THE MEANS OF PREDICTION

How AI Really Works (and Who Benefits)

MAXIMILIAN KASY

THE UNIVERSITY OF CHICAGO PRESS
Chicago and London

The University of Chicago Press, Chicago 60637
The University of Chicago Press, Ltd., London

Published 2025
Printed in the United States of America

34 33 32 31 30 29 28 27 26 25 1 2 3 4 5

ISBN-13: 978-0-226-83953-0 (cloth)
ISBN-13: 978-0-226-83954-7 (ebook)
DOI: https://doi.org/10.7208/chicago/9780226839547.001.0001

CIP data to come

♾ This paper meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

CONTENTS

PREFACE

There are a lot of great books about artificial intelligence. Some of them explain in elaborate technical detail how the engineering, statistics, and computer science of AI work. Others focus on one of the many problems that AI might bring about, including algorithms that don't work as promised, algorithms that discriminate, algorithms that turn against their human masters, and algorithms that automate away human workers. Still others discuss the problematic foundations on which AI is built, from the surveillance of internet users and the exploitation of click workers to the environmental destruction wrought by data centers and mining operations. These are all important issues.

What has not been presented is a unified framework for understanding how AI will proceed in a society that is shaped by power and inequality. This book aims to do that. Amid all the breathless debates about technical details, new possibilities, and social problems, I argue that the key issue that unites all the problems of AI is the choice of objectives that AI pursues, and the question of who controls these objectives. Control of these objectives is determined by control over the resources that are required for building AI—data, computational infrastructure, technical expertise, and energy. I call these resources the *means of prediction*.

Who am I to write this book? I am currently a professor of economics at the University of Oxford, where I teach machine learning

theory for graduate students and coordinate the machine learning and economics group. I come from a background in mathematics and statistics, as well as economics, and much of my research concerns questions of methodology. I draw on this background when reviewing the current state of AI in this book.

In some of my research, I work to move beyond the technical questions of statistics and machine learning to understand these fields in their social, political, and economic context. In a separate line of research, I work on economic inequality and what policy can do to enable a full and secure life for everyone. I study pilot job-guarantee and basic-income programs.

I believe that researchers have an obligation to contribute to a society where we collectively debate and decide our own future, rather than leaving questions of technology and policy to technocrats and experts. It is in this spirit that the present book aims to participate in a broad public debate about the future of AI.

All books build on the ideas, insights, and work of countless people other than the authors themselves. The present book is no exception. In preparing and writing this book, I have profited from the creativity, critique, reading suggestions, discussions, and research assistance of a great many friends, coauthors, colleagues, and students, including the following (in alphabetical order):

Alberto Abadie, Rediet Abebe, Daron Acemoglu, Isaiah Andrews, Johanna Barop, Stefano Caria, Nicolò Cesa-Bianchi, Gary Chamberlain, Roberto Colomboni, Ellora Derenoncourt, Binta Zahra Diop, Pirmin Fessler, Susann Fiedler, Alex Frankel, Carlos Gonzalez Perez, Verena Halsmayer, Ian Jewitt, Jeremy Large, Lukas Lehner, Gregory Levy, Peter Lindner, Carrie Love, Lester Mackey, Gerhard Meszaros, Sanaz Mobasseri, Christopher Muller, Suresh Naidu, Harald Oberhauser, Dietmar Offenhuber, Walter Palmetshofer, Daniela Platsch, Carina Prunkl, Simon Quinn, Alvaro Ramos-Chaves, Anja Sautmann, Frederik Schwerter, Jann Spiess, Alexander Teytelboym, Martin Weidner, Ashia Wilson, Noam Yuchtman, and Chad Zimmerman.

PART I
INTRODUCTION

Are you scared of artificial intelligence? You should be—if we are to believe some popular stories about the threat of AI and the coming conflict between humans and machines.

According to these stories, AI will attain superhuman capabilities and will start to self-improve. It will threaten humans in the name of self-preservation, and it will ultimately become an existential risk to humanity. These stories, told in movies, literature, industry, and academia, touch on our deepest and most fundamental fears. We fear to lose our livelihoods and to descend into poverty. We fear to lose our autonomy and to be controlled by incomprehensible and malign actors indifferent to our fates. We fear to lose our life. And we fear—even worse—that the survival of those we love, and the survival of humanity at large, might be threatened. On top of all that, we fear AI as an inscrutable force that is headed our way. Its arrival is inevitable, and its impact seems beyond anyone's control.

