CONTENTS

INTRODUCTION

What is artificial intelligence (AI), and why does it raise philosophically interesting ethical questions? Why is it important to control AI, and what human values should AI technologies be aligned with? Who is responsible—who can be praised or blamed—when AI technologies produce good or bad outcomes? When people use generative AI to produce texts, images, music, or other outputs, who should be considered the author(s) of those outputs? If AI technologies continue to become increasingly "smarter," is there a risk that human beings will become "dumber" if we begin to rely very heavily on artificial intelligence in all areas of life? What if it becomes possible to create AI with some form of consciousness? What will the future of (meaningful) work look like in a world in which we hand over more and more tasks to AI technologies? Does having more AI in all areas of life necessarily mean a better world for all of us?

These kinds of questions are discussed within philosophical debates about the ethics of AI. This book is about such questions. It explores different ways we can reflect on—and begin to answer—these and related questions from a philosophical point of view. The book is not directly concerned with concrete policy making, nor with legal or economic questions about our present and future life with AI. Nor is it concerned with sociological or psychological questions. It is about distinctively philosophical questions about the ethics of AI.

What does this mean? What makes something a distinctively philosophical question about a topic (such as the ethics of AI)?

Philosophers sometimes shy away from saying what exactly philosophy is. It can be many different things. Yet, there are some things that many philosophical discussions have in common. Here are some examples.

Philosophical discussions are often about basic concepts and basic ideas we have about a subject matter. A typical philosophical question about AI, for instance, could regard what exactly we should understand by the term

"intelligence" and whether it is possible to create technologies with something that we could call intelligence.

Philosophical questions are also often about basic principles or fundamental norms related to different parts of life. Which basic ethical principles, for instance, should we follow in life in general? And which ethical principles should govern our creation, deployment, and use of AI technologies?

Philosophical questions are sometimes also about apparent conflicts among different beliefs and convictions that many people have. For example, many people believe that we have a free will that is not determined from the outside. At the same time, they might also believe that everything that happens in the universe is determined by natural laws. Is this a clash of ideas or a contradiction, or could both things be true at the same time?

Philosophical inquiry is also sometimes about different ways in which things could be or not be in hypothetical scenarios (in different "possible worlds"). For instance, is it conceivable that there is a possible world where human beings have brains and nervous systems just like ours, but where these alternative human beings are complete "zombies" in the sense that they lack any subjective consciousness? Would we have different ethical duties toward such philosophical zombies than those that we have toward normal human beings? Or, to give an example related to AI: Could AI technologies that function in the ways that current AI technologies do become conscious, and if so, what would that mean for how we should interact with them?

A final example of a classic philosophical topic is what it is to live a good and meaningful human life. In the context of AI, we can ask whether AI technologies might create new ways in which human beings could live good or meaningful lives, or whether more and more AI in all parts of life might, in certain ways, make it harder for us to live good and meaningful lives.

This book relates these types of philosophical questions to the ethics of AI. It encourages you (the reader) to make up your own mind about these topics. And it will try to provoke you to reflect deeply on—and perhaps even worry about—these topics.

The book is not a neutral introduction to philosophical aspects of the ethics of AI. It is an opinionated introduction. I have chosen topics I find particularly interesting. And my discussion is partly driven by my own

philosophical convictions in what follows. The aim of making this an opinionated introduction is not to convince all readers that they should necessarily share all my views. Instead, the idea is that a book that offers arguments and critically assesses others' ideas is more interesting and more engaging than a book that tries to remain completely neutral.

The reader is encouraged to question and disagree with the ideas and arguments discussed in these chapters. That is what philosophy is all about. It is about mutually respectful and careful critical reflection—which sometimes involves deep disagreements—about the most basic questions of human life.

* * *

The book will consider three different kinds of questions:

1. Ethical questions related to actual AI technologies that already exist in society
2. Ethical questions raised by how people perceive AI technologies, or the assumptions people make about AI, whether or not these are completely accurate or well-founded
3. Ethical questions raised by AI technologies that do not yet exist, but that may come into existence in the near or distant future

The first two types of questions will be discussed at greater length than the third type. But the third type of question will be discussed as well. Such questions must be discussed with care. Otherwise, philosophical discussions about them may veer off too far in the direction of science fiction. At the same time, we should not forget that good science fiction can be very philosophical. Reflection on what is, at present, best considered as science fiction can be interesting from a philosophical and ethical point of view. Still, most of the book will be about real people, real events, actually existing technologies, and AI of the sorts we already have in society. But we will not shy away from occasionally reflecting on philosophically interesting questions about possible future forms of AI.

There are some topics often discussed within the ethics of AI—such as explainability, biases in data sets and outputs from AI technologies, or the environmental impact of AI technologies—that we will address to some

extent, but less than might be found in other books about the ethics of AI. There are so many different interesting and important topics we could discuss in a book like this. We cannot fit them all into one book, so I have had to make choices about what to focus on.

Speaking about making choices, in this context, we also have to make decisions about what types of AI technologies to discuss. When I first started researching the ethics of AI a decade ago, the hottest topic was the ethics of self-driving cars. Now, at the time of writing, the hottest topic within the ethics of AI is generative AI, and large language models in particular. A decade from now—or even sooner than that—the hottest topic might be something else. For instance, some people have recently suggested that the next big thing in the ethics of AI is the topic of "advanced AI assistants," sometimes simply referred to as "agents" in the tech industry. At any rate, "agentic AI" is the latest hype in the tech world. So, how do we choose what AI technologies to discuss in a book like this?

One thing to notice is that a lot of philosophical questions about different AI technologies are actually similar in nature. In other words, the same questions tend to arise regarding different forms of AI within the ethics of AI at different points in time. For instance, will we lose control over these forms of AI? And who is responsible if something goes wrong? Those questions always come up. Then there are some questions specific to—or at least more clearly relevant in relation to—specific forms of AI. For instance, there are questions about language that are particularly relevant to current forms of large language model-based generative AI technologies. Such questions might be less relevant to, say, self-driving cars. This book features a mix of ethical questions that come up in relation to all forms of AI technologies, and ethical questions specifically related to particular types of AI.

From Chapter 1 onward, I will seldom mention myself. But as noted above, this is an opinionated introduction to the ethics of AI. So, what follows reflects the views and philosophical interests of the book's author. I am passionate about and deeply fascinated by the topics I cover here. My hope is that readers will be inspired to think more about these and related topics, and perhaps themselves go on to write about them.

This book is partly meant to be used in introductory courses about the ethics of AI at universities. But it is also intended to be of interest to fellow

philosophy researchers who work on the ethics of AI, or anybody interested in the topic. Accordingly, I am assuming that you, the reader, are interested in artificial intelligence and philosophy, including the branch of philosophy that deals with ethics. So, if you are ready for some ethics of AI from a philosophical perspective, let's dive right in!