to the size of the code. We will then have the ability to apply knowledge in a way that is beyond our human comprehension.

9.3.5 Stage 5 AI - use AI to create or improve AI

The next logical step will be to use the Stage 3 knowledge and the Stage 4 algorithms not only to play chess or to predict Alzheimer’s, but to create and improve AI systems. We do not know what Stage 5 AI systems will look like, but because the procedures to create and to improve them will be automated to a certain extent, we can expect the speed of AI performance improvement to be significantly higher than it is today.
Figure 2 summarizes the 5 stages of the AI evolution.

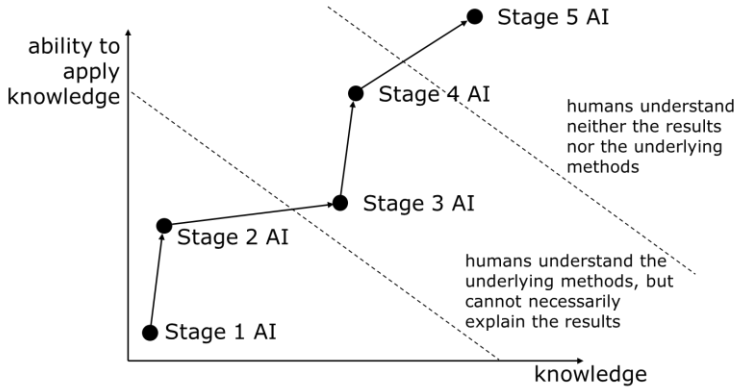

***Figure 2**: the 5 AI stages shown along the two axes of the passive and active component.*

9.4 Dealing with incomprehensible systems

Interacting with systems that we cannot understand leaves some discomfort in most of us. What are our options to deal with that situation?

9.4.1 Creation of technical alternatives

The problem resolves itself if it is possible to somehow enhance the existing systems to provide an explanation. There is a research branch Explainable AI trying to do just that, and any results towards that goal are very valuable. The problem is that in general that is not possible without making compromises. The incomprehensibility of deep learning networks has its roots in the amount of data and of calculations involved, but this is on the other hand exactly what makes them work so well. Any restrictions, such as cutting down or simplifying the connections of the network degrade its performance. In other words, asking for an explanation, or more precisely for an explanation so simple that humans can understand it, deteriorates the system's results.

9.4.2 Ban the development of Stage 3 AI

One extreme proposal would be to ban the development of AI of Stage 3 or higher. The problem with that proposal is that it is practically impossible to enforce. AI systems can be built without anyone noticing it. All it takes is a few talented engineers and some processing power. Plenty of the know-how and of the software tools required to build such systems are freely available.

9.4.3 Ban the use of Stage 3 AI

Monitoring the use of certain AI systems may be easier than monitoring its development, so we could at least ban its use if that makes sense. However, what would be a good reason for a ban? The fact that some people feel uncomfortable seems to be too weak, at least in a liberal society, and particularly when such systems provide benefits that are not achievable in any other way.

Dealing with items that nobody fully understands is not entirely new. For example, the exact mechanisms of how some of the drugs we use can cure diseases may not be fully understood, but by carefully measuring the effects and side effects, we are still confident enough to allow them to be used. We should apply a similar attitude towards AI systems.

9.4.4 Unrestricted use of Stage 3 AI

If we are cautious enough with AI systems that we do not understand, do we need any restrictions at all?
That answer is probably yes. In some circumstances we require an explanation for decisions. For example, were a judge in a court case to rely on an AI that nobody understands to find his verdict, it would be unsatisfactory and could subsequently undermine a country's juridical system. The reasoning for the judge's decisions provides us with the means to find flaws in his judgement and thus a reference point to dismiss a verdict if appropriate.
The European Union in its General Data Protection Regulation (GDPR) enforces a "right to explanation" in some circumstances. The examples listed in that law are decisions on accepting or refusing a credit application and job applications.

9.4.5 Define restrictions for the use of Stage 3 AI

There is one fundamentally new aspect when comparing incomprehensible AI systems with other systems that we do not fully understand such as drugs: AI systems can control things, and with that ability new risks are introduced. A system that does not directly control anything, but only gives recommendations obviously introduces much lower risks than autonomous systems. Furthermore, the potential impact of a failure must be taken into account.
A well-tested medical diagnostic AI that only presents results and leaves the final decision to doctors and patients has little potentially damaging impact. A self-driving car has the potential to kill people. Obviously, we need to take measures to limit that risk. On the other hand, cars driven by humans are probably an even bigger risk, therefore allowing well-tested self-driven cars on our road makes sense.
Figure 3 shows some examples of systems with different AI capabilities and different levels of controls is a diagram indication the corresponding risk levels.

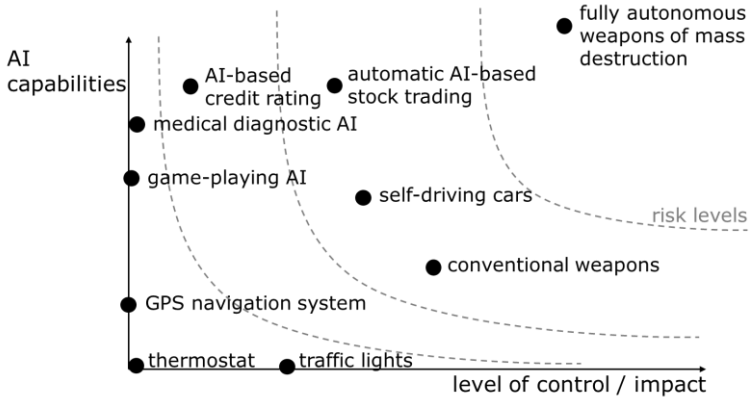

***Figure 3**: AI capabilities vs control levels*

9.5 Conclusions

Humanity faces a new challenge with AI systems beyond our comprehension. The case presented in this paper is that this in itself is not a problem. We need to find regulations that consider the level of risks when we pass control to AI systems and assess the use of Stage 3 AI systems on a case by case basis. Future Stage 5 AI systems will bring new challenges not only due to the much wider range of potential applications, but also due to the speed of developments. It will furthermore intensify ethical and political questions that are beyond the

scope of this paper. Coping with that situation from a regulatory point of view will prove immensely challenging.

9.6 References

[1] Murray Campbell, A. Joseph Hoane Jr., Feng-hsiung Hsu, “Deep Blue”, Artificial Intelligence, Vol 134, Issue 1-2, Jan 2002

[2] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks”, Communications of the ACM, Vol 60, Issue 6, Jun 2012

[3] David Silver et al., “Mastering the game of Go with deep neural networks and tree search“, Nature Vol. 529, 28 Jan 2016

[4] David Silver et al., “Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm”, arXiv:1712.01815v1 [cs.AI] 5 Dec 2017

[5] Mathotaarachchi S. et al, “Identifying incipient dementia individuals using machine learning and amyloid imaging”, Neurobiol Aging. Vol 59, Nov 2017

[6] Andre Esteva et al, “Dermatologist-level classification of skin cancer with deep neural networks”, Nature, Vol 542, pages 115–118, Feb 2017

[7] Riccardo Miotto, Li Li, Brian A. Kidd, Joel T. Dudley, “Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records”, Scientific Reports, Vol 6, Article 26094, May 2016