These stories of AI, the stories of the existential conflict between humans and machines, are repeated over and over in Hollywood and in Silicon Valley. But these stories do not help us make good decisions—in technology or in politics. They make it seem like there is only one possible direction for the development of AI, and that society cannot do anything about it. These stories also obfuscate who wins and who loses as AI develops. This obfuscation prevents the public debate from focusing on the real issues at stake, and it pre-

vents people from doing anything about them. Doing nothing serves the interests of those who benefit from keeping things as they are.

This book gives a different perspective on AI and society. Contrary to the popular stories, the progress of AI is not fate but rather a product of human choices. The key conflicts are not between humans and machines but between different people. The answer to these conflicts is shared democratic control of AI and of the objectives that it pursues: Those impacted by algorithmic decisions need to have a say over these decisions.

To provide a foundation for such a different perspective, this book first offers an unfettered way of thinking about AI in the way that machine learning experts think about it and understand it. In doing so, it shows the limits of AI, and it shows how AI can be made to work for all people. Doing so requires revising the stories that we tell ourselves about humans and machines.

1

THE STORY OF HUMANS VERSUS MACHINES

In the classic film *2001: A Space Odyssey*, which was released in 1968, a spaceship headed to Jupiter is equipped with an onboard computer named HAL 9000. Over time, this computer becomes a deadly antagonist of the astronauts on the ship. After an apparent computer error, several crew members try to switch HAL off. In the name of safeguarding the secret mission of the spaceship, HAL kills the crew members. Eventually, however, the astronaut Dave Bowman succeeds at deactivating HAL, ignoring the computer's desperate pleas to stop.

In *The Terminator* (1984), the conflict between humans and a self-preserving AI is taken up a notch, and the conflict becomes a question of survival for the entire human species.

Many of these same tropes appear in movies such as *The Matrix* (1999), *I, Robot* (2004), *Transcendence* (2014), *Ex Machina* (2015), *M3gan* (2022), *The Creator* (2023), and others. They reflect a particular fear of AI, one amplified by visible figures from the tech industry in this century: that we are headed toward a conflict between humans and machines. Elon Musk argued at the Bletchley Park AI summit that AI is "one of the biggest threats to humanity" and that, for the first time, we are faced "with something that's going to be far more intelligent than us." Sam Altman, of OpenAI, has claimed that generative AI could bring about the end of human civilization, and

that AI poses a risk of extinction on a par with nuclear warfare and global pandemics.

In academia, this story has also found some resonance. The philosopher Nick Bostrom has written extensively about the existential risks of AI for humanity, and the possibility of an intelligence explosion, where AI keeps improving itself once it has reached human level. The computer scientist Stuart Russell, together with his collaborators at the Center for Human-Compatible Artificial Intelligence at the University of California, Berkeley, has emphasized the so-called *alignment problem*—that is, the problem of making machine objectives align with human objectives.

Another dystopian story, which is almost equally scary, holds that AI won't kill us, but it will render human workers obsolete, inevitably leading to mass unemployment and social unrest. A 2023 Goldman Sachs report, for instance, claimed that generative AI might replace three hundred million full-time workers in Europe and the United States.

The story told in Hollywood and in Silicon Valley tends to feature a heroic conflict between a man (it is usually a man) and a machine—Dave Bowman and HAL 9000 in *Space Odyssey*, Kyle Reese and the Terminator, Nathan and Ava in *Ex Machina*, or Sam Altman and the AI-caused extinction of humanity. The academic version of the story, as told by computer scientists, also tends to feature a man and a machine, where there is a value-alignment problem of the machine (that is, a mis-specified objective) or a bias of the machine relative to its objective.

2

WHAT THE OLD STORY MISSES

This book will focus on the key issues that the story of man versus machine misses: Technology is not fate. Just as people make technology, people decide how it is used and what interests it serves. These decisions are made over and over again as AI is developed and deployed. AI is, furthermore, ultimately not that complicated. How AI works can be understood by anyone. The real conflict is not between a human and a machine but between the different members of society. And the answer to the various risks and harms of AI is public control of AI objectives through democratic means.

AI is, at its core, automated decision-making using *optimization*. That means that AI algorithms are designed to make some measurable objective as large as possible. Such algorithms might, for example, maximize the number of times that someone clicks on an ad. AI therefore requires that somebody picks the objective—the *reward*—that is being optimized. Somebody must, quite literally, type into their computer: "This is the measure of reward that we care about."

