

Contents

1	Introduction: Why AI Ethics?	1
1.1	Introduction	1
1.2	What Is Artificial Intelligence?	2
1.3	And What, Precisely, Is Ethics?	5
1.4	What Does This Book Cover?	8
1.5	Why AI Ethics Is So Exciting	9
1.6	Why Should Anyone Want to Study AI Ethics?	13
1.7	Bottom-Up and Top-Down Approaches to AI Ethics	17
1.8	And What Are You Bringing to This Book?	18
1.9	Frequent Criticisms of Ethics	19
1.10	The Boundaries and Limits of Ethics	22
1.11	The Approach of This Book	25
1.12	Developing Ethical Autonomy	26
1.13	Why Examine Underlying Concepts in AI Ethics?	29
1.14	Why Am I Writing This Book?	30
1.15	What Is to Come	31
1.16	Key Points	32
1.17	Educator Notes	32
	References	33
	Further Reading	33
2	The Rise of AI Ethics	35
2.1	AI Ethics: Why Now?	35
2.2	Same or Different? What's Special About AI Ethics?	38
2.3	Common Themes and Varieties of Questions in AI Ethics	40
2.4	Ethical Questions in AI Currently Receiving Attention	43
2.4.1	Freedom and Autonomy	45
2.4.2	Transparency and Explanation	48
2.4.3	Justice and Fairness	49
2.4.4	Beneficence/Nonmaleficence	51

2.4.5	Responsibility	53
2.4.6	Privacy	54
2.4.7	Trust	55
2.4.8	Sustainability	57
2.4.9	Dignity	58
2.4.10	Solidarity	59
2.5	A Brief History of AI Ethics	60
2.5.1	Why Consider History?	60
2.5.2	Playing with Fire: Historical Concerns About Technologies	64
2.6	AI Ethics Now: Who, What, Where, and Why?	75
2.7	AI Ethics, Law, Policy, Regulation, and Ethics Guidance	77
2.8	Looking Forward: What Is Our Job Now?	78
2.9	Case Study: An Indigenous AI Report	79
2.10	Key Points	81
2.11	Educator Notes	82
	References	83
	Further Reading	87
3	AI, Philosophy of Technology, and Ethics	91
3.1	Artificial Intelligence, Ethics, and AI as a Technology	91
3.2	Comparing AI to Humans	93
3.2.1	Case Study: Attributing Agency to Humans and to AI	95
3.3	A Brief Introduction to the Philosophy of Technology	96
3.3.1	Preliminaries	96
3.3.2	A Little History	98
3.3.3	Knowledge and Technology	102
3.3.4	The ‘Technological Fix’	104
3.3.5	Boundary Questions and Ethical Conundrums: Technology, Metaphysics, Ontology, and Ethics	107
3.3.6	Some Implications for the Conduct of AI Ethics	109
3.4	Technology and Attitudes Towards Nature	110
3.5	The Implicit Values of Technology	115
3.6	Communication Through New Technologies	120
3.6.1	AI and Communication Technologies	120
3.6.2	The Reproduction of Ideas and Images	121
3.6.3	Artificial Intelligence as a Medium	124
3.7	Key Points	125
3.8	Educator Notes	126
	References	127
	Further Reading	129

