BOOK REVIEW

AI Ethics by Mark Coeckelbergh (2020) MIT Press, Cambridge MA, 248pp., $US15.95 (paperback) ISBN: 9780262538190

In the late 1980s, I got into artificial intelligence (AI) in the final stages of my study at Twente University. This was the era of symbolic AI, and although some people had ambitious ideas and projects (Douglas Lenat's cyc project, for instance), many were engaged with developing mundane rule-based expert systems (like me). Others were finding out that many things we humans take for granted (like moving around in a room without bumping into things or picking up an egg without breaking it) are actually hard for a machine. There was much enthusiasm at the time about the prospect of artificial intelligence. This was the era where computing became affordable and slowly more powerful, albeit that computer memory for most researchers was measured in kilobytes and disk space in megabytes. The second AI Spring would surely evolve into a bright AI Summer.

Throughout the history of AI, philosophers have been part of the debate. Originating with the question of whether machines can think (e.g., Turing, 1950), many philosophers, psychologists, computer scientists, cognitive scientists and others have wondered what intelligence is and what AI could aspire to be (Boden, 2016). Questions such as whether decisions made by AI systems would be fair and unbiased were, as far as I can remember, not part of the debate.

And then the second AI Winter came. AI could, again, not deliver on its promises and funding dried up. Until the 2010s. Computing machinery, thanks to Moore's Law, had made rapid progress, storage space became abundant thanks to Kryder's Law and the internet super-connected everything. Data became a resource, became Big Data, both enabling and necessitating AI. Only with sufficient intelligence in the machinery could these huge data resources be brought to bear and monetized. Big Data made way for Data Science and nowadays AI is the new catch-all term for everything involving large amounts of data and machines. AI is fashionable and everything seems to be AI enabled (just like many products around 2006–7 contained nanoparticles because that sounded cool at the time).

Winter is a'coming in?

Controversy over whether machines can have intelligence, whether they will take over jobs and even humanity now seems a lot more pressing and realistic than it was in the late 1980s. An entirely new set of ethical questions has emerged and has made its way into policy circles and even into the curricula of data science and AI students. FAT (now ACM FAccT) is short for fairness, accountability and transparency and has become a mantra in the AI business. AI systems monitor our everyday lives – even through cameras in billboards – assess our behaviour and base decisions on algorithms that have been trained by large numbers of cases. As a result, Amazon recommends what we are likely to buy, Google purports to know what we're looking for and the police identify culprits on the basis of vague surveillance camera images. How do these automated decisions affect our lives? Are they fair and unbiased? Many people, both within the AI community and beyond, have genuine concerns whether AI is developing in the right direction. And if these concerns are treated inadequately, we may be steering towards a third AI Winter.

The book

AI Ethics is a welcome, accessible contribution to a growing literature and essential reading for anyone who wants a brief overview of the history and future of AI and the ethical challenges it raises. The book is part of the MIT Press Essential Knowledge series, which aims at delivering

'expert overviews of subjects that range from the cultural and the historical to the scientific and the technical'. It is set in a small pocket book format (13 × 18 cm) with some 248 pages, including sources, further reading and a glossary. The audience for the series, I assume, is the general (educated) reader and so the book has to balance accessibility with what we expect from an academic book, thoroughly annotated and referenced. The book does reasonably well in this respect. It contains a fair number of references and thus starting points for exploring further the various topics addressed. To me it feels a bit unsatisfactory not to have references to all (or most) of the many examples mentioned in the book. For instance, in passing, Coeckelbergh mentions that 'already (non-AI) drones can handicap a big London airport'. The scholar in me would want to have a reference to the Gatwick drone incident between 18 and 21 December 2018.

Mark Coeckelbergh is well positioned to write an overview of AI ethics. He is a professor of philosophy of media and technology at the University of Vienna and has been active in the field of robotics and AI for a significant period. He is a member of various ethics advisory boards for robotics and AI as well as policy advisory boards, such as the EU high-level expert group on artificial intelligence. Producing a book on AI ethics is challenging because AI is a difficult phenomenon, being simultaneously the science of studying the development of human-like intelligence, a set of techniques to create and program software, the product of such programming and a buzzword for anything that needs to sound fancy. There is consequently a lot of ground to cover. Another reason why producing the overview is challenging is that the range of philosophical and ethical aspects to cover range from the relatively well known and mundane, such as privacy concerns, to such questions as whether AI will transcend humans (Kurzweil, 2005).

