The road to conscious machines: AI through failed ideas

Roman Krzanowski
Pontifical University of John Paul II in Krakow

Michael Wooldridge. 2021 [1st ed. 2020]. *The Road to Conscious Machines: The Story of AI*. Dublin: Pelican, pp.394.

The book *The Road to Conscious Machines: The Story of AI* is a seemingly simple, introductory book about AI, but that's not quite the case. What sort of a book on AI opens with a declaration like this: "much of what is published about AI in the popular press is ill-informed or irrelevant. Most of it is garbage, from a technical point of view, however entertaining it might be" (Wooldridge, 2021, p.3)? Wooldridge clearly has something different to share than just another AI story. One of the very first statements in the introduction explains what this is: "It's the story of AI through failed ideas" (Wooldridge, 2021, p.1).

In fact, the book offers much more. It presents a philosophical exploration of the AI-related quest to understand the essence of our minds, ethics, and consciousness. (AI has always sought the full spectrum of human agency.) As he observes "... we do not remotely understand what it is we want to create, or the mechanisms that create it in people" (Wooldridge, 2021, p.2). Woodridge message is this: We have not constructed artificial minds, automated moral machines, or artificial consciousness because we have never understood how the mind works, the essence of ethics, and what consciousness actually is. Initial and subsequent waves in AI have been informed by ideas (philosophies) that were simply wrong, and we have no clear path ahead of us either. The insightful discussions of prior and current failures in AI technology and a sobering look at its so-called "success stories" is what makes Wooldridge`s book different from other introductions to AI and what makes it worthwhile to read.

Michael Wooldridge is an old hand at AI. He has been around from the early days of the technology, taught at Oxford, published numerous research papers, led international conferences, and been president of the European Association for AI. It's therefore fair to say that he knows what he is talking about. We may, however, argue where the beginnings of AI fit on the timeline: Is it with Charles Babbage or Alan Turing, or is it with the early neural network (NN) systems of McCulloch or Hebb.[1] Wooldridge joined the AI crowd around the time of the second AI winter (i.e., the late 1980s and early 1990s). The seasonal analogy of "AI winters" represents the drying up of funding due to a lack of progress in this research field, and there were also long periods of frantic search for the next breakthrough development, although there may have been some overlap, as some suggest.

On a first reading, the book seems rather simple and non-technical. It begins with an attempt to define AI, which is no simple task in itself. Wooldridge posits that we should differentiate between what the dream of AI is and what is actually delivered. AI definitions are mostly a fusion of these two ideas (Wooldridge, 2021, pp.1–3) in varying proportions, thus obfuscating the AI concept rather than clarifying it.

After the anticlimactic introduction, Wooldridge takes us through the history of AI, the past endeavors, and the current achievements before moving onto a speculative future. AI for Wooldridge begins with

[1] The origins of NNs go back to 1940s and 1950s and are associated with names of Warren McCulloch, Walter Pitts, D. O. Hebb, Wesley A. Clark, and Frank Rosenblatt.

Turing's paper on machine intelligence (Turing, 1950). This first wave of AI lasted from the early 1950s to the early 1980s, with it ending with the first AI winter. After this, in a period called the Golden Age of AI, we inherited technologies like symbolic AI, ELIZA, SHRDLU, LISP, and the concept of narrow AI. (Wooldridge positions this as a layman's term, because all AI is narrow in this sense (Wooldridge, 2021, p.42)). The death knell to this wave of AI came with Lighthill's report (Lighthill, 1973). The main problem was that Lighthill was essentially right about AI research at the time: It was working from faulty assumptions, so it was going nowhere.

