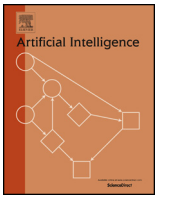
Artificial Intelligence 289 (2020) 103386

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint



Artificial Intelligence requires more than deep learning — but what, exactly?

Michael Wooldridge

Department of Computer Science, University of Oxford, United Kingdom of Great Britain and Northern Ireland

ARTICLE INFO

Article history:
Received 6 August 2020
Accepted 7 September 2020
Available online 29 September 2020

Review of Rebooting AI: Building Artificial Intelligence We Can Trust, by Gary Marcus and Ernest David (Pantheon Books, New York, 2019)

These are dizzying times to be an AI researcher. Just twenty years ago, many outsiders treated our field with suspicion, even ridicule. Beyond the confines of university computer science departments, AI as a scientific/engineering discipline was barely acknowledged at all.

Today, things could not be more different. AI is in the press on a daily basis, with new AI systems regularly making global headlines. Researchers from other sciences are queuing up at the doors of AI researchers, desperate to sprinkle some AI fairy dust over their own work. The leaders of our field are revered as heroes and heroines, garlanded with honours and invited to meet the heads of government and industry — even the Pope. Surveys in some countries show that aspirational teenagers, who might previously have dreamed of finding glory through being a musician, film star, or football player, now see a career in AI as the path to fame and riches. The level of funding being directed at AI right now, and the continuing frenzy around AI startups and acquisitions, truly defies comprehension. Who could possibly have predicted anything like this at the turn of the century?

Of course, while all AI researchers are basking in the bright warm sunshine of a rejuvenated and celebrated field, the truth is that for many of us, this is nothing more than reflected glory. While the media generically uses the label "AI" in their coverage of our field, the headline advances have been in machine learning (ML), and more specifically, in the field of neural networks. Advances in algorithms for training neural nets, new neural network structures (particularly convolutional neural nets, generative networks, and adversarial networks), coupled with the availability of rich, carefully curated data sets and the availability of cheap computer power for training neural nets, have made possible systems that seemed firmly out of reach just two decades ago.

So what does it all mean? Have we really found a silver bullet for AI? What are the limitations of the new techniques? And above all: *Where is it all going?* No veteran AI researcher can have failed to have mulled over these questions repeatedly since everything went crazy, back around about 2012 [1]. And it is these questions that the present book aims to address.

The authors will be well-known to many in the AI community. Ernest Davis, a professor at New York University, has a long research track record in commonsense reasoning, knowledge representation, and the philosophical foundations of AI. Gary Marcus is an author and entrepreneur, formerly a psychology professor at New York University, who studied under Steven Pinker. Marcus has published a string of books on topics relating to AI, perhaps most notably his 2003 book

E-mail address: mjw@cs.ox.ac.uk.

https://doi.org/10.1016/j.artint.2020.103386

The Algebraic Mind, in which he addresses the relationship between symbolic reasoning, and connectionism [2]. *The Algebraic Mind* attempts a reconciliation of these two perspectives — a theme which resonates with the current volume.

Rebooting AI is structured into eight chapters, and weighs in at just 200 pages. The book wastes no time in establishing it's first main thesis, in Chapter 1 ("Mind the Gap"): while we have, in a meaningful sense, made progress with *some aspects* of AI, that progress has in fact been limited to very narrow tasks only, and "real" AI requires much more than that:

> 'The central problem, in a word: current AI is *narrow*; it works for particular tasks that it is programmed for, provided that what it encounters isn't too different from what it has experienced before. That's fine for a board game like Go — the rules haven't changed for 2,500 years — but less promising in most real-world situations. Taking AI to the next level will require us to invent machines with substantially more flexibility.' (p. 13)
> 'What is missing from the field today [...] is *broad* (or "general") intelligence.' (pp. 15–16)

The narrowness, it is suggested, is an inherent, unavoidable consequence of the very techniques that have driven the current AI rejuvenation: data driven machine learning techniques. The authors identify what they call the "*AI chasm*" – the gap between what AI has actually achieved (success with very narrow and often crisply defined problems, backed up by large sets of training data), and the overinflated beliefs and expectations about it (mistakenly believing that success with these tasks implies some broader intelligence is at work). They point to the problems caused by this gap: for example, the naive trust placed in smart cruise control systems like Tesla's Autopilot, leading to at least one fatal crash.

