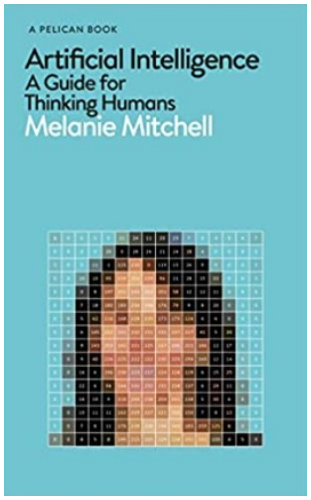

Title, Author: Artificial Intelligence: A Guide for Thinking Humans, Melanie Mitchell

Published: 2019, Pelican

Questions such as will AI take over the world, cause mass unemployment, make humans obsolete or create a Marxist utopia[1] feel like they are in abundance today. We all must have a view but the relatively recent nature of machine learning compared to other occurrences during our history, combined with its fast developments and general complexity often results in such a view being uninformed. Enter stage left Melanie Mitchell with her new book *Artificial Intelligence – A guide for thinking Humans*. It covers the intellectual history of the field with characters such as Alan Turing and Ray Kurzweil, alongside the practical successes (and continued challenges). There is also a healthy dose of involved but digestible overviews of the technical aspects driving AI achievements. For instance the explanation of techniques such as image recognition software or natural language processing using "deep learning" through convoluted neural networks (or convnets for short!), was certainly worth persevering with.

The most valuable aspect of the book was how she contextualised the overall discussion with an examination of intelligence itself. How can we answer some of the questions above if we don't actually understand what constitutes *thinking?* Put differently – to understand artificial intelligence we first need to understand what understanding itself means and how we do it as humans.

The history of AI since its beginning in the 1950s has been somewhat seasonal. Certain developments about a machine programme performing a human function create significant high-profile promise of impending world changing developments which attract considerable funding and media coverage. The resulting developments don't materialise or underwhelm, causing an *AI winter* of intellectual and financial disinterest. We are certainly in the midst of considerable promise and investment, and within the context of mass availability of data and significant improvements in computing power, there is a compelling argument at the very least to say that the length of the cycle of interest and disinterest is narrowing. Computers can beat human beings at Chess and Go, recognise images, automate large aspects of driving, and predictions from certain quarters believe that a general level of intelligence amongst machines could be with us by the end of the decade.

The reality according to Mitchell is that AI is a long way from being at a general human level. How long? "Take your estimate, double it, triple it, quadruple it. That's when" is the view she shares of Oren Etzioni, Allen Institute for AI (launched by Microsoft co-founder Paul Allen). But what about Ray Kurzweil's view of the singularity whereby machines learn to do everything better than humans – crucially including making machines? This is based, amongst other things, on Moore's law of the doubling of computing power every two years resulting in an exponential lift off at some point, due to the power of compounding. If these trends continue, a high-street computer will "achieve human brain capability (10^{16} calculations per second) …around the year 2023. At that point, Kurzweil believes that

[1] Artisan carpentry in the morning, theatre going in the evening.

achieving human-level AI *will just be a matter of* reverse engineering the brain." Whilst much of the successes of AI have been based on neuroscience, our understanding of the brain is too superficial. Moreover, a new-born baby's brain does not have human intelligence until it has interacted over a period of time with a complex world. Kurzweil has a bet with Mitchell Kapor that a computer will pass the Turing Test by 2029 on the website Long Bets. Kapor asserts that without the equivalent of a human body, and everything that goes along with it, a machine will never be able to learn all that is needed to pass his and Kurzweil's strict version of the test.

This is compelling when viewed in the context of the *types* of mistakes that AI currently makes. For instance, despite the significant advances in translation software when Google Translate tells you in English that the French diner "forgot to ask for his piece of legislation" rather than "forgot to ask for the bill", it reveals more than an amusing mistake. The *artificial intelligence* programme is just following pattern recognition in its software and by getting the context so wrong shows it doesn't truly understand the task it has been asked to solve. Until it can understand the context of a situation it can never be viewed as intelligent – at least as we humans define it.

Beyond poor transferability, AI software can be manipulated to detect things hacking software wants it to detect. Having initially felt that that our entire legal, financial and Governmental bureaucratic systems could effectively dealt with by AI, I am now more cautious due to learning how easily manipulatable simple image recognition software is. For instance stickers which are undetectable to the human eye can be placed on stop signs causing driverless cars to continue rather than come to a halt. Despite the sign being large, red and having STOP written on it the software detects the image and understands it differently to how we do. We understand from a very early age that a large red thing in that shape probably means something important even if we never came across the word STOP before.

Our understanding of more complex concepts, particularly abstractions, is made possible by using more earthly ones in the form of metaphor. Mitchell references George Lakoff's book *Metaphors we Live By* to demonstrate how AI would struggle to understand the world in its current unevolved form. For instance, to understand something like *time* we use the more concrete concept of money. *Time* can be spent, saved, wasted, worthwhile, or is often used profitably. We often use physical directions to understand emotional states – we *fall* into depression, have *lifted* spirits, feel *up* or *down*. It's difficult to programme software to learn like this. For instance would a computer think to use temperature to better understand social interactions? *Icy stares, cold shoulders,* and *warm welcomes*, allow us to quickly describe a photograph whilst AI struggles to simply identify the objects therein.

For those thinking that they can use recent advances as a way to predict stock markets or solve inefficient government should note that considerable *engineering muscle power* goes into programming AI. In terms of directing that power "*it's almost like an artform to get the best out of these systems…there's only a few hundred people in the world that can do that very well*", according to Demis Hassabis, CEO of Google DeepMind,there's a lot of human driven trial and error involved, calling into question the nomenclature of "machine learning".

Microsoft's Eric Horvitz describes recent advances as taking us only to the stage of alchemy, but what's clear is that the availability of big data and ever-increasing computing power has

been a game changer for progress in attempting to improve the ability of machines to do things that were the sole preserve of humans. Reading the book has given me a better understanding of why and how this has come about, whilst informing me of the dangers and opportunities that lie before us as we embrace developments that the vast majority of us are blissfully unaware of. To stand a better chance of not making catastrophic mistakes, having an informed opinion is of vital importance and Melanie Mitchell has made an excellent effort to allow us take the first steps.

Amul Pandya
July 2020