Artificial Intelligence: Foundations, Theory, and Algorithms

Virginia Dignum

Responsible Artificial Intelligence

How to Develop and Use AI in a Responsible Way

Springer

Artificial Intelligence: Foundations, Theory, and Algorithms

Series editors
Barry O'Sullivan, Cork, Ireland
Michael Wooldridge, Oxford, United Kingdom

More information about this series at http://www.springer.com/series/13900

Virginia Dignum

Responsible Artificial Intelligence

How to Develop and Use AI in a Responsible Way

Springer

Virginia Dignum
Department of Computing Science
Umeå University
Umeå, Sweden

ISSN 2365-3051 ISSN 2365-306X (electronic)
Artificial Intelligence: Foundations, Theory, and Algorithms
ISBN 978-3-030-30370-9 ISBN 978-3-030-30371-6 (eBook)
https://doi.org/10.1007/978-3-030-30371-6

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The last few years have seen a huge growth in the capabilities and applications of Artificial Intelligence (AI). Hardly a day goes by without news about technological advances and the societal impact of the use of AI. Not only are there large expectations of AI's potential to help to solve many current problems and to support the well-being of all, but also concerns are growing about the role of AI in increased inequality, job losses and warfare, to mention a few.

As Norbert Wiener said already in 1960, as often quoted by Stuart Russell: "[W]e had better be quite sure that the purpose put into the machine is the purpose which we really desire". But what is this purpose, and who are those addressed by the pronoun 'we'? In my view, we refers to us all: researchers, developers, manufacturers, providers, policymakers, users and all who are directly and indirectly affected by AI systems. We all have different responsibilities, but we all have the right, and the duty, to be involved in the discussion of the purpose we want AI technology to have in our lives, our societies and our planet because AI and its impact are too important to be left to the technocrats alone.

This means that we all need to understand what AI is, what AI is not, what it can do, and most importantly, what we can do to ensure a positive use of AI, in ways that contribute to human and environmental well-being and that are aligned with our values, principles and priorities.

Moreover, we need to ensure that we put in place the social and technical constructs that ensure that responsibility and trust for the systems we develop and use in contexts that change and evolve. Obviously, the AI applications are not responsible, it is the socio-technical system of which the applications are part of that must bear responsibility and ensure trust. Ensuring ethically aligned AI systems requires more than designing systems whose result can be trusted. It is about the way we design them, why we design them, and who is involved in designing them. This is work always in progress. Obviously, errors will be made, disasters will happen. More than assigning blame for these failures, we need to learn from them and try again, try better.

It is not an option to ignore our responsibility. AI systems are artefacts decided upon, designed, implemented and used by people. We, people, are responsible. We are responsible to try again when we fail (and we will fail), to observe and denounce when we see things going wrong (and they will go wrong), we are responsible to be informed and to inform, to rebuild and improve.

This book aims at providing an overview of these issues at undergraduate level and for readers of different backgrounds, not necessarily technical. I hope that you find its contents useful, because there is work to be done to ensure that AI systems are trustworthy and those who develop and use them do so responsibly. And we (people) are the ones who can and must do it. We are all responsible for Responsible AI.

This book would not have been possible without the invaluable discussions I've had with colleagues, friends and participants at the many events where I've spoken. Their questions, ideas and, in many cases, divergent ideas have been a main source of inspiration for my work. It is therefore not possible to list here everybody I would like to thank. However, I would like to say a special thanks to Catholijn Jonker, Jeroen van den Hoven, and all my past and current PhD students and postdocs. I also thank Michael Sardelić Winikoff and Francesca Rossi, for their careful and critical review of this manuscript. Without them this book would not have been possible. Finally, a special thanks to Frank, always.

Virginia Dignum
May 2019

Contents

Chapter 1
Introduction

> "As the use and impact of autonomous and intelligent systems (A/IS) become pervasive, we need to establish societal and policy guidelines in order for such systems to remain human-centric, serving humanity's values and ethical principles."
>
> The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

Where we introduce Responsible Artificial Intelligence and discuss why that is important.

As advances in Artificial Intelligence (AI) occur at a rapid pace, there is a growing need for us to explore and understand what impact these will have on society. Policymakers, opinion leaders, researchers and the general public have many questions. How are biases affecting automated decision-making? How is AI impacting jobs and the global economy? Can, and should, self-driving cars make moral decisions? What should be the ethical, legal and social position of robots?

Many are also worried about the consequences of increasing access by government, corporations and other organisations to data that enables extensive and intrusive predictions concerning citizen behaviour.

The underlying concern in all these questions is: Who or what is responsible for decisions and actions by AI systems? Can a machine be held accountable for its actions? What is our role as we research, design, build, sell, buy and use these systems? Answering these and related questions requires a whole new understanding of socio-technical interactions, the ethical aspects of intelligent systems, and the novel mechanisms for control and autonomy of AI systems.