Coeckelbergh takes us through the history of AI and the landscape in which it develops. He pays attention to the big philosophical questions, as well as to the smaller practical ones. He mentions and treats many of the relevant (open) questions, provides answers when they are available, outlines what policy initiatives exist to steer AI in the right direction, gives policy advice and raises questions that should be of major concern to policy makers and developers. And, not least, he provides a nuanced view of the potential and limitations of AI. He positions AI ethics as an important field while noting that there are more pressing issues to be resolved than AI ethics (the climate crisis and the effects of the Anthropocene are clear examples). The book generally maintains a careful balance, showing that on the one hand AI does not raise entirely new concerns (e.g., all technologies have affected the labour market and AI is just the newest incarnation) while on the other hand that there is indeed something new under the sun (e.g., if humans don't understand what happens under the hood of AI systems, then how does this affect such concepts as dignity?).

A walk through

The first chapter shows us that AI is not science fiction, but is already here; it argues that we need to look at ethical and societal problems surrounding AI developments. Instead of first describing what AI is or isn't, chapter 2 puts on the big pants and outlines the science fiction futures from past and present. It brings back the images of Frankenstein's monsters and takes us along the futurist projections of superintelligence and transhumanism to show that AI, at least according to some, is something other than your garden variety technology. AI, in the vision of Nick Bostrom and Ray Kurzweil, will surpass human intelligence and may cause existential risks for humanity. After all, why would AI care about human goals?

What Coeckelbergh does here is focus on one strand of AI, 'general AI', the kind that aims to mimic, or surpass, full human capacity and that, according to many scientists in the field, is far away. Whether strong AI is indeed possible, the concept raises a number of philosophical and ethical questions that are addressed throughout the book. For instance, what does machine intelligence mean? What does it mean to be human? Should/could AIs receive legal status (personhood, for instance)? Who is responsible for AI's behaviour, particularly in cases where many hands (artificial minds) are involved?

Throughout the book, Coeckelbergh tries, and often succeeds, to provide a multicultural perspective on issues. For instance, he contrasts the dystopian and negative images surrounding AI and robotics prevalent in the West (e.g., Frankenstein) with the Japanese attitude towards robots. In the Japanese animistic way of thinking, AIs can, at least in principle, have spirit or soul and can be sacred. Raising awareness among readers that AI and robot ethics are not universal, even in a globalised world, is important. Chapter 3 considers in further detail whether general (or strong) AI is possible and what distinguishes humans from machines (if anything). The well-known works of John Searle (for instance, his Chinese room experiment) and Hubert Dreyfus are introduced and placed in current context of post-humanism, and post-phenomenology. For non-philosophers this chapter will provide an interesting introduction to the position of humans versus machines.

Still without knowing exactly what we're dealing with, chapter 4 tackles a number of the big questions from the general AI debates. What is the moral status of AI? The chapter covers a lot of ground, not only on an abstract level, but also exploring the trolley dilemmas (a set of situations involving a run-away trolley (tram) on its way to kill one or more people inexplicably tied down on the tracks). It is for you, moral agent, to decide how many die by flipping levers, pushing overweight people from bridges and so on to change the course of the trolley. These dilemmas are used to teach people moral reasoning. Coeckelbergh makes a clear case that trolley dilemmas have practical bearing for self-driving vehicles. We may have to embed some form of functional morality in these cars to make them cope in a humanly acceptable way with choices between bad and worse. He also uses this and other examples to show there are no right and wrong moral answers; ethics is not a checklist.

What does AI stand for, really?

Chapter 5 finally takes us deeper into the technology. After the futuristic account of AI as the potential new overlords and reasons why we should not worry too much about their rule, we now learn that there is much AI that does not qualify as strong AI. Coeckelbergh takes us through the different forms of AI, from narrow AI to embodied AI (robots) to softbots. He also introduces the reader (briefly) to the different AI paradigms (symbolic, connectionism and its subfields, such as machine learning). The technology is further explored in chapter 6, which talks about machine learning, data and data science as the most important components of current narrow AI developments and applications.

On to ethics

Chapter 7 picks the low hanging fruit in terms of ethics concerns, such as privacy, manipulation, exploitation, vulnerable users, fake news, safety and security. Many of the issues are not new, as Coeckelbergh points out, but there are quantitative and qualitative differences resulting from AI entering the field. For instance, video manipulation is as old as the video camera, but new AI tools make it easy for any script kiddie to manipulate videos to show politicians saying what they would never say in real life. Changing the outcome of elections is no longer far-fetched.

The book moves on to one of the core fields of AI ethics (in my view as a law and technology scholar) – responsibility and explainability of AI. How to allocate responsibility in situations where AI does all the work, as with the self-driving vehicle? Who is responsible (or liable) for damages? This is one of the areas where lawyers and ethicists come together. Again, the issues are not entirely novel; we have experience with children and animals who are not considered moral agents. In such cases, handlers and custodians are responsible for the actions of these non-moral agents. Should the same hold for AIs? Coeckelbergh seems to hesitate here if only because the proposed responsible human doesn't know what happens within AI. But the same holds true for my cat. I have no idea what she is up to most of the time, though I am certain she likes chasing (and eating) mice. Still, I am responsible (liable at least) for her actions, even though often I don't have control over her.

