

Contents

Preface	ix
Online Materials	xiv
Acknowledgments	xv

1 Introduction 1

Demographic Disparities	3
The Machine Learning Loop	5
The State of Society	6
The Trouble with Measurement	8
From Data to Models	11
The Pitfalls of Action	13
Feedback and Feedback Loops	14
Getting Concrete with a Toy Example	17
Justice beyond Fair Decision Making	20
Our Outlook: Limitations and Opportunities	22
Bibliographic Notes and Further Reading	23

2 When Is Automated Decision Making Legitimate? 25

Machine Learning Is Not a Replacement for Human Decision Making	26
Bureaucracy as a Bulwark against Arbitrary Decision Making	28
Three Forms of Automation	30
Mismatch between Target and Goal	36
Failing to Consider Relevant Information	38
The Limits of Induction	41
A Right to Accurate Predictions?	43
Agency, Recourse, and Culpability	44
Concluding Thoughts	47

3 Classification 49

Modeling Populations as Probability Distributions	50
Formalizing Classification	51
Supervised Learning	56
Groups in the Population	58

	Statistical Nondiscrimination Criteria	60
	Independence	61
	Separation	63
	Sufficiency	67
	How to Satisfy a Nondiscrimination Criterion	70
	Relationships between Criteria	71
	Case Study: Credit Scoring	74
	Inherent Limitations of Observational Criteria	79
	Chapter Notes	80
4	Relative Notions of Fairness	83
	Systematic Relative Disadvantage	83
	Six Accounts of the Wrongfulness of Discrimination	85
	Intentionality and Indirect Discrimination	87
	Equality of Opportunity	88
	Tensions between the Different Views	93
	Merit and Desert	95
	The Cost of Fairness	98
	Connecting Statistical and Moral Notions of Fairness	100
	The Normative Underpinnings of Error Rate Parity	105
	Alternatives for Realizing the Middle View of Equality of Opportunity	109
	Summary	110
5	Causality	113
	The Limitations of Observation	114
	Causal Models	117
	Causal Graphs	120
	Interventions and Causal Effects	123
	Confounding	124
	Graphical Discrimination Analysis	127
	Counterfactuals	132
	Counterfactual Discrimination Analysis	138
	Validity of Causal Modeling	143
	Chapter Notes	149
6	Understanding United States Antidiscrimination Law	151
	History and Overview of US Antidiscrimination Law	152
	A Few Basics of the American Legal System	157
	How the Law Conceives of Discrimination	163
	Limits of the Law in Curbing Discrimination	167
	Regulating Machine Learning	172
	Concluding Thoughts	182
7	Testing Discrimination in Practice	185
	Part 1: Traditional Tests for Discrimination	186
	Audit Studies	186
	Testing the Impact of Blinding	190

Revealing Extraneous Factors in Decisions	192
Testing the Impact of Decisions and Interventions	193
Purely Observational Tests	194
Taste-Based and Statistical Discrimination	198
Studies of Decision-Making Processes and Organizations	200
Part 2: Testing Discrimination in Algorithmic Systems	202
Fairness Considerations in Applications of Natural Language Processing	203
Demographic Disparities and Questionable Applications of Computer Vision	204
Search and Recommendation Systems: Three Types of Harms	206
Understanding Unfairness in Ad Targeting	208
Fairness Considerations in the Design of Online Marketplaces	210
Mechanisms of Discrimination	212
Fairness Criteria in Algorithmic Audits	214
Information Flow, Fairness, Privacy	215
Comparison of Research Methods	217
Looking Ahead	219
Chapter Notes	219
8 A Broader View of Discrimination	221
Case Study: The Gender Earnings Gap on Uber	221
Three Levels of Discrimination	225
Machine Learning and Structural Discrimination	229
Structural Interventions for Fair Machine Learning	234
Organizational Interventions for Fairer Decision Making	238
Concluding Thoughts	245
Chapter Notes	247
Appendix: A Deeper Look at Structural Factors	248
9 Datasets	251
A Tour of Datasets in Different Domains	252
Roles Datasets Play	260
Harms Associated with Data	271
Beyond Datasets	274
Summary	282
Chapter Notes	282
References	285
Index	311