To Better Grasp Human Nature

Dr. Aleksandra Przegalinska explains why we find humanoid robots so creepy and considers whether watching machines play football is actually fun.

ACADEMIA: **What exactly is artificial intelligence?**

ALEKSANDRA PRZEGALINSKA: That's a good question, as it is still up for debate. Recently, Dominika Maison published a report after having asked many experts to define AI – all the answers began with "it's hard to say," so there is something to it. I can tell you how I define AI. For me, it's a discipline of knowledge, a field of R&D, that has been developing dynamically (albeit with periods of stagnation) since the 1940s. It is an interdisciplinary umbrella, embracing at least several other sub-fields. Generally speaking, AI could be defined as a field in which we try to create artificial systems that can simulate human behavior, such as exhibiting intelligence, adapting to changing conditions, and operating effectively in a changing world.

Some say that AI is about simulating human mental functions. In my opinion, that's not the case. In a simulation, we pretend like there is something that in reality is not there. When AI first came about, its first creator, a man named John McCarthy, said that AI is when machines perform an activity that we would consider to demonstrate intelligence if it were performed by a human. If a system does the job as well as a human, that's considered AI. This is a behavioral approach. I think that these days we're not looking to pretend that something works like a human, but to actually try to recreate certain cognitive, mental functions of humans or intelligent living systems in artificial environments so that they can also act, adapt, and solve problems.

These are the general goals and assumptions of this field. Under that umbrella, the most important subfields are machine learning, or the ability to process huge volumes of data, followed by machine vision, where an image is processed and the system is able to identify what is happening around it, such as seeing objects and people. The third most important subfield is natural language processing. This involves textual data being processed by various methods, but the main idea is to synthesize human speech so that it is understood and reproduced. In short, the goal is for these systems to work well operating not only with formal language, but with natural languages as well, such as Polish, English, or German. In addition to these main sub-fields, AI is also involved in some robotics, especially sensor technology, where a system adapts to its environment. Boston Dynamics [a leading American company specializing in engineering and robotics – editor's note] creates systems that are able to deal with a changing environment, such as overcoming obstacles and standing stably on various surfaces. This is the part of robotics that is considered to involve artificial intelligence.

It's interesting that you talk more about the capacity to adapt to the environment, not necessarily about simulating human traits.

Simulation is also important. For example, genetic algorithms simulate how a specific population functions. Through such simulation we can demonstrate something that would be difficult to demonstrate in the physical world. It would be very difficult to create a real neural network able to process data efficiently, but such a network can be implemented as a computer simulation. But this is not simulation in the sense of pretending or cheating, but rather doing something in a virtual space when it cannot be done in physical space. That is different. When I talk about simulation in a negative sense, I mean a simple system that

pretends to be something, but does so by means of simplistic tricks. This often happens in robotics, and for me this is not artificial intelligence, but simply pretending.

So, non-simplistic simulation of human behavior would involve AI being able not only to analyze large amounts of data, but also to make autonomous decisions, for example?

A good way to explain this difference is speech. The point is not to have a system capable of imitating the human voice, but of generating it, learning to speak. The point is to have a system that understands what we are saying at the semantic level, meaning it is able to map the meanings of words. It cannot simply be a box that processes phrases, tasked with forming a question from every declarative sentence. Such early bots did not understand content in language. They simply processed phrases, though they did so quite convincingly. Whatever was typed in, the system would respond like a psychotherapist, turning the statement into a question. So if you typed in "I have a problem," the system would ask: "You have a problem?" A therapist might do the same thing, but this does not involve understanding human statements or their content, just simple algorithmic operations. There are also many tricks to make it look like a system is processing language, understanding the content, synthesizing human speech, when in fact that is not the case at all. This type of simulation was once legitimate, but these days it is no longer useful because we can do it in a much more comprehensive, sophisticated, and complex way.

Is it possible for an AI to be empathetic during a conversation?