The fact that we generally don't know what happens within AI, the so-called black-box problem, also brings legal scholars and ethicists together. In many situations, we demand or have a right to an explanation when decisions are taken about us or affecting us. This is one of the thorny areas in AI. Explaining what the neural network in an algorithm used for the decision looks like is obviously unhelpful. Far less clear is what would do as an explanation. Coeckelbergh addresses a number of the legal and ethical questions and rightly points out that AI also triggers us to explore what we mean by explanation with or without AI. Thinking about what to expect from AI may teach us about non-AI phenomena as well.

Chapter 9 addresses two other important ethical concerns: (1) bias in machine learning and what this means for decision-making, and (2) what AI will do to jobs and ultimately to the meaning of life. In the section about bias, Coeckelbergh gives a comprehensive overview of the issues regarding training sets for machine learning and how these can lead to unfair treatment of certain groups, such as poor and non-white people, in predictive policing tools. These are based on historic data and so police resources are directed to neighbourhoods predominantly populated by poor and non-white people, increasing the likelihood that people in these neighbourhoods will be stopped more than those in richer neighbourhoods. This may further skew the data, while the crime rates may actually be much the same in the two neighbourhoods. Bias in data sets is a complex topic and much research is allocated to understanding the issues.

The second topic of this chapter takes us back to one of the large-scale concerns. If AIs replace us in the workplace, then what place is there for humans? Again, not a new problem but one that resurfaces with every major new technology (steam engine, computer, robot). Maybe I am too optimistic, but we have always adapted to new job markets. There are hardly any steam engine drivers left; they have found other work.

What to do?

Chapters 10 and 11 take a closer look at Coeckelbergh's more applied work in advisory boards and policy councils. How should society cope with undesirable effects of AI developments? This is where policy and regulation come into view. Chapter 10 discusses a number of the policy initiatives that have been taken in the last five to ten years. Coeckelbergh looks at national, international, supra-national initiatives, but also at industry and civil society initiatives and discusses the core of some of these proposals and frameworks. He calls attention to the many questions in this area, such as what to regulate, when and by whom. One of the most prominent questions is whether regulation needs to be enacted, or whether ethical principles and self-governance are the way forward. Chapter 11 provides a number of challenges for policy makers. One of these is how to embed ethics into design and use of AI. Ethics, like law, is seen as hampering innovation. To some extent, they are both constraints, but then design freedom is also constrained by gravity.

On balance

This book provides a balanced overview of many topics and outlines many of the important questions in the field of AI ethics. Mark Coeckelbergh knows what he is talking about and is able to explain the core ideas to an audience of non-experts. The book provides valuable starting points for reflection and further study. All this I like. What I like less is that, although the table of contents suggests a clear structure of the text and the chapters do follow the structure, each chapter jumps back and forth through the topics. It feels a bit like the chapters are essays, rather than part of a bigger narrative.

While I completely subscribe to Coeckelbergh's call to arms to take the ethical questions seriously in education, policy making, development and use of AI (I, myself, have taught ethics to data science students), I find the underpinning of this relatively weak. He argues that ethics provides barriers to certain developments, but that these barriers are necessary. He also tries to present a case

for positive ethics, developing a vision of the good life and good society and how AI can contribute to these. But instead of providing compelling examples of why both negative and positive ethics are essential to prevent a new AI Winter, Coeckelbergh's argument is basically, trust me, you need this. This will hardly convince the sceptics.

A final remark concerns the closing chapter, provocatively called 'It's the climate, stupid! On priorities, the Anthropocene, and Elon Musk's car in space'. Here Coeckelbergh sees AI as a significant socio-technical system in the context of the big societal challenges we currently face. He rightly points out that ethical issues in AI are small beer compared with the challenges that global warming presents and that we should get our priorities right. He also shows that AI can affect the grand challenges both negatively and positively. It can increase existing inequalities among people and disadvantage those already hard hit by climate change. But AI can also help build a smart electricity grid with fewer energy losses. The message in the chapter is sound, but Coeckelbergh sounds more like an activist than a scientist. I hope this does not obscure the message for the mainstream reader.

References

Boden, M. (2016) *AI: Its Nature and its Future*, Oxford University Press, Oxford.

Kurzweil, R. (2005) *The Singularity is Near*, Viking, New York.

Turing, Alan (1950) 'Computing machinery and intelligence', *Mind*, 49, 236, pp.433–60.

Ronald Leenes
Tilburg Institute for Law, Technology, and Society (TILT)
Tilburg University, Netherlands
R.E.Leenes@tilburguniversity.edu