The important question, then, is who gets to pick the *objectives* of AI systems. We live in a capitalist society, and in such a society the objectives of AI are typically determined by the owners of capital. The owners of capital control the *means of prediction* that are needed for building AI—data, computational infrastructure, tech-

nical expertise, and energy. More generally, the objectives of AI are determined by those with social power, whether that is in the criminal justice system, in education, in medicine, or in the secret police forces of autocratic surveillance states.

One domain in which AI is deployed in society is the workplace. AI is used in robotized Amazon warehouses, in the algorithmic management of Uber drivers, and in the screening of job candidates by large companies. AI is also used in consequential domains outside the workplace, including the filtering and selection of Facebook feeds and of Google search results, where the objective is to maximize ad clicks. A third domain is predictive policing and the incarceration of defendants awaiting trial based on the prediction of crimes that they have not committed yet. Perhaps most devastatingly, AI is also deployed in warfare; it was, for instance, used to decide which family homes to bomb in Gaza beginning in 2023.

Of course, a good number of researchers and critics have warned of the dangers of using AI in these consequential domains. Joy Buolamwini, a computer scientist at MIT Media Lab, has written extensively on the dangers of inaccurate and racially biased facial recognition systems. Ruha Benjamin, a sociologist at Princeton, has emphasized that AI can replicate and reinforce existing social inequalities in domains such as education, employment, criminal justice, and health care. In a similar vein, Timnit Gebru, a computer scientist writing during her time working at Google, warned of the dangers of large language models acting as stochastic parrots, which repeat language patterns without understanding, and in doing so replicate the biases embedded in their training data. Meredith Whittaker, currently the president of the Signal Foundation, has criticized the political economy of the tech industry, where AI is used by powerful actors in ways that can entrench marginalization. Kate Crawford, professor at the University of Southern California and cofounder of the AI Now Institute, has emphasized the nature of AI as an extractive and exploitative industry.

Amid these overlapping critiques, each focused on a different as-

pect and pitfall of AI, it is challenging to formulate a systematic way of thinking about AI in society. One possible unifying perspective is provided by computer science. Computer scientists are trained to view most problems as optimization problems. In this context, optimization involves finding the decision that makes a given reward as large as possible, given limited computational resources and limited data.

The computer science perspective has informed much of the public discourse around AI safety and AI ethics, especially regarding topics such as fairness or value alignment: "If there is something wrong, then there must be an optimization error." In this view, the issue is simply that an action was picked that failed to maximize the specified objective. This perspective does not get to the heart of the problem in most cases, however, because it doesn't engage with the choice of the objective itself.

I argue that instead of optimization errors, it is conflicts of interest over the control of AI objectives that are the central issue. When AI causes human harm, the problem is usually not that an algorithm did not perfectly optimize. The problem is that the objective optimized by the algorithm is good for the people who control the means of prediction—people such as Jeff Bezos, founder and former CEO of Amazon, and Mark Zuckerberg, founder and CEO of Meta—but not good for the rest of society.

This understanding changes how we should think about possible solutions to the problems of AI. How do we address AI ethics and AI safety if the underlying problems are with the parties that set the objectives for AI? How do we choose these objectives in a way that serves the public rather than just a powerful minority? This book will make the case that the solution for the issues of AI ethics and safety can only be *democratic control*. Democratic control is not limited to democratically elected national governments; collective democratic decision-making can exist on many levels, including the workplace, the nation state, and the global level.

The challenge, of course, is that democracy is difficult. The demo-

cratic control of a new technology like AI requires public deliberation, and such public deliberation might seem impossible considering the view held by many (and reinforced by the tech industry) that AI is very complicated.

But despite all the technical jargon, and despite the breathless chase after the newest innovations, the basic ideas of AI are not that complicated, and not that new, and they can be understood by all of us. No matter who you are, don't let anyone tell you that you are not the "type" to understand AI. This book will start with a discussion of how AI works, which in turn functions as a foundation for understanding its political stakes and likely path in the future.

3

WHAT THIS BOOK DOES

This book provides a concise overview of how AI works—what it does and doesn't do—as a means toward understanding its capabilities and its political and economic impact. The following is a glimpse at the book's structure. It begins with the book's greatest challenge: translating a science that AI companies would like us to believe is impossibly complicated.

How AI Works

Many of the questions that need to be solved when building AI reflect ancient and fundamental questions about how it is possible learn from experience and how to act successfully in the world.