Then came knowledge-based AI systems, or as they were known to lay audiences, expert systems. In these times, we got MYCYN, PROLOG, logic-based AI, Cyc, knowledge graphs, knowledge bases, inference engines, and knowledge engineering. All this went fine until people started to ask these logic-based systems, as these expert systems were, simple common-sense questions that did not necessarily follow logic, as we often do in real life (Wooldridge, 2021, pp.89–123). Expert systems failed miserably in these tests, so the concept of knowledge as a set of clear logical relationships between clearly defined concepts failed as well. Then came robots. Logical AI at the heart of robotic systems could only cope with specific tasks in a box-like world, because a symbolic and logical representation of the environment did not work well beyond simple situations. Nevertheless, human ingenuity in the quest for funding is boundless, so a new AI paradigm was created called behavioral AI, which was proposed by Rodney Brooks.[2] In Brooks's view, robotic systems must be situated directly in the environment, with their intelligence not being based on a preprogrammed set of rules but rather an emergent property arising from interaction with the environment (Wooldridge,

[2] Rodney Brooks. Available at http://people.csail.mit.edu/brooks.

2021, pp.126–127). His manifesto was published under the title "Intelligence without representation" (Brooks, 1991). The idea was right, but the realization lagged behind somewhat. Next came agent-based AI systems, SIRI, AI assistants, and reasoning under uncertainty, while Deep Blue beat Gary Kasparov at Chess (Wooldridge, 2021, pp.125–165). This was at the cusp of the new millennium, and some thought we were done.

But with the advance of hardware, new possibilities opened up. New artificial neural network (ANN) systems based very roughly on the model of neural nets found in biological systems were conceptualized and put to work. This technology had begun with the older model of the perceptron (Neural Net version 1), gone through connectionism (Neural Net version 2), arrived at Deep Learning (Neural Net version 3), and culminated in DeepMind, with AlphaGo winning a game of Go against a Korean Go black belt. And this is where we are now. We have created self-learning systems that seem to learn tasks by themselves and exceed even what humans are capable of. For example, DeepMind learned to play Atari 2600 console games better than any human, and AlphaGo learned Go strategies (Wooldridge, 2021, pp.167–210). These systems program themselves in a way, so they are more than just Turing machines in this sense. But the question is this: Where are we really in relation to the ultimate goals of AI? Are we touching the dream of artificial general intelligence (AGI)? Wooldridge claims that we are not even close, because human intelligence and life is not like an Atari game or the game of Go.

We then get chapters (Wooldridge, 2021, pp.237–303) on the potential harmful effects of AI technology on the global scale, such as the problem of killer robots, algorithmic bias, fake AIs, massive surveillance, and mental manipulation undermining human agency, democracy, and the bases of democratic societies. The final chapter discusses conscious AI. I will omit the details and merely repeat Wooldridge's claim that conscious AI is more of a dream, but he thinks it cannot be totally excluded. It may happen that like with the Turing test, we will have a test of consciousness. If a system passes such a test, we will say it is conscious. Whether it is conscious in the same way we are is another matter. The conclusions are similar to those implied in the Turing test for human intelligence.

Next comes Wooldridge's list of AI blunders, conceptual phantasms, and simply poorly-thought-out ideas, and this is where the story of AI gets really interesting. In the section "The Singularity is Bullshit," Wooldridge resolves the issue of artificial brain capacities exceeding human intelligence, as embodied in Ray Kurzweil's concept of the singularity. He writes that there is no chance of this happening soon, if at all. In the section titled "We need to talk about Asimov," Wooldridge disposes with the so-called Three Laws of Robotics (TLR), which are often used by aspiring AI researchers when they venture into ethical domains and seasoned, and sometimes not so seasoned, philosophers when they explore the realities of AI. Wooldridge says these laws are clearly not realizable, period! He writes that we should drop them from serious discussion before they do too much harm. Indeed, it seems that Asimov reached the same conclusion after thinking about them a bit more. The next section is entitled "We need to stop talking about the trolley problem." For Wooldridge, this problem is too simplistic to be meaningful yet too complex to be solved by current synthetic systems. After all, how can we ask an AI agent to solve something that we cannot solve ourselves? This is of course very thoughtful advice. Wooldridge is not a philosopher of ethics, however, so he does not add the ethical dimension to the naiveté of the trolley problem. Ultimately, all dreamed up laboratory-like ethical problems are never real ones—they do not come even close to real-life situations.

