BOOK REVIEW

Artificial Intelligence: Modern Magic or Dangerous Future? by Yorick Wilks (2019) Icon Books, London, 176pp., £9 (paperback) ISBN: 9781785785160

Artificial intelligence has left the worlds of academia and the research laboratory and spread to the worlds of business and public policy. The past several decades have seen many technological advances that have offered ever more capable and convenient prosthetic and orthotic devices that have simplified a lot of everyday tasks, saved a lot of human effort and in many cases done better than humans. This progress has, more often than not, been attributed to advances in AI. So, it is not surprising that AI has become a staple in the output of mass media. This avalanche of news is driven in very large part by public fascination with technology that seems to defy imagination, as well as by the *frisson* of doomsaying prophecies of machines taking over the world. In his recent book, *Artificial Intelligence: Modern Magic or Dangerous Future?,* Yorick Wilks presents concise and cogent arguments why AI is not modern magic and why it should not be feared.

The genre of this book is difficult to pinpoint. It is neither a technical introduction to the field nor a practical 'how to' manual for people considering a career in AI. That territory is staked, respectively, by such volumes as *Artificial Intelligence: A Modern Approach* by Stuart Russell and Peter Norvig, and *Artificial Intelligence for Dummies* by John Mueller and Luca Massaron. Nor is this book just a collective biography of AI notables and flagship projects. The book is a sort of intellectual teaser – as Professor Wilks puts it, 'a quick and painless introduction to its history, achievements and aims – immediate and ultimate'. Like a movie trailer that is intended to help people decide whether they want to watch the flick, this book will give serious but busy readers enough background and knowledge of core AI issues and trends to decide whether they want to delve deeper (for example, by following the excellent list of suggested readings on pp.157–8).

The first chapter, quite naturally, tackles defining what artificial intelligence is. This is not an easy task because the term means different things to different people. Wilks cites influential opinion on this topic, notably that of the AI pioneer John McCarthy, who posited that AI 'should be chiefly about getting computers to do things humans do easily and without thinking' and that of the philosopher Hubert Dreyfus, who believed that only entities that learned while growing up could be considered true AI. Another axis of tension in AI is between the goal of creating general, human-like, intelligence and making partial practical advances in developing reasonably useful applications. The latter attitude is traced back to Alan Turing and the intentions behind his famous Turing test of machine intelligence. Wilks (p.8) writes:

> Turing was not trying to say that computers did or ever would think: he was trying to shut down what he saw as useless philosophical discussion and present a practical test such that, if a machine passed it, we could just agree that they thought and so could stop arguing fruitlessly about the issue.

The core definitional issue, then, is whether AI is intended to imitate human capabilities or seek to imitate how people operate. That is, 'should AI be based on building representations inside computers of how the world is, or should it just be manipulating numbers so as to imitate our behaviour?' (p.11). Throughout the book, Wilks traces the historical development of two paradigms of AI research – representation-oriented (classical) AI and empirical, machine learning-based AI. He offers important insights into the reasons underlying the changing R&D fashions, as well as each paradigm's strengths and stumbling blocks.

Wilks then considers the role of logic in AI. In the early years of AI research, most practitioners worked on writing computer programs that emulated formal logical reasoning, which was

considered the mainstay of human rationality. Wilks spent a significant portion of his early career dispelling this belief as misguided or, at best, impractical. Here he explains the genesis of this belief and describes the many issues that it brings about, such as ample evidence from psychological research that people do not think logically in everyday life, the less than impressive results of the once-prominent theorem-proving research program, and the theoretical limits to what logic can do to support reasoning and decision-making, most famously introduced by Kurt Gödel. Wilks traces this argument ultimately to Hume, who stated that 'the power of logic was overstated where everyday life was concerned' (p.23). In the final analysis, AI would have a much more limited impact if it addressed only the well-behaved world inhabited by fully rational entities. In reality, as Wilks correctly points out, it is not so much logical reasoning as acquisition of formally represented knowledge and a set of decision heuristics that propelled much AI progress, from the expert systems of the 1970–80s to IBM's Jeopardy-winning Watson system.

The excitement generated by expert systems ultimately ebbed because of the difficulty of obtaining enough formally -represented knowledge, the so-called 'knowledge bottleneck'. Faced with this obstacle, AI had a choice of looking for ways to overcome the bottleneck or ways to bypass it. The paradigm-changing decision was influenced by two factors. First, spectacular advances in computing power and storage capacity made practical the use of huge amounts of raw data, for example, to learn about regularities and cooccurrences in it and use these to reach conclusions, bypassing logic or human judgement. Second, a sea change occurred in what was considered success in AI. Having noticed that big data and statistics-based findings incurred significant error rates, the field decided to turn the tables and view situations in which an AI system yielded correct results in, say, 60% of cases as success and rather than failure. This was acceptable – and remains universally accepted, as can be seen from the popularity of leaderboards in AI R&D – since the new criterion of R&D success is Insert doing better than your competitors on a particular (typically narrowly defined) task. As a result, little attention is being paid to the objective criteria of success or general needs of a comprehensive application. This is the mode of R&D that eventually led to Watson and the championship Go program.

Chapter 3 illustrates how ingeniously the material in the book is presented. Wilks explains how readers actually encounter AI in their lives by using an example with which they are quite familiar, the worldwide web. The discussion then flows into the idea of markup languages, the semantic web, connections with databases, information retrieval and the use of language processing to support web-based applications. Chapter 4 is devoted to demystifying the way AI is actually implemented in digital computers. The terminology introduced here (e.g., coding, algorithm, software, recursion, iteration) will be useful for people outside the field. In addition, Wilks explains the motivation behind the introduction of the classical AI programming languages, Lisp and Prolog.

