BOOK REVIEW

Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way, Virginia Dignum (2019) Springer Nature, Cham, Switzerland, 129pp., £29 (hardback) ISBN 9783030303709

The first decades of the twenty-first century are marked by artificial intelligence entering the mainstream of public attention. The question is not whether AI will impact various domains of human interest and endeavour, but how AI will impact them and how it should do so.

Three emphases emerged as national policies about AI developed in the second decade of the century. In China, AI for government control; in the US, AI for commercial gain; and in Europe, what is termed 'responsible AI' (Dutton, 2018; Ding, 2018). This book is a monograph authored by one of the leading figures in the European responsible AI movement. As we might expect, the book opens with a summary of AI's rise and a description of what AI is and is not. The author, Virginia Dignum, agrees with other leading voices that AI is more than machine learning. AI is defined as 'artefacts that perceive the environment and take actions that maximise their chance of success at some goal' (p.3).

The media view of AI is quickly de-bunked:

> Contrary to what some may want us to believe, the algorithms used by AI are not a magic wand that gives their users powers of omniscience or the ability to achieve any and everything. (p.102)

Nor, as the book treats briefly in its final chapter, is the rise of evil world-ruling AI robots imminent. Plenty of excellent layman's books on AI have been written by expert AI researchers, birthed out of necessity and opportunity with societal awakening to AI. Dignum's book is not one of these. It has a more academic flavour, and is published by Springer Nature, a major academic publisher of computer science. *Responsible Artificial Intelligence* is Dignum's second book, following on from the tongue-twisting *Handbook of Research on Multi-Agent Systems: Semantics and Dynamics of Organizational Models*, which she edited. Dignum's first book would be inaccessible techno-magic for the general public: her second book, while more accessible, remains academic in tone.

Dignum held a position at Delft University of Technology for many years, researching multi-agent systems, AI and values. The book follows upon her move to a professorship at Umeå University in Sweden, where she holds a chair in AI and society. Dignum sits on the European Commission's high level expert group in artificial intelligence, and is a member of several other prominent institutions. That the author is well qualified and that *Responsible Artificial Intelligence* is timely there is no doubt. That the book also fits with the TU Delft agenda on AI is also clear. The university – the oldest, largest and most highly ranked technical university in the Netherlands – puts it thus:

> AI research and education at Delft University of Technology focuses on understanding, designing, and engineering the responsible automation of complex systems, involving people as well as technical components. (Delft University of Technology, 2020)

This emphasis on responsible engineering is the mantra of Dignum's book. Indeed, the subtitle is *How to Develop and Use AI in a Responsible Way*. To reach this end, the book's opening chapter tells us what AI is, from both computer science and philosophy perspectives. Dignum dwells on AI as autonomy, adaptability and interaction; she helpfully distinguishes among AI, machine learning, and data analytics. In presenting the main ethical theories, she argues that responsible AI involves ethics in various ways that the book later develops. Dignum also presents a summary

of value ethics – yes, the trolley problem receives a mention – and an outline of the computational steps, at a high level, an AI agent with ethical reasoning must undertake.

With the reader informed about what AI and ethical reasoning are, and with the book's motivation well illustrated, the book moves on to the author's ART (accountability, responsibility, transparency) framework. Dignum has published the framework before and has published extensively about value-sensitive design and ethics in AI. *Responsible Artificial Intelligence* devotes two chapters to these topics and then a chapter to the governance of AI.

But what is *responsible* AI? Perhaps the concept is as slippery to define as AI itself. Dignum (p.93) concedes that responsible AI means different things to different people, serving as an overall container for many diverse opinions and topics. Depending on the speaker and on the context, it can mean any one of the following:

- policies concerning the governance of R&D activities and the deployment and use of AI in societal settings,
- the role of developers, at individual and collective level,
- issues of inclusion, diversity and universal access, and
- predictions and reflections on the benefits and risks of AI.

Nonetheless, we are told that 'Responsible AI is thus about being responsible for the power that AI brings' (p.2); that it is 'the development of intelligent systems according to fundamental human principles and values' (p. 6); and, notably for the thrust of the book, that (p.48):

> Responsible artificial intelligence is concerned with the fact that decisions and actions taken by intelligent autonomous systems have consequences that can be seen as being of an ethical nature. These consequences are real and important, independently of whether the AI system itself is able to reason about ethics or not. As such, responsible AI provides directions for action and can maybe best be seen as a code of behaviour – for AI systems, but, most importantly, for us. (p.48)

This is arguably the book's central point – that AI systems should be designed, developed and deployed responsibly in the context of human societies. Who would disagree with this? Thus, 'Responsible artificial intelligence is about human responsibility for the development of intelligent systems along fundamental human principles and values, to ensure human flourishing and well-being in a sustainable world' (p. 119). 'Responsible AI is not about the characteristics of AI systems, but about our own role' (p.7). The assertion that 'At the core of AI development should lie the idea of "AI for good" and "AI for all"' (p.48) aligns with the European view of AI (European Commission, 2020).