Chapter 2 ("What's at Stake") focusses on the potential risks raised by data-driven narrow AI. The authors appear unconcerned about Skynet-style robot takeovers and potential existential risks raised by AI (quite rightly, in my opinion). They focus instead on the more imminent (and more realistic) concerns raised by AI: issues such as bias in training data, and more generally the issue of communicating our desires to machines that lack an understanding of the norms and conventions of human life and everyday human communication.

Chapter 3 ("Deep learning and beyond") presents a very readable summary of the evolution of deep learning, and how it has come to be so successful, and gives a critique of the approach, with respect to the narrowness argument. They identify three fundamental limitations of current deep learning techniques: greed, opacity, and brittleness. Deep learning is greedy in the sense that huge amounts of data are required to make it work. DeepMind's celebrated AlphaGo system required 30 million games to reach superhuman performance. Humans, by contrast, can learn basic competences very quickly. Opacity relates to the well-known problem that a neural network ultimately amounts to a (long) list of numeric weights. We have no way of recovering the expertise embodied within such a network net. Finally, brittleness means that neural networks can fail in surprising and (more problematically) unpredictable ways. Famously, image classification networks can be fooled into misclassifying pictures in a way that no human would ever be (a turtle identified as a rifle, for example). Not so much of an issue when classifying your vacation photos; much more problematic for driverless cars trying to make sense of road signs. Deep learning, the authors conclude, really isn't deep at all:

> '"Deep" ... doesn't mean that the system has learned anything particularly conceptually rich about the data it has seen. ... At best, deep learning is a kind of idiot savant, with miraculous perceptual abilities, but very little overall comprehension.' (pp. 62–64)

Chapter 4 ("If computers are so smart, how come they can't read?") looks at how deep learning copes with language comprehension. The chapter presents a detailed analysis of just how difficult it is for computers to be able to correctly interpret even the most simple everyday communications. The authors conclude that purely data driven approaches are not likely to lead to effective language understanding:

> 'Virtually every sentence that we encounter requires that we make inferences about how a broad range of background knowledge interrelates with what we read. Deep learning lacks a direct way of representing that knowledge, let along performing inferences over it in the context of understanding a sentence. ... Statistics are no substitute for real-world understanding.' (pp. 88–90)

Chapter 5 ("Where's Rosie?") moves in a slightly different direction to the previous chapters, pausing on the discussion of deep learning, and focussing instead on robotics. Again, the purpose of the chapter is essentially to systematically point out how fundamentally unsolved many of the challenges are in obtaining general purpose robotics. They conclude:

> '[R]oboticists have done an excellent job of getting robots to figure out where they are, and a fairly good job of figuring out how to get robots to perform individual behaviours.
> But the field has made much less progress in ... assessing situations, predicting the probable future, and deciding, dynamically, as situations change, which of the many possible actions makes the most sense in a given environment.' (p. 113)

The main conclusion the authors draw is that general purpose intelligent robots inhabiting our world will require 'rich cognitive models and deep understanding' (p. 114).

Chapter 6 ("Insights from the human mind") sets out some key principles that the authors believe will be required for successful general purpose intelligence. In this respect, while previous chapters are for the most part review and analysis, this chapter sets out 11 key principles that can be interpreted as the authors' manifesto for progress towards general AI. Probably the most important, and I believe least contentious of these, is the first: *there are no silver bullets*. That is, a simple dogmatic insistence that one crucial idea will take us to general AI is almost certainly mistaken.