V. Dignum, *Responsible Artificial Intelligence*, Artificial Intelligence: Foundations, Theory, and Algorithms, https://doi.org/10.1007/978-3-030-30371-6_1

This book is not about the future. It does not present scenarios of doom nor visions of heaven on earth. It also does not focus on super-intelligence, singularity or the other potential areas of AI. Instead, this book is about the present. In particular, it is about responsibility: our responsibility for the systems we create and use, and about how, and whether, we can embed responsibility into these systems. It is also about the accountability and transparency mechanisms that can support taking responsibility.

This book aims to introduce a responsible approach to AI design, development and use. One that is centred on human well-being and that aligns with societal values and ethical principles. AI concerns all of us, and impacts all of us, not only individually but also collectively. We thus need to go further than the analysis of benefits and impacts for individual users, but rather to consider AI systems as part of an increasingly complex socio-technical reality.

Responsible AI is thus about being responsible for the power that AI brings. If we are developing artefacts to act with some autonomy, then *"we had better be quite sure that the purpose put into the machine is the purpose which we really desire"*. (Stuart Russell quoting Norbert Wiener in [103]). The main challenge is to determine what responsibility means, who is responsible, for what, and who decides that. But given that AI systems are artefacts, tools built for a given purpose, responsibility can never lie with the AI system because as an artefact, it cannot be seen as a responsible actor [26]. Even if a system's behaviour cannot always be anticipated by designers or deployers, chains of responsibility are needed that can link the system's behaviour to the responsible actors. It is true that some, notably the European Parliament[1], have argued for some type of legal personhood for AI systems. However, these suggestions are more guided by a science-fiction-like extrapolation of current expectations on AI capabilities than by scientific truth. Moreover, AI systems operate on behalf of or under the mandate of corporations and/or people, both of which already have legal personhood in many countries, which is sufficient to deal with potential legal issues around the actions and decisions of the AI systems they operate. We will discuss this issue later in this book.

For example, where lies the responsibility for a parole decision, for a medical diagnosis or for the refusal of a mortgage application, when these decisions are made by AI systems or based on the results provided by an AI system? Is the developer of the algorithm responsible, the providers of the data, the manufacturers of the sensors used to collect data, the legislator that authorised the use of such applications, or the user who accepted the machine's decision? Answering these questions and distributing responsibility correctly are no simple matters.

A new and more ambitious form of governance of AI systems is a most pressing need. One that ensures and monitors the chain of responsibility across all the actors. This is required to ensure that the advance of AI tech-

[1] http://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html?redirect

nology is aligned with societal good and human well-being. To this effect, policymakers need a proper understanding of the capabilities and limits of AI in order to determine how issues of accountability, responsibility and transparency should be regulated.

But what is AI? AI refers to artefacts that perceive the environment and take actions that maximise their chance of success at some goal [104]. The emphasis here is on the 'artificial' as the counterpart to *natural* intelligence, which is the product of biological evolution. Minsky defines AI as *"the science of making machines do things that would require intelligence if done by men"*. Or, according to Castelfranchi paraphrasing Doyle *"AI is the discipline aimed at understanding intelligent beings by constructing intelligent systems"* [44]. Indeed, one important reason to study AI is to help us better understand natural intelligence.

AI represents a concerted effort to understand the complexity of human experience in terms of information processes. It deals not only with how to represent and use complex and incomplete information logically but also with questions of how to see (vision), move (robotics), communicate (natural language, speech) and learn (memory, reasoning, classification).

Although the scientific discipline of Artificial Intelligence has been around since the 1950s, AI has only recently become a household term. However, in its current use, AI generally refers to the computational capability of interpreting huge amounts of information in order to make a decision, and is less concerned with understanding human intelligence, or the representation of knowledge and reasoning.

Within the AI discipline, Machine Learning is the broad field of science that deals with algorithms that allow a program to 'learn' based on data collected from previous experiences. Programmers do not need to write the code that dictates what actions or predictions the program will make based on a situation, but instead, the system takes appropriate action based on patterns and similarities it recognises from previous experiences.

AI systems use algorithms to reach their objectives, but AI is more than the algorithms it uses. An algorithm is nothing more than a set of instructions, such as computer code, that carries out some commands. As such, there is nothing mysterious about algorithms. The recipe you use to bake an apple pie is an algorithm: it gives you the instructions you need to achieve a result based on a bunch of inputs, in this case the ingredients. The end result of your apple pie is as much dependent on your skills as a baker, on the ingredients you choose, as it is on the algorithm itself. And, more importantly, never by itself will the apple pie recipe transform itself into an actual pie! The same holds for AI algorithms: the outcomes of an AI system are only partly determined by the algorithm. For the rest, it is your choice of data, deployment options and how it is tested and evaluated, amongst many other factors and decisions, that determine the end result.

