CONTENTS

Lists of Figures, Tables, and Boxes xv

1. Past developments and present capabilities 1
Growth modes and big history 1
Great expectations 3
Seasons of hope and despair 5
State of the art 11
Opinions about the future of machine intelligence 18

2. Paths to superintelligence 22
Artificial intelligence 23
Whole brain emulation 30
Biological cognition 36
Brain–computer interfaces 44
Networks and organizations 48
Summary 50

3. Forms of superintelligence 52
Speed superintelligence 53
Collective superintelligence 54
Quality superintelligence 56
Direct and indirect reach 58
Sources of advantage for digital intelligence 59

4. The kinetics of an intelligence explosion 62
Timing and speed of the takeoff 62
Recalcitrance 66
Non-machine intelligence paths 66
Emulation and AI paths 68
Optimization power and explosivity 73

5. Decisive strategic advantage 78
Will the frontrunner get a decisive strategic advantage? 79
How large will the successful project be? 83
Monitoring 84
International collaboration 86
From decisive strategic advantage to singleton 87

6. Cognitive superpowers 91
Functionalities and superpowers 92
An AI takeover scenario 95
Power over nature and agents 99

7. The superintelligent will 105
The relation between intelligence and motivation 105
Instrumental convergence 109
Self-preservation 109
Goal-content integrity 109
Cognitive enhancement 111
Technological perfection 112
Resource acquisition 113

8. Is the default outcome doom? 115
Existential catastrophe as the default outcome of an intelligence explosion? 115
The treacherous turn 116
Malignant failure modes 119
Perverse instantiation 120
Infrastructure profusion 122
Mind crime 125

9. The control problem 127
Two agency problems 127
Capability control methods 129
Boxing methods 129
Incentive methods 131
Stunting 135
Tripwires 137
Motivation selection methods 138
Direct specification 139
Domesticity 140
Indirect normativity 141
Augmentation 142
Synopsis 143

10. Oracles, genies, sovereigns, tools 145
Oracles 145
Genies and sovereigns 148
Tool-AIs 151
Comparison 155

11. Multipolar scenarios 159
Of horses and men 160
Wages and unemployment 160
Capital and welfare 161
The Malthusian principle in a historical perspective 163
Population growth and investment 164
Life in an algorithmic economy 166
Voluntary slavery, casual death 167
Would maximally efficient work be fun? 169
Unconscious outsourcers? 172
Evolution is not necessarily up 173
Post-transition formation of a singleton? 176
A second transition 177
Superorganisms and scale economies 178
Unification by treaty 180

12. Acquiring values 185
The value-loading problem 185
Evolutionary selection 187
Reinforcement learning 188
Associative value accretion 189
Motivational scaffolding 191
Value learning 192
Emulation modulation 201
Institution design 202
Synopsis 207

13. Choosing the criteria for choosing 209
The need for indirect normativity 209
Coherent extrapolated volition 211
Some explications 212
Rationales for CEV 213
Further remarks 216
Morality models 217
Do What I Mean 220
Component list 221
Goal content 222

Decision theory 223
Epistemology 224
Ratification 225
Getting close enough 227

14. The strategic picture 228
Science and technology strategy 228
Differential technological development 229
Preferred order of arrival 230
Rates of change and cognitive enhancement 233
Technology couplings 236
Second-guessing 238
Pathways and enablers 240
Effects of hardware progress 240
Should whole brain emulation research be promoted? 242
The person-affecting perspective favors speed 245
Collaboration 246
The race dynamic and its perils 246
On the benefits of collaboration 249
Working together 253

15. Crunch time 255
Philosophy with a deadline 255
What is to be done? 256
Seeking the strategic light 257
Building good capacity 258
Particular measures 258
Will the best in human nature please stand up 259

Notes 261
Bibliography 305
Index 325