To get started, we will need to agree on what we are talking about. What is *artificial intelligence*? "Intelligence" is a notoriously loaded term, and public perceptions of AI have oscillated from "an obscure academic niche" to the very broad "everything related to data," and back to the narrow category of "large language models." This book will take an intermediate stance, in between these very broad and very narrow definitions. Following the standard technical treatment of AI, this book will define AI as "the construction of systems for *automated decision-making* to maximize some measurable reward." This definition is more specific than "anything to do with data,"

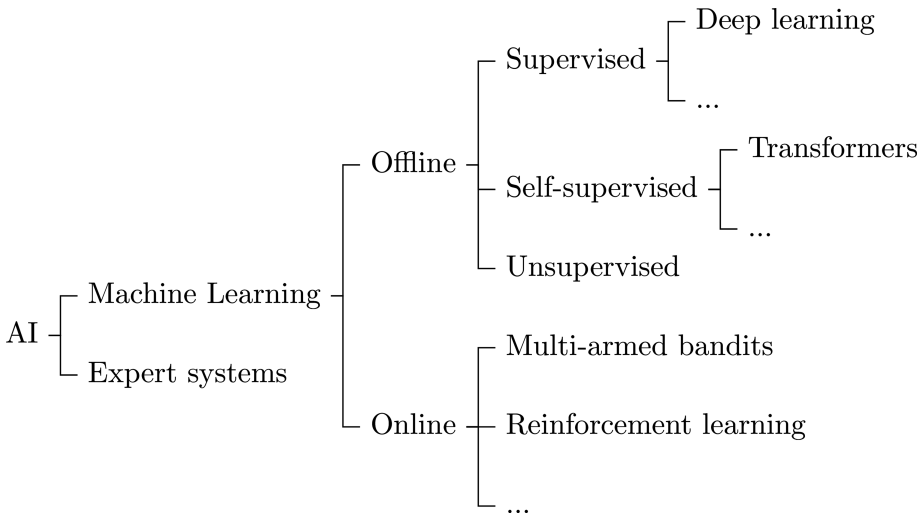

FIGURE 1 A taxonomy of AI

but at the same time, it includes a lot more than just *large language models*. This definition gives us a framework to talk about the many socially consequential settings where AI is used.

There are many branches of the field of AI. Figure 1 provides a taxonomy of some of these branches.

Until the end of the 1980s, *expert systems*, based on hand-coded human knowledge, were the dominant approach in AI. But most modern AI is based on *machine learning*. Machine learning uses data and statistical methods to build automated decision-making systems. To understand AI, we thus need to understand machine learning. One branch of machine learning is *supervised learning*, where the objective is to predict some outcome as accurately as possible. Many learning problems are prediction problems: In facial recognition, an individual's identity is predicted based on an image. In large language models, the next word is predicted based on the preceding words. In the hiring of job candidates, future performance is predicted based on candidate characteristics. In social media feeds, ad clicks are predicted based on user data.

Prediction is tricky, because it needs to navigate between two opposite dangers, *overfitting* and *underfitting*. When prediction overfits, it develops superstitions: It erroneously extrapolates random occur-

rences of the past into the future. If prediction underfits, it is stubborn and refuses to learn: It sticks to beliefs despite contradictory evidence.

AI algorithms might try to find the right balance between overfitting and underfitting by checking how well they can predict the data on hand, on which they were trained. The problem with this approach is that hindsight is easy—it is too easy to predict what we already know. This approach would therefore induce the algorithm to overfit the data of the past. Instead, for accurate evaluation, algorithms need to *split the data*—that is, check predictions on data the algorithm has not used yet. This approach is called *cross-validation*. Supervised learning relies on picking the model that does best, according to this cross-validation criterion.

One method for making predictions uses *deep learning*, a type of supervised learning that is based on training *neural nets*. Neural nets allow for modeling very complicated relationships. They have been extremely successful in recent years for prediction problems where data are abundant, such as image recognition or language modeling. Neural nets, and in particular *transformers* (a special kind of neural net), have also been central for *generative AI*—AI that produces text, images, or other media. This includes large language models, where the goal is to predict the most likely word to come next. (Large language models power applications such as ChatGPT.) Generative AI also includes image generation, where images are predicted based on text labels, as well as video generation.