Not in terms of it actually feeling, but we can give it a sense of empathy. It is interesting that you ask about this, because DeepMind, which is a subsidiary of Google (the same company that created AlphaGo Zero, which defeated a human in a game of go) announced that they would create something called ToMnet – the Theory of Mind Network. This is to be a network that develops the ability to attribute mental states to minds. When we enter into various interactions with each other, we are able to understand that the other person, sending different messages, is experiencing certain emotional states, that there are intentions behind his or her words, etc. In philosophy, this is known as the "theory of mind." However, I always thought

Dr. Aleksandra Przegalinska (PhD)
works in the fields of philosophy and artificial intelligence. She currently works at Kozminski University in Warsaw, and in 2020 she will join the Labor & Worklife Program at Harvard Law School. She is the co-initiator of Poland's first program of study on artificial intelligence in management.
aprzegalinska@kozminski.edu.pl

that this was a purely philosophical concept, never thinking anyone would seriously attempt it in practice. DeepMind has said that it wants to build a network that will develop a theory of mind. So to answer your question: this AI will not feel anything or be aware of anything. It will be a neural network that will process non-verbal messages and therefore have an increased ability to process the context of the speaker's message. So it may understand that I am nervous or calm, that I am impatient or am expecting something by saying certain words. I don't know exactly how it will work, because DeepMind isn't revealing much. The intention is to get as close as possible to the human style of communication, which assumes that if conversation is involved, there is something beyond the layer of verbal communication and the meanings of words.

What is the Holy Grail for those working with AI, the ultimate success? What do researchers dream about?

That would of course be machine consciousness. Because consciousness is something we don't understand, even in ourselves. But I don't know if I truly want this Holy Grail, to be honest. We can't understand our own mind, figure out how thoughts are formed, how the brain works. The methods we have today, like neuroimaging the brain itself, are not very advanced. The mind is part of the physical brain that processes thoughts, and it works a little differently in everyone. There are many unsolved questions in this area. Consciousness itself is a vague concept. It assumes that various living things are conscious on some level, with humans certainly having a high level of consciousness. They also have self-awareness, resourcefulness, they can refer to themselves, including across time, they have integrity of identity across time, all of which certainly somehow interacts with the consciousness. These are all issues that have been identified but not yet clearly understood. Neuroscience has no answers to them right now. One may wonder what the greatest goal of robotics and AI is. If you ask a businessman, he would like to use AI in processes to improve efficiency. But if you ask a scientist, he would say that artificial intelligence should give us insight into human reasoning and thinking. For both neuroscience and the philosophy of mind, the goal of AI is to help us to understand consciousness, and even recreate it, and this is the Holy Grail. I think AI is a beautiful and wonderful field. Of course, it can be misused, but it is very noble in itself. It is an attempt to understand the mechanisms of the biological world, to recreate them in some way, an attempt to understand ourselves. Its objective is to understand the most important layer of humanity, the mental layer. That does not mean that I would like AI to have awareness. Fortunately, I don't think that will be possible in the near future.

> For both neuroscience and the philosophy of mind, the goal of AI is to understand consciousness, or even recreate it, and this is the Holy Grail.

That would indeed be very scary. So replacing humans is not the goal of robotics?

Absolutely not. The purpose of robotics is to partially understand human psychomotor functions and recreate them well. For example, the field of bionics deals with the creation of artificial limbs able to interact with the human body. Their goal is not to replace people at all, but to better understand them. And if it does involve replacing, then only in those functions that humans don't want to perform. Why should people in the 21st century carry heavy packages and break their backs if we are clearly not adapted to do so? It's better to create a robot, such as Atlas by Boston Dynamics, that will carry packages for us wherever we want. So it is meant to be used to perform certain activities that are difficult for humans. The human ability to think logically is much greater. Machines have their domains and we have ours. Artificial intelligence excels at operating with formal language and processing huge amounts of data. The same applies to boring and difficult tasks, such as continuously filling out the same forms. This is also something that AI can take over for humans. So these days the main goal is not to try to replace humans, but rather to make our lives more comfortable and identify areas where AI systems can support natural intelligence.

A recent issue of *The Economist* carried an article saying that soon patients in clinics will be seen by robots, not doctors. The headline is controversial, but the article itself talks about people being replaced in areas where it is easy to make mistakes, where typical human errors unfortunately occur.

