About the book

This book gives a perspective on machine learning that treats fairness as a central concern rather than an afterthought. We'll review the practice of machine learning in a way that highlights ethical challenges. We'll then discuss approaches to mitigate these problems.

We've aimed to make the book as broadly accessible as we could, while preserving technical rigor and confronting difficult moral questions that arise in algorithmic decision making.

This book won't have an all-encompassing formal definition of fairness or a quick technical fix to society's concerns with automated decisions. Addressing issues of fairness requires carefully understanding the scope and limitations of machine learning tools. This book offers a critical take on current practice of machine learning as well as proposed technical fixes for achieving fairness. It doesn't offer any easy answers. Nonetheless, we hope you'll find the book enjoyable and useful in developing a deeper understanding of how to practice machine learning responsibly.

Why now?

Machine learning has made rapid headway into socio-technical systems ranging from video surveillance to automated resume screening. Simultaneously, there has been heightened public concern about the impact of digital technology on society.

These two trends have led to the rapid emergence of Fairness, Accountability, and Transparency in socio-technical systems (FAT*) as a research field. While exciting, this has led to a proliferation of terminology, rediscovery and simultaneous discovery, conflicts between disciplinary perspectives, and other types of confusion.

This book aims to move the conversation forward by synthesizing long-standing bodies of knowledge, such as causal inference, with recent work in the FAT* community, sprinkled with a few observations of our own.

How did the book come about?

In the fall semester of 2017, the three authors each taught courses on fairness and ethics in machine learning: Barocas at Cornell, Hardt at Berkeley, and Narayanan at Princeton. We each approached the topic from a different perspective. We also presented two tutorials: Barocas and Hardt at NIPS 2017, and Narayanan at FAT* 2018. This book emerged from the notes we created for these three courses, and is the result of an ongoing dialog between us.

Who is this book for?

We've written this book to be useful for multiple audiences. You might be a student or practitioner of machine learning facing ethical concerns in your daily work. You might also be an ethics scholar looking to apply your expertise to the study of emerging technologies. Or you might be a citizen concerned about how automated systems will shape society, and wanting a deeper understanding than you can get from press coverage.

We'll assume you're familiar with introductory computer science and algorithms. Knowing how to code isn't strictly necessary to read the book, but will let you get the most out of it. We'll also assume you're familiar with basic statistics and probability. Throughout the book, we'll include pointers to introductory material on these topics.

On the other hand, you don't need any knowledge of machine learning to read this book: we've included an appendix that introduces basic machine learning concepts. We've also provided a basic discussion of the philosophical and legal concepts underlying fairness.[1]

[1] These haven't yet been released.

What's in this book?

This book is intentionally narrow in scope: you can see an outline here. Most of the book is about fairness, but we include a chapter[2] that touches upon a few related concepts: privacy, interpretability, explainability, transparency, and accountability. We omit vast swaths of ethical concerns about machine learning and artificial intelligence, including labor displacement due to automation, adversarial machine learning, and AI safety.

[2] This chapter hasn't yet been released.

Similarly, we discuss fairness interventions in the narrow sense of fair decision-making. We acknowledge that interventions may take many other forms: setting better policies, reforming institutions, or upending the basic structures of society.

A narrow framing of machine learning ethics might be tempting

to technologists and businesses as a way to focus on technical interventions while sidestepping deeper questions about power and accountability. We caution against this temptation. For example, mitigating racial disparities in the accuracy of face recognition systems, while valuable, is no substitute for a debate about whether such systems should be deployed in public spaces and what sort of oversight we should put into place.

About the authors

Solon Barocas is an Assistant Professor in the Department of Information Science at Cornell University. His research explores ethical and policy issues in artificial intelligence, particularly fairness in machine learning, methods for bringing accountability to automated decision-making, and the privacy implications of inference. He was previously a Postdoctoral Researcher at Microsoft Research, where he worked with the Fairness, Accountability, Transparency, and Ethics in AI group, as well as a Postdoctoral Research Associate at the Center for Information Technology Policy at Princeton University. Barocas completed his doctorate at New York University, where he remains a visiting scholar at the Center for Urban Science + Progress.

Moritz Hardt is an Assistant Professor in the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley. His research aims to make the practice of machine learning more robust, reliable, and aligned with societal values. After obtaining a PhD in Computer Science from Princeton University in 2011, Hardt was a postdoctoral scholar and research staff member at IBM Research Almaden, followed by two years as a research scientist at Google Research and Google Brain. Together with Solon Barocas, Hardt co-founded the workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) in 2014.

Arvind Narayanan is an Associate Professor of Computer Science at Princeton. He studies the risks associated with large datasets about people: anonymity, privacy, and bias. He leads the Princeton Web Transparency and Accountability Project to uncover how companies collect and use our personal information. His doctoral research showed the fundamental limits of de-identification. He co-created a Massive Open Online Course as well as a textbook on Bitcoin and cryptocurrency technologies. Narayanan is a recipient of the Presidential Early Career Award for Scientists and Engineers.

Thanks and acknowledgements

This book wouldn't have been possible without the profound contributions of our collaborators and the community at large.

We are greatful to our students for their active participation in pilot courses at Berkeley, Cornell, and Princeton. Thanks in particular to Claudia Roberts for lecture notes of the Princeton course.

Special thanks to Katherine Yen for editorial and technical help with the book.

Moritz Hardt is indebted to Cynthia Dwork for introducing him to the topic of this book during a formative internship in 2010.

We benefitted from substantial discussions, feedback and comments from Andrew Brunskill, Aylin Caliskan, Frances Ding, Michaela Hardt, Lily Hu, Ben Hutchinson, Lauren Kaplan, Niki Kilbertus, Kathy Kleiman, Issa Kohler-Hausmann, Eric Lawrence, Zachary Lipton, Lydia T. Liu, John Miller, Smitha Milli, Shira Mitchell, Robert Netzorg, Juan Carlos Perdomo, Claudia Roberts, Olga Russakovsky, Matthew J. Salganik, Carsten Schwemmer, Ludwig Schmidt, Annette Zimmermann, Tijana Zrnic.