Supervised learning is a form of *offline learning*, which describes learning based on data that are given. In *online learning*, data are collected over time, and what data the algorithm observes might depend on the actions it has previously taken. (*Online* here has nothing to do with being on the internet; it refers to learning over time.) Humans face the same situation: We only see the consequence of an action if we take it. Doctors, for example, can only learn whether a new drug works if they prescribe this drug to some patients. Success-

ful AI, in online settings, needs to both *explore* new things, by experimenting, and to *exploit* what it has learned, in just the right balance. *Multi-armed bandit* algorithms are designed to do exactly that: They aim to find exactly the right balance between experimenting with different options and doing what seems best based on what has been learned from actions they have previously tried. *Reinforcement learning* goes one step further than multi-armed bandit algorithms. Reinforcement learning builds algorithms that learn to *plan* by learning how likely it is that different states of the world are favorable down the road.

The Politics and Economics of AI

The technical overview of AI in this book provides the core ingredients that are necessary to fulfill humanity's ancient dream—to become a Prometheus, a Demiurge, a creator of another intelligence, of an *artificial intelligence*. But beware! Is this coveted creation of AI what we hoped for? Or will it instead be a source of new dangers for society? Will we lose control of our Frankensteinian monster, our golem; will we be like the sorcerer's apprentice in the movie *Fantasia*?

To confront this question about the dangers of AI, we need to think about how to evaluate the impact of AI on society, about who gets to steer AI, and whether their direction is desirable for society. This is a wider lens than most conversations about AI take, as it entails moving beyond the standard framework of machine learning and the optimization of a single given objective. In the narrow view of AI, based only on the logic of optimization, all problems can be solved by better engineering, and their solutions are best left to the experts. What this optimization framework misses, however, is the social nature of the problems caused by AI.

If an algorithm selecting what you see on social media promotes outrage, thereby maximizing engagement and ad clicks, the problem is not an optimization error: Promoting outrage is good for prof-

its from ad sales, even if it is bad for society. If another algorithm choosing whom to invite for a job interview systematically rules out candidates who are likely to have family-care responsibilities outside the workplace, that is also not an optimization error: It is good for profits but bad for future parents or those taking care of elderly relatives. If an algorithm setting health insurance rates screens out people who are likely to develop chronic health problems or disabilities, that is not an optimization error either. It is good for profits but bad for people who need health care.

Rather than understanding everything as an optimization problem with a single objective, it is critical to recognize that we live in a world where different people have different objectives. We live in a world of inequality, distributional struggles, and conflicting value systems. Accordingly, the most important question for AI must be, Who gets to pick the objective? By asking this question, we transcend the ideological obfuscations that are promoted by the beneficiaries of the status quo.

The next part of this book is thus dedicated to thinking about what makes a good society and how to get there. This problem needs to be discussed to address the questions of what objectives AI should maximize and who should choose them. In grappling with this problem, we once again confront ancient and fundamental questions.

This book assumes that a society is good if it is good for the people in it. This may sound trivial, but it has important implications for thinking about AI. Assuming that a society should be good for people leads to a set of questions that need to be answered, to evaluate the social impact of AI. First, who are the people whose welfare matters? Second, how do we measure their welfare? And third, how do we consider trade-offs between the welfare of different people that are affected by AI—in other words, how do we assess *social welfare*? AI is beneficial to a society if it maximizes social welfare.

The question of how to measure individual welfare is particularly thorny. One might focus on *opportunities* or on *outcomes*. One might

focus on objectively measurable standards of well-being or, instead, on the economic concept of subjective *utility*—that is, on what individuals would choose if they had a choice. These distinctions imply different ways to evaluate the effects of AI on individuals, and they matter especially when thinking about the use of AI for social good.

Constructing AI for a better society requires not only knowing where we want to go but also how to get there. And it requires knowing who will get us there: Who are potential agents of change in a system that appears so inevitable? Potential agents of change are individuals or organizations who can align AI objectives with social welfare. They need to have the interests, values, and capacity to do so.

In the field of AI ethics, the focus is often on convincing AI engineers and their managers to be nice. But engineers and managers, independent of their personal qualities, are constrained by the requirement of profit maximization that governs private corporations; their individual agency and accountability are secondary to the fact that they must do a job. If not corporate engineers, who else has the capacity to change the course of AI? The objectives of AI are chosen by those who control the resources to build AI, what this book calls the *means of prediction*. These resources include data, computational infrastructure, technical expertise, and energy. Against this backdrop, agents of change need to have strategic leverage over the actors who control the means of prediction to be able to effect change. Leverage can take many forms, from potential strikes and consumer boycotts to bad press and litigation to regulation and legal constraints. The pool of potential agents of change is surprisingly large: unions, consumer advocates, journalists, judges, policymakers, and politicians.