Chapter 5 addresses the processing of spoken and written language in AI systems. The topic is one of the central pillars of AI, one to which Professor Wilks has made many important contributions over his long career. The chapter discusses the relation between language processing and logic, approaches to knowledge representation and the impact of machine translation (MT), for many years the foremost application area for language processing. The chapter describes the historical move from logic-based AI to 'shallow' methods in both MT and natural language processing (NLP) overall. The influence of automatic speech recognition and information extraction is also highlighted. These discussions neatly underscore the tension between the scientific and technological concerns within AI.

In the next chapter, Wilks ponders on the nature of learning. A brief description of learning with neural nets introduces readers to the principles underlying spectacular claims about how machines can learn. Current deep learning algorithms work very well on a number of specific tasks, but typically require a lot of training using annotated datasets. When machine learning relies on annotated training data, it is called 'supervised' learning. In contrast, 'unsupervised' learning bypasses the need for people to prepare these datasets. Although it is a hot area of research, results are inconclusive at best. What is clear – and is well argued in the book – is that deep learning

algorithms can detect correlation but not causation. This means that the reasons for their decisions are typically obscure, even for people. Also, machine learning does not attempt to imitate human processing, because, unlike humans, these programs cannot learn from just a handful of examples. Still, machine learning methods are attractive, in large part because they allow AI practitioners to dispense with coding the knowledge to support an AI system. Of course, annotating training datasets also requires a lot of human labour, but for some reason the practitioners of machine learning approaches to AI do not consider this a hindrance.

The chapter ends with a discussion of four issues important for understanding the current state of deep, learning-oriented AI: opacity, novelty, fusion and limits. Opacity refers to the baffling state of affairs when deep learning AI generates results that 'not even its designers and programmers always understand fully'. The discussion of novelty laments the unfortunate 'lack of scholarly memory in AI and the relentless emphasis on novelty'. Fusion hints at the need for integrating machine learning methods with knowledge-based ones, specifically with respect to the provenance of the features of the world and situations that are relevant to statistical decision-making. The limits of machine learning relate to the inability of this approach to deal with causality. Wilks presents the idea, inspired by the work of Judea Pearl, that 'causation is never in the data of the world, it is something we impose on it in order to understand'.

Vision and robotic action are the topics of the next chapter. As with other topics, Wilks presents a concise, non-technical overview. This discussion forms a natural segue into the topic of AI companions, truly intelligent devices aimed at supporting humans in a variety of applications. In a sense, this is the central part of the book, as companions are presented as a core AI project of the present and the future. Wilks has led a number of projects in this area over the years and has made significant scholarly and community-building contributions to it. The discussion, though still non-technical, is detailed and presents a variety of past, current and potential future approaches, applications and options in developing AI companions. Chapter 8 follows with a programmatic discussion of how a true AI companion will provide a qualitative leap in people's interaction with the worldwide web: how it will go way beyond the capabilities of Siri and other such apps in having an identity, being discreet, safeguarding the information it has and providing help in dealing with what Wilks calls 'the military-industrial-information complex' on a variety of issues related to privacy. This reckoning is expanded in Chapter 9, in which Wilks claims that 'the web may become unusable for non-experts unless we have companion-like agents to manage its complexity for us' (p.120). He proceeds to discuss issues that arise in connection with this development, such as identity and identity shielding and safeguards for the information content of a companion.

The penultimate chapter looks at a sampling of 'philosophical and political issues that AI has affected or created'. There is no consensus on these issues, and Wilks presents only his personal views. The issues covered include automation and loss of jobs caused by AI, and such ethical issues as the allocation of responsibility, the ethics of automated warfare, the relation of AI to religion, machine consciousness and fake news. Wilks neatly summarizes the main topics and concludes by underscoring the fact that, even with all the recent progress, current AI systems are still not able to conduct a conversation. Turing suggested this in 1950 as a test for judging whether a machine can really think. Still Wilks's overall assessment of the current state of AI and future developments in the field is not at all pessimistic. There are good reasons for today's AI to concentrate on narrow tasks and the low-hanging fruit in terms of applications. But the field is increasingly aware that to crack the hard nut of human language understanding and to have machines emulate human learning capacity, some kind of hybrid of knowledge-based and machine learning-oriented approaches must emerge in the future.

This book is very useful in that it occupies the usually neglected space between science journalism and academic surveys of a scientific discipline. In just about 40,000 words, it manages to comment on all the major scientific and historically prominent issues in AI and connect them with the relevant developments in philosophy, psychology, logic and computer science. When Wilks offers projections into the future, they are informed by a long-term, historically motivated

view of the field that distils a lifetime of work in AI and its applications. For example, the concept of semantic ascent (p.38), introduced by Richard Braithwaite in the 1950s, was the subject of the author's 1971 paper. One must also stress the remarkable even-handedness in the book, particularly commendable for an author who has always been an active polemicist on the topic of AI methodologies. Finally, what makes the book eminently readable is the author's ability to describe very complex issues in simple terms. The discussion of Gödel and the inconsistency of logical systems is an excellent example.

This reviewer has assigned the book to new graduate students of cognitive science at Rensselaer Polytechnic Institute. These students already had some background in AI, having taken regular AI courses. Still, it was gratifying to see that the book significantly expanded their understanding of the AI enterprise.

Sergei Nirenburg
Rensselaer Polytechnic Institute, Troy NY
nirens@rpi.edu