As guides to practical solutions, they are useless, ethically speaking, in practice. Finally, we get a section titled “Be careful what you wish for.” Wooldridge reminds us of that engineering systems, which AI systems are *par excellence*, always fail in unpredictable ways, have unforeseen errors, or pursue unexpected actions. When faced with the complexities of life, we can never be sure what a complex AI system may decide to do. The paradigm case is that of the paper clip factory. So be aware! I wish every paper on autonomous robots would have a warning like this.

And what do we learn about AI from the promised second reading? We face the fact that thus far, all AI revolutions have failed to achieve the dream of AI (i.e., AGI). This is the yardstick by which we can judge AI and its status and progress, or lack of it. AI systems have so far failed because they have been based on incorrect assumptions about the mind, reality, reasoning, thinking, and human nature and its embodiment in the world. We have developed AI systems that exceed human capacities but only in specific, narrowly defined tasks. Even if these capacities appear immensely impressive to us, they are not what we are after.

We have begun developing AI under Descartes’ shadow, with Descartes’ views about reality being composed and presented to us in clear, logical, distinct ideas. Turing was also wrong because people do not “reckon” as he thought. This was the first and second wave of AI. The reality is not that of Descartes, nor is our thinking like that of Turing’s machine. The reality is messy, unpredictable, and fluctuating. Clear distinct ideas and logical reasoning only go so far if they go far at all. The realities of life are too complex to be programmed in clear, distinct steps, and the ontology of reality is not that of the box world. If we see the world’s problems like stacking boxes to reach some hanging bananas, we only get bananas and little else. Monkeys know this well. Life is much more complex than the multiplication of floating-point numbers and the deterministic universe of 8x8 chess board or even a 19x19 Go grid. Knowledge is not a set of prescribed, deterministic, static rules, complex or otherwise. This is why the second AI wave with its knowledge systems failed. With neural nets, we got self-learning systems, yet comparing artificial neural nets to the structure of the brain is again a mistake. Human neural systems do not need millions of cases to learn how to recognize a cat, for example. We do not actually have a theory of the mind that would be implementable or even one that would explain "how the mind does it." Nor do any reductionist theories add up to the human mind, as they are reductionist theories, at least at present. So, the failure of AI technology is a failure of our philosophy. We simply do not know how the mind works. AI research seems a bit like groping in the dark. As a careful researcher, Wooldridge is not saying that AGI is impossible, just that it is not close, and we are not on the right track to achieve it anyway. Yet how we can replicate something when we do not know what it is?

Wooldridge's book also shows us how to look at various AI developments, such as ethical autonomous robots, social robotics, emotional robotics, and whole-brain emulation. To evaluate what these or other AI techniques offer, we need to get behind technology and algorithms and explore the philosophical bases and assumptions, whether implicit or explicit, of proposed systems. The gap between what is expected or claimed under the big headings of AI ethics, AI social robotics, and AI emotion systems and what is actually proposed will tell us what a given technique is really about. One AI system may be better than another, but they all fail miserably in terms of what people think these systems offer.

There are many books on AI, so is it worth reading this one? I think we may say, "Yes." You could read Wooldridge to get an initiation into AI, but there are plenty of other books that are more technical and more detailed (see e.g. Russell and Norvig, 2020). Wooldridge's book is published under the Pelican book series, which is known for quick, simple introductions to a topic, perhaps less technical than the "A very short introduction…" series by OUP.[3] Even Wooldridge's technical section—which is intended to provide more in-depth information on topics like Prolog, Bayes's, and deep neural networks—is rather elementary in my personal opinion.

But Wooldridge's book offers insights into the science and technology of AI that other more technical AI books tend to overlook. The uniqueness of this book is embodied in the already quoted opening statement: "It's the story of AI through failed ideas." I would, however, add "and the philosophical blunders" to this. It is really unexpected, and surprising, that engineering has shown us where our philosophy went wrong. From this perspective, this book seems unique. Thus, even if you know a lot about AI, Wooldridge's book can give you a perspective on AI technology that is usually only available through a careful critical synthesis of tons of research and years of experience. You will also get insights into our knowledge, or lack of it, of the mind and intelligence, which informed AI research for better or worse, and be pre-warned of the fads and so-called received wisdoms that the AI field and popular press are crowded with. On reading the book, you will end up with a much clearer vision of AI technology and the failures of our philosophies for the mind and of AI, because

[3] Very Short Introductions. OUP Web site https://www.veryshortintroductions.com Accessed on 2021.08.27.

we posit that AI technology is driven, implicitly or explicitly, by our philosophical ideas. AI is philosophy in practice, as all engineering is. These insights come only on a second reading, however.