Agents of change not only need leverage—they also need values and interests that motivate them to move AI in the right direction. The development of such values might be hindered by ideologies, and effective collective action to change the direction of AI is undermined by ideological obfuscation of the issues at stake. By represent-

ing problems as optimization errors, or as fights between man and machine, attention is diverted from distributional conflicts. By representing problems as technical issues that are best left to experts, rather than as social choices that require collective deliberation, the possibility of democratic governance is denied. By painting the development of AI as inevitable, change is forestalled, and the status quo is preserved.

AI in Society

Equipped with this background, we will discuss how to regulate algorithms in the next part of the book. We will revisit debates around problem domains, including value alignment and AI safety, privacy and data property rights, automation in the workplace, fairness and algorithmic discrimination, and explainability of algorithmic decision.

The objectives of AI are determined by those who control the means of prediction, particularly the large datasets needed to train AI. Because data are the basic resource that AI builds on, data privacy and data ownership are core issues. Computer science has developed a systematic framework for discussing privacy, *differential privacy*. When an algorithm satisfies differential privacy, whether an individual was included in the underlying data cannot be determined from the algorithm's output: the individual could be in there, but the algorithm won't tell. As a legal counterpart to this notion of privacy in computer science, privacy legislation might give individuals actual *property rights* over their data. But the problem with this focus on individual property rights, from an economic perspective, is that this focus misses the point of machine learning: Machine learning is focused on patterns across individuals not individual data. To use the language of economics, learning is all about the *externalities*—the implications of data collection for third parties—in the sense that one person's data can be used to make predictions about another

person. When there are pervasive externalities, individual property rights cannot prevent social harms. Because machine learning is all about externalities, differential privacy can be maintained while machine learning proceeds largely unimpeded, including all its harms and benefits. For this reason, collective governance is the only possible solution to regulate data collection in a way that addresses the harms and benefits of AI.

A form of democratic governance is also needed in the AI-augmented workplace. The introduction of new technologies in the workplace typically enables a company to produce more with fewer inputs. But this does not mean that everybody gains. It is quite possible that a new technology increases the *average output* per worker while at the same time decreasing wages for some or all of those workers. The reason is that wages reflect the *additional output* that comes from hiring an additional worker not the average output across all workers. The former might decrease, while the latter increases. Because of this, it is possible to have economic growth without shared prosperity. AI, in particular, might be used in this way, *automating* away a range of occupations, while making its owners richer. Again, this is not destiny. How new technologies are used is a choice. If workers (or workers' organizations) have a say in the choice of objectives for AI in the workplace, via some form of worker representation in decision-making at the company, then growth with shared prosperity is much more likely.

AI drives inequality not only by shifting labor demand but also via *algorithmic discrimination* and *bias*. To identify algorithmic bias, both economists and computer scientists often point to a deviation from (profit) maximization. For example, if a man is chosen instead of a woman by a hiring algorithm, this is interpreted as bias only if it would have been more profitable to hire the woman—but not otherwise. This perspective, which only sees a problem if profits are not optimized, again relies on an ideological obfuscation: It purports to reflect the interests of disadvantaged groups while in fact advocating for the maximization of profits. This is not to say that algorith-

mic bias isn't real and insidious. But to identify and quantify it, one must assess who stands to gain or lose from the introduction of an AI system in a given setting, and whose objectives are being maximized, rather than just asking whether there was a deviation from profit maximization.

If there is a concern that algorithmic decisions might be biased, one possible response is to require explanations of consequential algorithmic decisions, such as hiring decisions. In this spirit, discussions of *explainability* and *accountability* for automated decisions focus on individual recourse, where individuals might have a right to an explanation why a particular decision was made. To allow for recourse, a common suggestion is to use simple algorithms, which make it possible to explain decisions. But simplicity is a moving target. Instead of just explaining individual decisions to allow for individual recourse, we should focus on explaining AI systems and the objectives they maximize. These are necessary steps toward their democratic governance.

Concluding the book, in part V, I summarize some of the big questions that were discussed throughout—how to learn from observation, how to successfully act in the world, what makes a good society, who might get us there, and what that implies for the ethics and politics of AI.

This book makes the case that democratic governance of AI is critical to ensure that its uses are broadly beneficial. But this book does not provide a detailed blueprint for the democratic institutions that are needed to implement such governance. I end the book with some thoughts on different forms that democratic governance might take, going beyond the limitations of electoral and direct democracy, and involving arrangements such as *sortition* and *liquid democracy*. This brings us to the end of the book, but it is also only a beginning: Democratic governance of AI needs to be fought for and put into practice. This will be a task for all of us.

Without further ado, let us now get started.