- 4 Methods in Applied Ethics 131**
 - 4.1 Introduction: A Simple Account of Methods in Applied Ethics 131
 - 4.1.1 Problems in the Methodology 134
 - 4.1.2 From Theory to Practice and Back Again: Fine-Tuning Responses in Applied Ethics 136
 - 4.1.3 Searching for Help: Can We Base AI Ethics on Ideas Borrowed from Medical Ethics? 139
 - 4.2 The Challenges from AI for Methodology in Applied Ethics . . . 143
 - 4.2.1 Describing and Identifying Ethical Issues 146
 - 4.3 Common Assumptions About the Nature of Ethics, and Their Influence on Methodology in Applied Ethics 148
 - 4.3.1 The Object of Morality 156
 - 4.4 Methodology in Applied Ethics: The Stories We Tell 158
 - 4.4.1 Introduction 159
 - 4.4.2 Stories About AI 161
 - 4.4.3 Science Fiction, Ethics, and AI 163
 - 4.4.4 Images of AI 165
 - 4.4.5 Describing Ethical Questions: Causation, Responsibility, and Ethics 169
 - 4.5 Key Points 173
 - 4.6 Educator Notes 174
 - References 175
 - Further Reading 178
- 5 Humans and Intelligent Machines: Underlying Values 181**
 - 5.1 Introduction 181
 - 5.2 The Nature and Value of Intelligence 183
 - 5.2.1 What Is Intelligence, and Why Do We Value It? 183
 - 5.2.2 Measuring Intelligence 191
 - 5.2.3 Intelligence and the Value of the Individuals 192
 - 5.2.4 A Quick Note on Intelligence and Speed 193
 - 5.2.5 Intelligence and Society 194
 - 5.2.6 Augmenting Human Intelligence with AI 196
 - 5.3 Values Underlying the Use of Data 197
 - 5.3.1 One View of the Value of Big Data 198
 - 5.3.2 Truth and Communication 199
 - 5.4 Human Nature 200
 - 5.4.1 Introduction 201
 - 5.4.2 Our Place in the Universe 201
 - 5.4.3 Claims About the Status of Humans Within the Natural World 204
 - 5.4.4 Humans and a Material Universe 205
 - 5.4.5 Is There an Essence to Human Nature? 210

5.4.6	The Divided Self and Notions of the ‘Higher Self’	213
5.4.7	Reason or Emotion?	215
5.4.8	The Boundaries of Human Nature	217
5.4.9	Boundaries and Limits	219
5.4.10	Is Human Nature Perfectible?	220
5.4.11	Myths and Origin Stories About Human Nature	221
5.4.12	Some Helpful Questions to Ask About Human Nature	222
5.5	Key Points	223
5.6	Educator Notes	224
References		225
	Further Reading	227
6	Normative Ethical Theory and AI Ethics	229
6.1	Introduction to Normative Ethical Theories	229
6.2	Normative Ethical Theories: Consequentialism	231
6.2.1	Outcomes of Actions: Pleasure, Pain, and Preferences	234
6.2.2	Measuring Outcomes	244
6.2.3	Whose Happiness Matters?	246
6.2.4	How Should We Aim at Happiness?	248
6.2.5	Consequentialism, Individuals, and Agency	253
6.3	Normative Ethical Theories: Deontological Approaches	256
6.3.1	Introduction	256
6.3.2	Immanuel Kant, Rule-Following, Agency, and Lessons for AI Ethics	259
6.3.3	Understanding and Interpreting Rules	261
6.4	Normative Ethical Theories: Virtue Ethics	264
6.5	Key Points	272
6.6	Educator Notes	273
References		274
	Further Reading	276
7	Philosophy for AI Ethics: Metaethics, Metaphysics, and More	277
7.1	Introduction	277
7.2	Metaethics	278
7.2.1	The Relevance of Metaethics for AI Ethics	278
7.2.2	Metaethics Behind AI Ethics: The MIT Moral Machine Experiment	279
7.2.3	Universal Values and Objectivity in Ethics	289
7.2.4	Empirical Approaches to Ethics	294
7.2.5	How Can We Gain Moral Understanding?	299
7.2.6	Technology, Attention, and Persuasion	304
7.3	The Theory of Knowledge	306
7.4	Philosophy of Language	308

