Quality info

Volume 18, Issue 19 *Fortnightly, Free soft copy* **1st January 2017**

Understanding the Four Types of Artificial Intelligence... ...and learning to overcome the boundaries between the machines and us

The common and recurring view of the latest breakthroughs in artificial intelligence (AI) research is that sentient and intelligent machines are just on the horizon. Machines understand verbal commands, distinguish pictures, drive cars, and play games better than we do. How much longer can it be before they walk among us?

A recent White House report on AI takes an appropriately skeptical view of that dream. It says the next 20 years likely won't see machines "exhibit broadly applicable intelligence comparable to or exceeding that of humans," though it does go on to say that in the coming years, "machines will reach and exceed human performance on more and more tasks." But, the assumptions about how those capabilities will develop, missed some important points.

The report focused almost exclusively on what one would call "the boring kind of AI." It dismissed in half a sentence most of AI research, into how evolution can help develop ever-improving AI systems, and how computational models can help us understand how our human intelligence evolved.

The report focuses on what might be called mainstream AI tools: machine learning and deep learning. These are the sorts of technologies that have been able to play *Jeopardy!* well, and beat human Go Masters at the most complicated game ever invented. These current intelligent systems are able to handle huge amounts of data and make complex calculations very quickly, but they lack an element that will be key to building the sentient machines we picture having in the future.

We need to do more than teach machines to learn. We need to overcome the boundaries that define the four different types of AI, the barriers that separate machines from us—and us from them.

Type I AI: Reactive machines

The most basic types of AI systems are purely reactive, and don't have the ability to form memories or use past experiences to inform current decisions. Deep Blue, IBM's chess-playing supercomputer, which beat international grandmaster Garry Kasparov in the late 1990s, is the perfect example of this type of machine.

Deep Blue can identify the pieces on a chess board and know how each moves. It can make predictions about what moves might be next for it and its opponent and can choose the most optimal moves from among the possibilities.

But it doesn't have any concept of the past or any memory of what has happened before. Apart from a rarely used chess-specific rule against repeating the same move three times, Deep Blue ignores everything before the present moment. All it does is look at the pieces on the chess board as they stand right now, and choose from possible next moves.

This type of intelligence involves the computer perceiving the world directly and acting on what it sees. It doesn't rely on an internal concept of the world. In a seminal paper, AI researcher Rodney Brooks argued that we should only build machines like this. His main reason was that people are not very good at programming accurate simulated worlds for computers to use, what is called in AI scholarship a "representation" of the world.

The current intelligent machines we marvel at either have no such concept of the world, or have a very limited and specialized one for its particular duties. The innovation in Deep Blue's design was not to broaden the range of possible moves the computer considered. Rather, the developers found a way to narrow its view, to stop pursuing some potential future moves, based on how it rated their outcome. Without this ability, Deep Blue would have needed to be an even more powerful computer to actually beat Kasparov.

Similarly, Google's AlphaGo, which has beaten top human *Go* experts, can't evaluate all potential future moves either. Its analysis method is more sophisticated than Deep Blue's, using a neural network to evaluate game developments.

These methods do improve the ability of AI systems to play specific games better, but they can't be easily changed or applied to other situations. These computerized imaginations have no concept of the wider world, which means they can't function beyond the specific tasks they're assigned and are easily fooled. They can't interactively participate in the world the way we imagine AI systems one day might. Instead, these machines will behave exactly the same way every time they encounter the same situation. This can be very good for ensuring an AI system is trustworthy: You want your autonomous car to be a reliable driver. But it's bad if we want machines to truly engage with, and respond to, the world. These simplest AI systems won't ever be bored, or interested, or sad.

Type II AI: Limited memory

This Type II class contains machines that can look into the past. Self-driving cars do some of this already. For example, they observe other cars' speed and direction. That can't be done in a just one moment, but rather requires identifying specific objects and monitoring them over time.

These observations are added to the self-driving cars' preprogrammed representations of the world, which also include lane markings, traffic lights, and other important elements, like curves in the road. They're included when the car decides when to change lanes, to avoid cutting off another driver, or being hit by a nearby car. But these simple pieces of information about the past are only transient. They aren't saved as part of the car's library of experience it can learn from, the way human drivers compile experience over years behind the wheel.

So how can we build AI systems that build full representations, remember their experiences, and learn how to handle new situations? Brooks was right in that it is very difficult to do this. My own research into methods inspired by Darwinian evolution can start to make up for human shortcomings by letting the machines build their own representations.

Type III AI: Theory of mind

We might stop here and call this point the important divide between the machines we have and the machines we will build in the future. However, it is better to be more specific to discuss the types of representations machines need to form and what they need to be about.
Machines in the next more advanced class not only form representations about the world but also about other agents or entities in the world. In psychology, this is called "theory of mind"—the understanding that people, creatures, and objects in the world can have thoughts and emotions that affect their own behavior.

This is crucial to how we humans formed societies, because they allowed us to have social interactions. Without understanding each other's motives and intentions, and without taking into account what somebody else knows either about me or the environment, working together is at best difficult, at worst impossible.

If AI systems are indeed ever to walk among us, they'll have to be able to understand that each of us has thoughts and feelings and expectations for how we'll be treated. They will have to adjust their behavior accordingly.

Type IV AI: Self-awareness

The final step of AI development is to build systems that can form representations about themselves. Ultimately, we AI researchers will have to not only understand consciousness, but build machines that have it.

This is, in a sense, an extension of the "theory of mind" possessed by Type III AI. Consciousness is also called "self-awareness" for a reason. ("I want that item" is a very different statement from "I know I want that item.") Conscious beings are aware of themselves, know about their internal states, and are able to predict feelings of others. We assume someone honking behind us in traffic is angry or impatient, because that's how we feel when we honk at others. Without a theory of mind, we could not make those sorts of inferences.

While we are probably far from creating machines that are self-aware, we should focus our efforts toward understanding memory, learning, and the ability to base decisions on past experiences. This is an important step to understand human intelligence on its own. And it is crucial if we want to design or evolve machines that are more than exceptional at classifying what they see in front of them.

Readers may please note that D. L. Shah Trust brings out two e-journals on a fortnightly basis. These are mailed to those persons or institutions who are desirous of receiving them: These two e-journals are:

1. Safety Info
2. Quality Info

If you or your friends or colleagues wish to receive these journals, you may send us an e-mail requesting for the same. There is no charge for these journals. Our e-mail address is:

dlshahtrust@yahoo.co.in or haritaneja@hotmail.com or dlshahtrust@gmail.com

You can also access these journals on our website: www.dlshahtrust.org

Sponsored by: **D. L. Shah Trust**
For Applied Science, Technology, Arts & Philosophy
Mumbai. email: dlshahtrust@yahoo.co.in
Ph: 022-22838890

Edited by Hari Taneja, Mumbai and
Published by R. Ramamurthy, Bangalore 560084.
email: dlshahtrust@yahoo.co.in