Chapter 1
Introduction

Abstract The introductory chapter describes the motivation behind this book and provides a brief outline of the main argument. The book offers a novel categorisation of artificial intelligence that lends itself to a classification of ethical and human rights issues raised by AI technologies. It offers an ethical approach based on the concept of human flourishing. Following a review of currently discussed ways of addressing and mitigating ethical issues, the book analyses the metaphor of AI ecosystems. Taking the ecosystems metaphor seriously allows the identification of requirements that mitigation measures need to fulfil. On the basis of these requirements the book offers a set of recommendations that allow AI ecosystems to be shaped in ways that promote human flourishing.

Keywords Artificial intelligence · Ethics · Human flourishing · Mitigation strategies · Innovation ecosystem

Artificial intelligence (AI) raises ethical concerns. Such concerns need to be addressed. These two statements are not too contentious. What is less clear is what exactly constitutes the ethical concerns, why they are of an ethical nature, who should address them and how they are to be dealt with.

AI is increasingly ubiquitous and therefore the consequences of its use can be observed in many different aspects of life. AI has many positive effects and produces social benefits. Applications of AI can improve living conditions and health, facilitate justice, create wealth, bolster public safety and mitigate the impact of human activities on the environment and the climate (Montreal Declaration 2018). AI is a tool that can help people do their jobs faster and better, thereby creating many benefits. But, beyond this, AI can also facilitate new tasks, for example by analysing research data at an unprecedented scale, thereby creating the expectation of new scientific insights which can lead to benefits in all aspects of life.

These benefits need to be balanced against possible downsides and ethical concerns. There are many prominent examples. Algorithmic biases and the resulting discrimination raise concerns that people are disadvantaged for reasons they should not be, for instance by giving higher credit limits to men than to women (Condliffe

B. C. Stahl, *Artificial Intelligence for a Better Future*,
SpringerBriefs in Research and Innovation Governance,
https://doi.org/10.1007/978-3-030-69978-9_1

2019), by referring white people more often than Black people to improved care schemes in hospitals (Ledford 2019) or by advertising high-income jobs more often to men than to women (Cossins 2018). AI can be used to predict sexual preferences with a high degree of certainty based on facial recognition (The Economist 2017), thereby enabling serious privacy breaches.

The range of concerns goes beyond the immediate effects of AI on individuals. AI can influence processes and structures that society relies upon. For example, there is evidence to suggest that AI can be used to exert political influence and skew elections by targeting susceptible audiences with misleading messages (Isaak and Hanna 2018). People are worried about losing their livelihoods because their jobs could be automated. Big multinational companies use AI to assemble incredible wealth and market power which can then be translated into unchecked political influence (Zuboff 2019).

A further set of concerns goes beyond social impact and refers to the question of what AI could do to humans in general. There are fears of AI becoming conscious and more intelligent than humans, and even jeopardising humanity as a species. These are just some of the prominent issues that are hotly debated and that we will return to in the course of this book.

In addition to the many concerns about AI there are numerous ways of addressing these issues which require attention and input from many stakeholders. These range from international bodies such as the United Nations (UN) and the Organization for Economic Cooperation and Development (OECD) to national parliaments and governments, as well as industry groups, individual companies, professional bodies and individuals in their roles as technical specialists, technology users or citizens. As a consequence, discussion of the ethics of AI is highly complex and convoluted. It is difficult to see how priorities can be set and mitigation strategies put in place to ensure that the most significant ethical issues are addressed. The current state of the AI ethics debate can be described as a cacophony of voices where those who shout loudest are most likely to be heard, but the volume of the contribution does not always offer an assurance of its quality.

The purpose of this book is to offer new perspectives on AI ethics that can help illuminate the debate, and also to consider ways to progress towards solutions. Its novelty and unique contributions lie in the following:

1. The book provides a **novel categorisation of AI** that helps to categorise technologies as well as **ethical issues**
   I propose a definition of AI in Chapter 2 that focuses on three different aspects of the term: machine learning, general AI and (apparently) autonomous digital technologies. This distinction captures what I believe to be the three main aspects of the public discussion. It furthermore helps with the next task of the book, namely the categorisation of ethical issues in Chapter 3. Based on the conceptual distinction, but also on rich empirical evidence, I propose that one can distinguish three types of ethical issues: specific issues of machine learning, general questions about living in a digital world and metaphysical issues.