Responsible AI thus means that besides choosing the proper algorithms, you also need to consider the ingredients (e.g. the data) to use and the com-

position of the team using it. To bake an apple pie, you have the choice between using organic apples or the cheapest ones on sale. You also can ask a starting cook or a star cook to bake it. The same holds for developing AI systems: which data are you using to train and to feed your algorithm? Does it take into account diversity and specific characteristics of the domain, or is it some set of training data that you downloaded for free from the Internet? And who is building and evaluating the system? A diverse and inclusive team that reflects the spectrum of stakeholders and users? Or the cheapest team you could put together and are you relying on poorly paid testers from Amazon Mechanical Turk to label your data? The choice is yours. The results will reflect those choices.

Responsible AI requires participation. That is, it requires the commitment of all stakeholders and the active inclusion of all of society. Which means that everybody should be able to get proper information about what AI is and what it can mean for them, and also to have access to education about AI and related technologies. It also means that AI researchers and developers must be aware of societal and individual implications of their work and understand how different people use and live with AI technologies across cultures. For this effect, the training of researchers and developers on the societal, ethical and legal impact of AI is essential to ensure the societal and ethical quality of the systems and the developer's awareness of their own responsibility where it concerns the development of AI systems with direct impact on society.

Looking solely at performance, AI seems to provide many advantages over naturally intelligent systems like humans. Compared to people, AI systems can generally make quicker decisions and operate at any time. They don't get tired or distracted and are more accurate than humans in those tasks they are built for. Moreover, software can be copied and does not need to be paid. On the other hand, there are many important advantages of natural intelligence. First, you don't need to go far to find it. There are billions of humans available and we don't need to 'build' them, we just need to educate them. The human brain is a miracle of energy efficiency, capable of managing a variety of skills and executing many different tasks at once, using only a fraction of the energy an artificial neural network uses to execute only one task. People are great at improvising and can handle situations they never encountered before in ways that we can only dream machines will ever do.

AI can help us in many ways: it can perform hard, dangerous or boring work for us; it can help us to save lives and cope with disasters; and it can entertain us and make each day more comfortable. In fact, AI is already changing our daily lives and mostly in ways that improve human health, safety and productivity. In the coming years we can expect a continuous increase of the use of AI systems in domains such as transportation, the service industries, healthcare, education, public safety and security, employment and workplace and entertainment[2].

[2] One Hundred Year Study on AI: https://ai100.stanford.edu/

It is easy to feel overwhelmed by these possibilities and the rapid pace of AI advances. Already, thought leaders and newspapers are voicing concerns about the potential risks and problems of AI technology[3]. Killer robots, privacy and security breaches, the impact of AI on labour and social equality[4], super-intelligence and existential risks[5] are ubiquitous in the media, making us wary about AI.

In reality, there are many reasons for optimism. According to the World Health Organisation, 1.35 million people die annually in traffic accidents, more than half of which are caused by human error[6]. Intelligent traffic infrastructures and autonomous vehicles can provide solace here. Even if these will inevitably still cause accidents and deaths, forecasts show they can significantly reduce overall casualties on the road. AI systems are also already being used to provide improved and earlier diagnostics for several types of cancer, to identify potential pandemics, to predict wildlife poaching and so improve ranger assignments, to facilitate communication by improved translation, or to optimise energy distribution.

We are ultimately responsible. As researchers and developers, we must make fundamental human values the basis of our design and implementation decisions. And as users and owners of AI systems, we must uphold a continuous chain of responsibility and trust for the actions and decisions of AI systems as they act in our society. Responsibility rests not only with those who develop, manufacture or deploy AI systems, but also with the governments that legislate about their introduction in different areas, educators, the social organisations providing awareness and critical assessment in their specific fields and all of us specifically to be aware of our rights and duties when interacting with these systems.

The ultimate aim of AI is not about the creation of superhuman machines or other sci-fi scenarios but about developing technology that supports and enhances human well-being in a sustainable environment for all. It is also about understanding and shaping technology as it becomes ever more present and influential in our daily lives. It's not about imitating humans, but providing humans with the tools and techniques to better realise their goals and ensure the well-being of all. From the perspective of its engineering roots, the focus of AI is on building artefacts. But it is more than engineering, it is human-centric and society-grounded. AI is therefore transdisciplinary, requiring not only technological advances but also contributions from the social sciences, law, economics, the cognitive sciences and the humanities.

---

[3] See e.g. http://observer.com/2015/08/stephen-hawking-elon-musk-and-bill-gates-warn-about-artificial-intelligence or http://www.theguardian.com/technology/2015/nov/05/robot-revolution-rise-machines-could-displace-third-of-uk-jobs

[4] http://www.express.co.uk/life-style/science-technology/640744/Jobless-Future-Robots-Artificial-Intelligence-Vivek-Wadhwa

[5] http://edition.cnn.com/2014/09/09/opinion/bostrom-machine-superintelligence/

[6] https://apps.who.int/iris/bitstream/handle/10665/276462/9789241565684-eng.pdf