Book review: *The Road to Conscious Machines. The Story of AI*

Author:
Michael Wooldridge. Professor of Computer Science, Head of the Department of Computer Science, University of Oxford

Publisher:
Pelican Books. Penguin Random House

Eva Valentí Ramírez
Head, Actuarial Review Department, Directorate for Risk Management
Consorcio de Compensación de Seguros

It is not uncommon for the media, cinema, and social media to spread ideas like:

1. Artificial intelligence (AI) will take away all our jobs in the near future. It will do anything we can do better than we can and will not have to be paid a salary.

2. Governments and certain powers that be will be able to manipulate the people easily through fake news made from AI-generated images and audio.

3. Superintelligent machines capable of rapid self-improvement will soon be here. They will evolve on their own and escape from our control.

Ideas like these are repeated so often that many people have ended up taking them as true, even though there is no actual technical basis for them.

Other times, instead of predicting terrible catastrophes, there is an overblown confidence in the capabilities of AI. For instance, as it applies to the insurance sector:

AI seeks to build machines that can "mimic" human behaviour so closely that they will ultimately be indistinguishable from ourselves. Machines that have the same range of abilities human intelligence has, known as **artificial general intelligence**, i.e., autonomous, self-aware intelligence capable of planning, reasoning, conversing, understanding jokes, telling stories, and so forth. This has not yet been achieved.

"Implementing AI tools will revolutionise the insurance industry. They will very shortly enhance the customer experience, sales efficiency, and claims processing. Insurance companies will undergo unprecedented growth."

Michael Wooldridge, the author, describes himself as *belonging to the first generation of humans that fooled around with computers as a teenager* and thinks that AI is neither as powerful or as advanced as one might be led to believe.

He has written this book to try to set the record straight and give a more realistic account of what AI is and is not, the technical successes that have been built up since its inception in the 1950s, and where things stand today.

Without shying away from the technical details but in a highly readable manner with lots of examples that are easy to understand, he sets out the basic concepts of AI and explains how it has grown into the powerful field we know today.

We know that computers excel at performing specific tasks, which they can do error-free at breakneck speed. A desktop computer can do in one second what it would take one person, working non-stop without slip-ups, 3,700 days to do.

Within the next ten years we will see completely autonomous cars, high-quality simultaneous interpreters, and software capable of detecting minute differences in the pixels on radiological scans and discerning tumours much better than any doctor can. There will even be mobile phone apps that can detect symptoms of dementia from how the user operates them.

All of this will improve our lives immensely. But that is not AI's final goal.

AI seeks to build machines that can "mimic" human behaviour so closely that they will ultimately be indistinguishable from ourselves. Machines that have the same range of abilities human intelligence has, known as **artificial general intelligence**, i.e., autonomous, self-aware intelligence capable of planning, reasoning, conversing, understanding jokes, telling stories, and so forth. This has not yet been achieved.

Scientists have gone down many a blind alley before realising that they need to retrace their steps and start over again down another path. They don't yet know if artificial general intelligence is feasible, and there is also no consensus as to whether it is even desirable.

The first part of the book tells the story.

As you might have guessed, it all started with Alan Turing and the "decision problem", the *Entscheidungsproblem* in German, the first step toward developing AI when it did not yet even have a name and there was no scientific community working on it.

Decision problems are mathematical problems with a yes or no answer, for instance: is 2+2=4? A decision problem is decidable if it can be answered following some finite steps (a rulebook); that is, a computer could solve it within a finite time. The question is, are all decision problems decidable or are some unsolvable by finite steps? That is, would a computer, no matter how fast it executed instructions, take an infinite amount of time to solve it? To answer that question, Alan Turing built the "Turing machine".

In this early period, between 1956 and 1974, known as the Golden Years, everything seemed possible. It was a time of unbridled optimism. The systems developed were given extravagant names. Scientists had to work at night because during normal working hours computers were used for more productive tasks. The idea was to build robots that could engage in something similar to a conversation or perform practical tasks like arranging a storeroom. But by the mid-1970s, after twenty years of research, only very basic progress had been made, and a portion of the scientific community began to think AI was a pseudoscience.