As a follow up on the future of AI and the past failures, one could also consult Russell's *Human Compatible* (2020), Brockman's *Possible Minds* (2019), Poole and Mackworth's *Artificial Intelligence: Foundations of Computational Agents* (2017), or Smith's *The Promise of Artificial Intelligence* (2019). While these books do not attempt to introduce AI, as they assume that the reader is already initiated into the topic, they provide valuable philosophical reflections on AI technology. They could be suggested as later, more advanced, supplements to Wooldridge's AI story. Nevertheless, all these efforts to rethink AI concepts are coming from engineers and computer scientists in the form of thought with some enlightened perspective. Their problem is that they perpetuate Descartes' miscalculation about the separation of mind and body, while some form of anti-Cartesian embodied mind may be the proper way to progress toward AGI.[4]

Abstract

Presented here is a review of the book by Michael Wooldridge *The Road to Conscious Machines: The Story of AI*. The book was published in 2021 by Pelican Publishing Company. The book presents a philosophical exploration of the AI-related quest to understand the essence of our minds, ethics, and consciousness. Nonetheless, the

[4] The proper relationship between the mind and body should be explored through non-technical sources rather than in the computer and engineering literature. See, for example, the rather old but still relevant (Abram, 2017, pp.31–73) [1st ed. 1997] or Hubert Dreyfus's ideas collected in (Dreyfus, 2016) [1st ed 2014]. Interesting biosemiotic perspective is shown in (Sarosiek, 2021).

book is unconventional as it focuses on failures of past AI programs rather than on AI success stories. The author also indicates the possible pathways to develop new, more efficient AI paradigms. The short technical section at the end of the book offers more in-depth information on topics like Prolog, Bayesian, and deep neural networks.

Keywords

artificial intelligence, history of artificial intelligence, artificial intelligence paradigms, future of artificial intelligence.

Bibliography

Abram, D., 2017. *The Spell of the Sensuous: Perception and Language in a More-Than-Human World*. 20th. New York: Vintage Books, a division of Penguin Random House LLC.

Brockman, J., ed., 2019. *Possible Minds: Twenty-Five Ways of Looking at AI*. 1st ed. New York: Penguin Press.

Brooks, R.A., 1991. Intelligence without representation. *Artificial Intelligence* [Online], 47(1-3), pp.139–159. Available at: https://doi.org/10.1016/0004-3702(91)90053-M.

Dreyfus, H.L., 2016. *Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action*. Ed. by M.A. Wrathall. First published in paperback. Oxford; New York, NY: Oxford University Press.

Lighthill, J., 1973. Artificial Intelligence: A General Survey. *Artificial Intelligence; a Paper Symposium* [Online]. London: Science Research Council. Available at: <http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm>.

Poole, D.L. and Mackworth, A.K., 2017. *Artificial Intelligence: Foundations of Computational Agents* [Online]. 2nd ed. Cambridge: Cambridge University Press. Available at: <https://artint.info/2e/html/ArtInt2e.html> [visited on 27 August 2021].

Russell, S.J., 2020. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Penguin Books.

Russell, S.J. and Norvig, P., 2020. *Artificial Intelligence: A Modern Approach*. Fourth edition, global edition, *Pearson series in artificial intelligence*. Harlow: Pearson.

Sarosiek, A., 2021. The role of biosemiosis and semiotic scaffolding in the processes of developing intelligent behaviour. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (70), 9–44.

Smith, B.C., 2019. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA: The MIT Press.

Turing, A.M., 1950. Computing Machinery and Intelligence. *Mind* [Online], 59(236), pp.433–460. Available at: https://doi.org/10.1093/mind/LIX.236.433 [visited on 13 July 2016].

Wooldridge, M., 2021. *The Road to Conscious Machines: The Story of AI*. London, UK: Penguin.