7.5	Questions in Metaphysics: What Is Reality?	310
7.6	Key Points	314
7.7	Educator Notes	314
	References	315
	Further Reading	317
8	Persons and AI	319
8.1	Introduction	319
8.2	An Account of Personhood: Respect for Persons as a Fundamental Ideal	323
8.3	Some Contemporary Models of Personhood and Their Ethical Implications	326
8.4	Interlude: Mind and Body	332
8.5	Are Conceptions of Personhood Too Focused on the Cognitive and the Rational?	339
8.6	Personhood Continued: Two Approaches to Personhood, Agency, and Machines	341
8.7	The Limits and Boundaries of Persons	344
	8.7.1 Personal Identity over Time	345
	8.7.2 The Distinction Between Persons and Things	347
	8.7.3 Understanding Persons from the Outside: Constructing Persons from Data	349
	8.7.4 Persons and the Significance of the Body: Faces and Facial Recognition	350
	8.7.5 Autonomy	351
	8.7.6 Privacy	354
8.8	Key Points	355
8.9	Educator Notes	356
	References	357
	Further Reading	361
9	Individuals, Society, and AI: Online Communication	363
9.1	Introduction	363
	9.1.1 Moderating Online Content: An Overview	365
9.2	What Concepts and Tools Can We Bring to These Questions?	368
9.3	Why Is AI Being Used to Address the Online Harms Question?	369
9.4	Content Moderation for Online Harms: What Is Needed?	372
	9.4.1 Online Content: Who Is Involved?	376
	9.4.2 What Is Online ‘Harm’?	378
9.5	Challenges for AI in Tackling Online Harms	381
	9.5.1 Introduction: Detecting Online Harms	382
	9.5.2 Language and Meaning	383
	9.5.3 Difficulties with Hate Speech	385

9.5.4	Problems with Algorithms	386
9.6	Bias and the Production of ‘AI Ethics’	389
9.7	AI, Online Content, Persons, Self, and Society	391
9.8	The Use of AI: Reshaping Ethical Questions?	396
9.9	And Just to Finish	400
9.10	Key Points	401
9.11	Educator Notes	402
	References	402
	Further Reading	406
10	Towards the Future with AI: Work and Superintelligence	409
10.1	Introduction	409
10.2	AI in the Workplace	411
10.2.1	AI in Work: Some Background	412
10.2.2	Why Use AI Rather Than Humans?	414
10.2.3	Assessing the Use of AI to Replace Human Roles or Tasks	415
10.2.4	Workplace Hierarchies and the Nature and Value of Work	417
10.2.5	The Claim that AI Will Relieve Humans of Drudgery	419
10.2.6	Seeing Work Through a Technological Lens	420
10.2.7	The Claim that the Use of AI Will Enable Work to Be More Humanised	422
10.2.8	Other Possible Outcomes of the Introduction of AI	422
10.2.9	A Word on Methods	426
10.3	AI, Employment and Unemployment, and Future Visions	428
10.3.1	Technological Unemployment and Social Divisions	431
10.3.2	Keeping Occupied in a Time of Unemployment	432
10.4	Superintelligence, Existential Risk, and the Control Problem	434
10.4.1	Introduction	434
10.4.2	How Might Superintelligence Come About?	437
10.4.3	Attempting to Find a Solution: Ensure Superintelligence Aligns with Human Values	440
10.4.4	How Do Debates About Superintelligence Relate to Current Issues in AI Ethics?	447
10.5	Longtermism in AI	447
10.6	Key Points	450
10.7	Educator Notes	451
	References	452
	Further Reading	455

11 Our Future with AI: Future Projections and Moral Machines	457
11.1 Introduction	457
11.2 The Purposeful Universe: Value Beyond the Human	458
11.3 Contrasting Frameworks Behind Visions of AI and Approaches to Ethics	465
11.4 The Moral Status of AI and Robot Rights	471
11.4.1 Treating AI 'As If' It Is a Person: Reciprocity and Its Dangers	475
11.5 Machines as Moral Agents: Could Artificial Intelligence Enhance Human Morality?	477
11.5.1 Why Would We Even Want a Machine to Help Us with Moral Issues?	480
11.5.2 Analysing the Different Tasks of Morality	481
11.5.3 Elements of Moral Competence: What We Need in Order to Act Like Moral Agents	482
11.5.4 Is Using a Machine as a Moral Adviser Permissible?	483
11.5.5 How Could We Build a Moral Machine? An Example	485
11.5.6 AI as Moral Assistant: Can It Help Us with the Knowledge Needed for Better Moral Judgements?	487
11.5.7 Data Gathering for Moral Judgements: How Problems May Arise	490
11.5.8 AI, Information, Ethics, and How We Communicate	493
11.6 Key Points	494
11.7 Educator Notes	495
References	496
Further Reading	498
Glossary	501
Index	513