> 'Truly intelligent and flexible systems are likely to be full of complexity, much like brains. Any theory that proposes to reduce intelligence down to a single principle ... is bound to be barking up the wrong tree.' (p. 119)

The remainder of the principles argue for the role of representations; the importance of the ability to abstract and generalise; the importance of structured representations; the role of theories; and the necessity of being able to understand causal relations.

Chapter 7 ("Common sense, and the path to deep understanding") explores the role of reasoning in AI. The authors acknowledge the limitations of previous attempts to formally capture common-sense reasoning (particularly logic-based approaches), but they argue that, nevertheless, the ability of a system to reason about its environment in a common sense way is essential. They conclude with a summary of their manifesto for general AI:

> 'Our recipe for achieving common sense, and ultimately general intelligence, is this: Start by developing systems that can represent the core frameworks of human knowledge: time, space, causality, basic knowledge of physical objects and their interactions, basic knowledge of humans and *their* interactions. ... Develop powerful reasoning techniques ... Connect these to perceptions, manipulation, and language. Use these to build rich cognitive models of the world. Then ... construct a kind of human-inspired learning system that uses all the knowledge and cognitive abilities that the AI has; that incorporates what it learns into its prior knowledge; and that, like a child, voraciously learns.' (pp. 178–179)

Finally, Chapter 8 ("Trust") turns to the issues of the extent to which we will be safe with emerging AI technologies, and how things might go wrong. It touches on issues such as Bostrom's famous paperclip scenario, and again argues that capabilities such as common sense reasoning will be required for trustworthy AI.

Overall, *Rebooting AI* has three main themes:

1. The first is a compelling, comprehensive exposition on what generality in AI means, and why it will be essential if we are ever to have intelligent machines that can safely co-exist with us.
2. The second is a lucid and detailed critique of data-driven approaches to AI, and in particular why such techniques are unlikely to achieve general AI.
3. The third is an argument for common sense reasoning, the role of knowledge and representation, and as a manifesto for progress towards general AI.

Where I believe the book succeeds most compellingly is in carefully demonstrating just what would be required to achieve general intelligence. Indeed, this book is, I believe, the best exposition on what general intelligence would mean and what capabilities it would require that I have read. A machine that could pass the many tests that the authors set out must, I think, be a pretty good candidate for general AI. If you ever find your students getting starry eyed and over excited about some amazing new game playing AI, then I urge you to point them at this book.

The book also serves tremendously well as a sober assessment of what data driven ML is likely to capable of — and what it isn't. While this territory will be more familiar to an AI audience (issues such as opacity and brittleness in deep learning), the book does an excellent job in explaining these limitations to a lay audience.

While I am convinced by the requirements that the authors set out for general AI in Chapter 7, I suspect they will have an uphill task to convince the newest generation of AI researchers with respect to some of the arguments in Chapter 8. It seems obvious (to me, at least) that knowledge and representation must have a role in general AI, and that the new AI must *somehow* connect with these to progress beyond the narrow tasks that it is currently proving so successful in. But how this will happen — and whether we will end up using the techniques and tools developed within symbolic AI — is much less clear.

Many popular science books on AI have appeared over the past few years: this is one of the best. It is eminently readable, extremely entertaining, and packed with thoughtful and original insights. Although aimed at a general audience, AI experts will nevertheless have much to learn from it. Above all, the book does an outstanding job of demonstrating just how rich human intelligence is, and just how far we are from achieving it.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks, in: Peter L. Bartlett, Fernando C.N. Pereira, Christopher J.C. Burges, Léon Bottou, Kilian Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012, Proceedings of a Meeting Held December 3–6, 2012, Lake Tahoe, Nevada, United States, 2012, pp. 1106–1114.
[2] G. Marcus, The Algebraic Mind: Integrating Connectionism and Cognitive Science, MIT Press, 2003.