Dignum distinguishes responsible AI from AI ethics:

> Ethics is the study of morals and values, while responsibility is the practical application of not only ethical concerns but also legal, economical and cultural ones to decide what benefits society as a whole. So, while with ethics, it suffices to observe what happens, responsible AI demands action. (p.6)

The book returns often to this point about our responsibility to act in order to ensure AI systems are responsible. If, then, 'AI systems are artefacts decided upon, designed, implemented and used by people', it follows that people are responsible for them. Further, and rising above the AI arms race, Dignum echoes Bostrom and Yudkowsky (2014):

> In reality, no intelligent artefact – however advanced and sophisticated – should be called 'autonomous' in the original philosophical sense, and therefore it can never be accorded the same form of moral standing as a human being nor inherit human dignity. (p.90)

With action for responsibility in mind, chapters 4 to 6 address in turn processes to design, develop, deploy and use AI; how to deal with ethical reasoning by the AI systems themselves; and mechanisms that can ensure that all involved take 'the responsible route'. So, we have the author's ART framework, the ethical and moral reasoning of AI systems (and an interesting discussion about the ethical and legal status of such systems, granted they are only artefacts), and the policy and governance of, and for, responsible AI in society.

It is worth quoting some extracts about the governance issues:

> Ensuring responsible AI is however more than setting up lists of desired principles, standards or recommendations (and there are a worthy set of these now). It requires action. Possible mechanisms for this action are regulation, certification and codes of conduct. (p.97)
>
> This is what responsible AI is about, the decisions taken concerning the scope, the rules and the resources that are used to develop, deploy and use AI systems. AI is not just the algorithm, or the data that it uses. It is a complex combination of decisions, opportunities and resources. (p.102)

Hence the conclusion of the book is already in the preface: 'We are all responsible for responsible AI' (p.vi). The details of how, in practice, we are to take action are not provided – and can hardly be expected from a book about principles. However, a set of case studies would have illustrated the principles espoused.

A second difficulty with *Responsible Artificial Intelligence* is the audience. The book talks about researchers, developers and users of AI systems, but who will actually read a high-level academic book like this? And who will identify with the relative responsibility (p.103)? Part of the public debate on AI is about the societal effects of the technology step change AI offers to many socio-economic sectors. The public – or at least the media – worries about killer robots and mass job losses, and in more sensible discussions, about algorithmic transparency and data privacy. Dignum does not predict what AI can do by the year 2050, or anything like that, but analyses the forces that will shape the societal impact of AI systems. The last chapter of the book connects the main discourse with the media view of AI. Dignum mentions the future of jobs and education, the societal risks of AI (foremost among these are risks to safety, democracy and human dignity) and superintelligence. Walsh (2018) provides a worthy and more popular treatment of these topics by another experienced AI researcher.

Responsible Artificial Intelligence is a valuable contribution to the debate about AI: not at a conceptual or futuristic level, but at the level of building principled, responsible AI systems, and the use of these systems. The further reading which the book suggests complements this technical monograph with accessible contributions about the nature and future of AI. If you want predictions about AI in 2050, read those books. If you want to know how a vision for responsible AI systems in the European fashion can be built from values, read this book.

References

Bostrom, N. and Yudkowsky, E. (2014) 'The ethics of artificial intelligence' in Frankish, K. and Ramsey, W. (eds) *Cambridge Handbook of Artificial Intelligence*, Cambridge University Press, Cambridge, pp.316–34.

Delft University of Technology (2020) 'AI, data and digitalization at TU Delft', available at https://ai.tudelft.nl (accessed August 2020).

Ding, J. (2018) *Deciphering China's AI Dream*, technical report, Future of Humanity Institute, University of Oxford, Oxford, available at https://www.fhi.ox.ac.uk/publications/deciphering-chinas-ai-dream-jeffrey-ding-2018-future-of-humanity-institute-university (accessed August 2020).

Dutton, T. (2018) 'An overview of national AI strategies', *Medium*, 28 June, available at https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd (accessed August 2020).

European Commission (2020) 'Shaping Europe's digital future: Commission presents strategies for data and artificial intelligence', available at https://ec.europa.eu/commission/presscorner/detail/en/ip_20_273, 19 February (accessed August 2020).

Walsh, Toby (2018) *Machines that Think: The Future of Artificial Intelligence*, Prometheus Books, Amherst NY.

Neil Yorke-Smith
Delft University of Technology
n.yorke-smith@tudelft.nl