Contents

List of Figures xi
List of Tables xiii
List of Cases xv
Preface xvii
Acknowledgements xxi

Introduction 1

I Basic Concepts 17

1 A Brief History of Artificial Intelligence 19
1.1 The Idea of AI 21
1.2 The Invention of AI 25
1.3 First-Wave AI: False Promises 28
1.4 Second-Wave AI: Empty Threats 34
1.5 Third-Wave AI: Deep Hype 38
1.6 Summary 40

2 Artificial Intelligence Today 43
2.1 Neural Network Architectures 45
2.2 Data and Datasets 50
2.3 Machine Learning Methods 51
2.4 Objectives, Goals, and Values 55
2.5 Learning Algorithms 59
2.6 Evaluation 61
2.7 Scaling Laws 62
2.8 Summary 63

3 The Value Alignment Problem **67**
3.1 The Standard Definition of Value Alignment 68
3.2 Adding Sophistication to the Standard Definition 72
3.3 The Principal-Agent Framework 76
3.4 The Value Alignment Problem for Artificial Intelligence 80
3.5 Benefits of the Structural Definition 84

II Axes of Value Alignment **87**
Introduction to Part II 89

4 Objectives **91**
4.1 Proxies and Abstractions 92
4.2 Insights from the Structural Definition 98
4.3 Bias and Fairness 101
4.4 Algorithmic Bias 104
4.5 The Social Character of Objectives 110
4.6 Summary 112

5 Information **115**
5.1 Informational Asymmetries, Economic and Artificial 116
5.2 Transparency and Opacity 120
5.3 Explainability, Interpretability, and Understanding 126
5.4 Data and Datasets 128
5.5 Interaction Effects 132
5.6 Summary 135

6 Principals **137**
6.1 Principals and Their Goals 139
6.2 The Values of Humanity 142
6.3 The Values Encoded in AI Research 144
6.4 The Human Costs of Artificial Intelligence 146
6.5 Interaction Effects 155
6.6 Summary 156

III Approaches to Value Alignment **159**
Introduction to Part III 161

7 AI Safety **163**
7.1 Adversarial Examples 164
7.2 Concrete Problems in AI Safety 166

7.3 Mitigating Risk 172
7.4 AI Safety and the Value Alignment Problem 178
7.5 Summary 180

8 Machine Ethics **183**
8.1 Artificial Moral Agency 185
8.2 Our Best Normative Theories 188
8.3 Technical Approaches to Artificial Moral Agency 194
8.4 Critiques of Artificial Moral Agency 197
8.5 Related Concepts 200
8.6 Summary 201

IV Mitigating Misalignment **203**
Introduction to Part IV 205

9 Measuring Degrees of Alignment **207**
9.1 Benchmarking 208
9.2 Benchmarking Ethics 213
9.3 Aligning Values 218
9.4 Degrees of Alignment 218
9.5 The Scaling Hypothesis for Value-Aligned AI 223
9.6 Summary 225

10 Normativity and Language **227**
10.1 Linguistic Communication 228
10.2 Language in Human Value Alignment 231
10.3 Language, Value Alignment, and Information Transfer 234
10.4 Objective Functions and Value Proxies 235
10.5 Implications 238
10.6 Summary 240

11 Values and Value-Ladenness **243**
11.1 The Value-Free Ideal of Science 244
11.2 Against the Value-Free Ideal 248
11.3 Values and Value Alignment 250
11.4 Optimism 253
11.5 Regulation 258
11.6 Summary 260

12 Conclusion **263**

V Appendix **269**

A Superintelligence and Control **271**
A.1 Superintelligence 271
A.2 Paths to Superintelligence 272
A.3 Forms of Superintelligence 276
A.4 Intelligence Explosion and the Singularity 278
A.5 Existential Risk 279
A.6 Intelligence, Motivation, and Goals 280
A.7 The Control Problem 281
A.8 Criticism 282
A.9 Summary 283

References 285
Index 321