AI then fell into a dark, stagnant period until research changed course and began to develop the first **expert systems**. Building an expert system involved giving a computer the knowledge needed to perform specific tasks, knowledge that people who are experts in those tasks gain only after extensive training, and the computer could do them much better than a human. For the first time there seemed to be a glimmer that AI would be able to be economically profitable.

So at the end of the 1970s a new period of enthusiasm emerged. But by the end of the 1980s no notable advances had been made. It turned out that it was not so easy to transfer human experience into coded instructions that a computer could execute. Scientists working on AI were again accused of selling smoke and mirrors, promising much and of not achieving any concrete results.

The direction of AI research shifted once again and would keep scientists busy for the next 10 years from 1985 to 1995. It was concluded that progress could be made only if systems gained information directly from the actual environment in which they were located. The idea was to set the behaviour the system should manifest in a given situation, organised into hierarchical levels, with one or another taking precedence – **behavioural AI**. The next step was to develop **agents**, self-contained AI systems that were autonomous and capable of holistically performing the tasks assigned by users.

In the meantime, since AI's inception, research had been moving forward down another revolutionary path: building machines that could learn.

The goal of building computers capable of learning is to design programs that achieve results starting from some input data, even though the software does not explicitly include a "rulebook" showing how to get there.

For that, the software has to be "trained". There are two types of training. The first is **supervised training**, achieved by feeding the machine a wide-ranging array of possible situations. This raises one of the ethical problems facing AI: if the dataset input for training is biased, the decisions taken by the computer will replicate that bias, giving rise to unfair outcomes.

For instance, a bank uses software to identify each customer's risk when granting bank loans. Software of this type is ordinarily trained using a set of records on previous customers labelled according to a risk classification of high or low. However, if the volume of data used for each customer is too large, training takes too long. So, what data can be omitted if we do not know which items are relevant for determining risk? For example, if the sole data item input for training is the customer's address, this could cause the software to discriminate against people living in certain neighbourhoods and prevent potentially good customers from obtaining loans.

The second type of training is **reinforcement learning**. The software is not fed explicit data but is allowed to take random decisions and is given negative or positive feedback according to whether the decisions are good or bad. The software takes that feedback into account when making its next decision.

One of the current challenges is to avoid bias within algorithms, because we do not know what path the algorithm takes to reach the decisions it makes.

So for the machine to take decisions, it only has to be told what it should do, and the machine itself will modify its behaviour, it will learn.

Then the question is, how does a program learn? The method of learning – deep learning – consists of giving the computer a **neural network** architecture capable of being trained. That structure is based on the nervous systems of animals in which nerve impulses are transmitted from one neuron to the next, which are triggered or not triggered by neurotransmitters released at the synapse. In the computer those chemical neurotransmitters are replaced by two numeric values, the **weight and the firing threshold**. The next neuron in the network fires or does not fire depending on the combination of those values.

The end of the book reviews AI's present and future.

It considers two achievements attained by AI that are already a reality, self-driving cars and AI applications in health monitoring.

Two chapters ironically headed *What we think might go wrong and Things that really could go wrong* poke fun at our fear of "the new" and what we are sure about without any reason to be and pick apart the risks AI actually poses that we would do well to bear in mind. There are thoughts on the future of work and human rights that go beyond the old cliché of disgruntled workers managed by an algorithm. The author puts forward interesting speculations on the changes AI will bring about in the work-driven society we know today. He also considers the problem of fake news and the moral dilemma of autonomous weapons systems guided by AI that makes its own decisions.

To conclude, the author amuses himself by fantasising about what a machine that we really could not tell apart from a human being would be like. What is consciousness? If a machine really was self-aware, would we even know it?

And what if artificial general intelligence just isn't possible?