Such automation is a bit of a far-fetched story. There's a certain fear factor, but such systems will take much longer to develop than it seems to those analysts, who are jumping the gun here. Yesterday I was at a conference in Berlin, where there were robots playing football. It was a really pathetic spectacle. Some ac-

tivities, such as wallpapering a room, are difficult for AI systems. Others are easy, such as processing huge amounts of financial or medical data. AI can undoubtedly help us detect rare diseases that we are not yet able to diagnose. Here, algorithms have great power. But to say that some jobs will be completely replaced by AI is an exaggeration. Especially since a particular job or profession involves many different activities. A job requires many different skills, often involves contact with people, communication. For the most part, we are not talking about assembly-line types of jobs, although they are still around and, in my opinion, will be replaced by robots, but most professions require many levels of different types of intelligence, including social and computational. So in this sense, people who talk about humans being replaced by machines don't know what they're talking about.

So what are some realistic threats posed by AI?

Market abuse in the financial world, hacking, malicious bots, bot traffic, things like election-manipulation by unleashing an army of bots into a network to spread false information, etc. There are a number of threats, but they are not related to automation or AI becoming conscious beings and destroying humanity. These threats lie elsewhere, such as when AI is used by a person for purposes other than those intended. I see problems when it comes to autonomous weapons. But the thought that AI is something that is spreading through our world, leaving people vulnerable and taking away their jobs, is a gross exaggeration.

Your work also focuses on the relationship between humans and technology. Why do people fear that machines are taking over the world?

There are several reasons for this fear. I think one is because people are afraid of machines that remind them of themselves. An AI system that looks like a robot is much less scary than one that looks like a human. Just think of Sophia the robot – such anthropomorphic systems that exhibit human features, generally not very successfully, instill a lot of anxiety in people, because people don't know how to classify them. They don't know what it is and they have to confront it. They ask themselves whether it is alive or dead. And this causes much confusion. So such systems are not very popular these days. Of course there are a lot of other problems. Many people in the United States have bought the Alexa system – robots that can help you run your entire household, turn off the lights, or even shop for you. They operate on the basis of voice commands, and so people are worried that Alexa may be eavesdropping on them. The western world is sensitive to privacy issues, which is why some people may not like these types of robots. Another issue is the fear of whether the system will take into consideration the type of information that is usually considered by humans. For example, if we have an AI system that calculates whether I qualify for a loan, I wonder if it might take only certain data into account while ignoring other information, making a decision that bank employees will follow blindly and won't allow me to challenge later. Perhaps I just need someone to believe in me. Artificial intelligence will not do so, as it does not feel jealousy, trust, or hope. It is a fully rational system and many people are also afraid of that: they will not have the chance to appeal what others take to be its expert decisions, and so the individual will lose out here.

One last question related to your academic background. You have worked at MIT, the world's top technical university. Can you give us some insight as to what working there is like? How does the quality of research work differ compared to Poland?

The first but key issue is a strong and very healthy collaboration with the business world. AI is a very popular tool in business, but MIT works in various fields. This is not business-specific work, it is more about solving fundamental questions that may later be useful. So it involves testing ideas in terms of practicality, but at the same time there is total intellectual freedom. This is not a scientific corporation that takes orders from the business world. Contracts with companies are worded in such a way as to protect the independence of the university, so this collaboration makes sense. The business world provides the data and financing. I am very impressed by this approach and I think it is very healthy. You don't work there solely for your own benefit, but in order to help make discoveries. Another thing is commitment to work. People who work there are highly motivated. They're not worn out, reluctant to work, or forced to do their jobs. On the contrary, these are often people who have quit their corporate jobs where they may have been paid better but are now fascinated by what they do. This commitment spreads to the entire team. Teamwork is taken very seriously there. There is a lot of cooperation, debriefing, sharing, challenging each other, and delving into each other's work.

We try to replicate that at Kozminski University in Warsaw, because this is a very effective way of working. Everyone strives towards helping each other become better. Of course, access to hardware, etc., is also a huge difference, but this is secondary. Working at MIT means teamwork, working towards a common goal, and looking into the future. There is much focus on working together, as well as on scientific optimism, the belief that we will succeed. On top of that we are always asking the question: Will society benefit from this?

Interview by Dr. **Justyna Orłowska**