

# Contents

Acknowledgments .....	xi
Author.....	xiii
<b>1 Introduction .....</b>	<b>1</b>
1.1 Introduction .....	1
1.2 The AI Control Problem.....	2
1.3 Obstacles to Controlling AI.....	3
1.4 Defining Safe AI.....	4
1.5 On Governability of AI .....	5
1.6 Conclusions.....	6
1.7 About the Book.....	7
References .....	8
<b>2 Unpredictability.....</b>	<b>11</b>
2.1 Introduction to Unpredictability .....	11
2.2 Predictability: What We Can Predict – A Literature Review.....	13
2.3 Cognitive Uncontainability .....	15
2.4 Conclusions.....	16
References .....	17
<b>3 Unexplainability and Incomprehensibility .....</b>	<b>21</b>
3.1 Introduction .....	22
3.2 Literature Review.....	22
3.3 Unexplainability.....	25
3.4 Incomprehensibility.....	27
3.5 Conclusions.....	30
Notes .....	30
References .....	31
<b>4 Unverifiability .....</b>	<b>36</b>
4.1 On Observers and Verifiers .....	36
4.2 Historical Perspective.....	37
4.3 Classification of Verifiers .....	39
4.4 Unverifiability.....	42
4.5 Unverifiability of Software .....	43
4.5.1 Unverifiability of Artificial Intelligence .....	44
4.6 Conclusions and Future Work .....	45

Notes .....	46
References .....	46
<b>5 Unownability</b> .....	<b>51</b>
5.1 Introduction .....	51
5.1.1 Proposals for Establishing Ownership.....	52
5.2 Obstacles to Ownership .....	52
5.3 Conclusions.....	54
References .....	55
<b>6 Uncontrollability</b> .....	<b>57</b>
6.1 Introduction .....	59
6.2 AI Control Problem.....	60
6.2.1 Types of Control Problems .....	60
6.2.2 Formal Definition.....	62
6.3 Previous Work .....	69
6.3.1 Controllable .....	69
6.3.2 Uncontrollable .....	72
6.4 Proving Uncontrollability .....	78
6.5 Multidisciplinary Evidence for Uncontrollability of AI.....	81
6.5.1 Control Theory .....	82
6.5.2 Philosophy .....	84
6.5.3 Public Choice Theory .....	85
6.5.4 Justice (Unfairness).....	86
6.5.5 Computer Science Theory.....	87
6.5.6 Cybersecurity .....	88
6.5.7 Software Engineering .....	88
6.5.8 Information Technology .....	89
6.5.9 Learnability .....	89
6.5.10 Economics .....	90
6.5.11 Engineering .....	90
6.5.12 Astronomy .....	90
6.5.13 Physics .....	91
6.6 Evidence from AI Safety Research for Uncontrollability of AI.....	91
6.6.1 Value Alignment .....	93
6.6.2 Brittleness.....	94
6.6.3 Unidentifiability.....	95
6.6.4 Uncontainability .....	96
6.6.5 Uninterruptability .....	97
6.6.6 AI Failures.....	98
6.6.7 Unpredictability .....	98
6.6.8 Unexplainability and Incomprehensibility.....	99
6.6.9 Unprovability .....	100
6.6.10 Unverifiability .....	101
6.6.11 Reward Hacking .....	103

6.6.12	Intractability .....	103
6.6.13	Goal Uncertainty .....	104
6.6.14	Complementarity .....	105
6.6.15	Multidimensionality of Problem Space .....	106
6.7	Discussion .....	106
6.8	Conclusions .....	109
	Notes .....	112
	References .....	112
<b>7</b>	<b>Pathways to Danger</b> .....	<b>128</b>
7.1	Taxonomy of Pathways to Dangerous AI .....	128
7.1.1	On Purpose – Pre-Deployment .....	128
7.1.2	On Purpose – Post-Deployment .....	130
7.1.3	By Mistake – Pre-Deployment .....	131
7.1.4	By Mistake – Post-Deployment .....	132
7.1.5	Environment – Pre-Deployment .....	133
7.1.6	Environment – Post-Deployment .....	133
7.1.7	Independently – Pre-Deployment .....	133
7.1.8	Independently – Post-Deployment .....	133
7.2	Conclusions .....	134
	References .....	135
<b>8</b>	<b>Accidents</b> .....	<b>139</b>
8.1	Introduction .....	139
8.2	AI Failures .....	140
8.2.1	Preventing AI Failures .....	146
8.3	AI Safety .....	148
8.4	Cybersecurity vs. AI Safety .....	149
8.5	Conclusions .....	151
	Notes .....	151
	References .....	154
<b>9</b>	<b>Personhood</b> .....	<b>158</b>
9.1	Introduction to AI Personhood .....	158
9.2	Selfish Memes .....	159
9.3	Human Indignity .....	160
9.4	Legal-System Hacking .....	161
9.5	Human Safety .....	162
9.6	Conclusions .....	165
	Notes .....	166
	References .....	166
<b>10</b>	<b>Consciousness</b> .....	<b>170</b>
10.1	Introduction to the Problem of Consciousness .....	170
10.2	Test for Detecting Qualia .....	173

10.3	Computers Can Experience Illusions, and so Are Conscious.....	176
10.3.1	Qualia Computing.....	177
10.4	Purpose of Consciousness.....	178
10.4.1	Qualia Engineering.....	179
10.5	Consciousness and Artificial Intelligence.....	180
10.6	Conclusions and Conjectures.....	182
	Note.....	184
	References.....	184
<b>11</b>	<b>Personal Universes.....</b>	<b>195</b>
11.1	Introduction to the Multi-Agent Value Alignment Problem.....	195
11.2	Individual Simulated Universes.....	196
11.3	Benefits and Shortcomings of Personalized Universes.....	199
11.4	Conclusions.....	200
	Note.....	201
	References.....	201
<b>12</b>	<b>Human <math>\neq</math> AGI.....</b>	<b>206</b>
12.1	Introduction.....	206
12.2	Prior Work.....	207
12.3	Humans Are Not AGI.....	209
12.4	Conclusions.....	211
	Note.....	213
	References.....	213
<b>13</b>	<b>Skepticism.....</b>	<b>217</b>
13.1	Introduction to AI Risk Skepticism.....	217
13.2	Types of AI Risk Skeptics.....	219
13.2.1	Skeptics of Strawman.....	221
13.3	Arguments for AI Risk Skepticism.....	222
13.3.1	Priorities Objections.....	223
13.3.2	Technical Objections.....	225
13.3.3	AI Safety-Related Objections.....	227
13.3.4	Ethical Objections.....	228
13.3.5	Biased Objections.....	229
13.3.6	Miscellaneous Objections.....	230
13.4	Countermeasures for AI Risk Skepticism.....	232
13.5	Conclusions.....	234
	Notes.....	235
	References.....	236
	<b>Index.....</b>	<b>243</b>