2. The book proposes **human flourishing** as the basis of an **ethical framework** to deal with the ethical challenges of AI.
   The three categories of ethical issues are descriptive, which means they are derived from observations of what people perceive as ethical issues. In order to move beyond description and find a basis for practice and intervention, a normative ethical position needs to be adopted. I argue that a suitable ethical theory that can be applied to AI and provide insights that guide action is that of flourishing ethics. Flourishing ethics has three considerable advantages. First, it covers the descriptive categories of AI ethics suggested in this book. Second, it is open to other ethical theories and allows for the integration of considerations of duty, consequences and care, among others. Third, it has a distinguished history, not only in ethics broadly, but also in the ethics of computing. What flourishing ethics requires is that AI, like any other technology and tool, should contribute to human flourishing. This position is not overly contentious, provides normative guidance and is sufficiently open to be applicable to the many technologies and application domains that constitute AI ethics.
3. The book offers a **novel classification of mitigation strategies** for the ethical challenges of AI.
   Classifying ethical issues and determining a suitable ethical theory can contribute to finding possible solutions. Such solutions do not develop in a vacuum but form part of an existing discourse. I therefore review the current discussion of mitigation measures that have been proposed to deal with these issues in Chapter 4. I distinguish between several categories of mitigation options, the first referring to policy and legislation, the second to options at the organisational level and the third to guidance mechanisms for individuals.
4. The book shows that the **metaphor of an ecosystem** helps us understand the complexity of the debate and offers **insights for practical interventions.**
   Based on a rich understanding of the AI landscape, I propose the interpretation of the AI ethics debate in terms of an ecosystem. The field of AI can be pictured as a set of interlinking ecosystems which consists of many different individual actors and groups interacting in complex ways that can influence the overall system unpredictably. Returning to the idea of flourishing, I suggest asking the question: how can the AI ecosystem as a whole be shaped to foster and promote human flourishing? This interpretation of AI ethics allows actions to be prioritised and bespoke advice to be developed for individual stakeholders and stakeholder groups. Perhaps most importantly, it leads to insights into higher-level activities, namely those that are conducive to the development of the ecosystem in the desired direction of promoting human flourishing.

This novel interpretation of the AI ethics debate not only offers conceptual insights and a theoretical basis allowing us to better understand, compare and contrast various issues and options, but also provides a foundation for practical actions. These are spelled out in more detail in Chapter 5. Following an introduction to the ecosystems view of AI and its limitations, I explore its implications for possible ways of addressing ethical issues. The ecosystems view implies that interventions into the

AI ecosystem clearly delineate the boundaries of the system they apply to. Such interventions need to support the development of the ecosystem by increasing the knowledge base and capacities of its members. A final requirement for any intervention into AI ecosystems is that it needs to employ governance mechanisms that are sensitive to the non-linear and often unpredictable dynamics of the system. On this basis I then propose some activities that are likely to shape the AI ecosystem in ways that are conducive to human flourishing.

Overall, this book offers a novel perspective on the AI ethics debate. It is based on empirical insights and strong concepts that help structure the debate in a transparent and constructive manner. Very importantly, I hope that the arguments I propose point beyond AI and offer guidance that is equally applicable to whichever technology succeeds AI when the current AI hype has subsided. It thereby offers a response to Floridi's (2018) call for ways to be found of governing the digital world.

References

Condliffe J (2019) The week in tech: algorithmic bias is bad. Uncovering it is good. The New York Times. https://www.nytimes.com/2019/11/15/technology/algorithmic-ai-bias.html. Accessed 21 Sept 2020

Cossins D (2018) Discriminating algorithms: 5 times AI showed prejudice. New Sci. https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/. Accessed 21 Sept 2020

The Economist (2017) Advances in AI are used to spot signs of sexuality. https://www.economist.com/science-and-technology/2017/09/09/advances-in-ai-are-used-to-spot-signs-of-sexuality. Accessed 21 Sept 2020

Floridi L (2018) Soft ethics and the governance of the digital. Philos Technol 31:1–8. https://doi.org/10.1007/s13347-018-0303-9

Isaak J, Hanna MJ (2018) User data privacy: Facebook, Cambridge Analytica, and privacy protection. Computer 51:56–59. https://doi.ieeecomputersociety.org/10.1109/MC.2018.3191268

Ledford H (2019) Millions of black people affected by racial bias in health-care algorithms. Nature 574:608–609. https://doi.org/10.1038/d41586-019-03228-6

Montreal Declaration (2018) Montréal declaration for a responsible development of artificial intelligence. Université de Montréal, Montreal. https://www.montrealdeclaration-responsibleai.com/the-declaration. Accessed 21 Sept 2020

Zuboff PS (2019) The age of surveillance capitalism: the fight for a human future at the new frontier of power. Profile Books, London