Table of Contents

Preface, xiii

Acknowledgments, xix

About the Author, xxi

Chapter 1 ■ AI-Completeness: The Problem Domain of Superintelligent Machines 1

1.1 INTRODUCTION 1

1.2 THE THEORY OF AI-COMPLETENESS 5

1.2.1 Definitions 6

1.2.2 Turing Test as the First AI-Complete Problem 7

1.2.3 Reducing Other Problems to a TT 8

1.2.4 Other Probably AI-Complete Problems 9

1.3 FIRST AI-HARD PROBLEM: PROGRAMMING 10

1.4 BEYOND AI-COMPLETENESS 11

1.5 CONCLUSIONS 14

REFERENCES 15

Chapter 2 ■ The Space of Mind Designs and the Human Mental Model 21

2.1 INTRODUCTION 21

2.2 INFINITUDE OF MINDS 22

2.3 SIZE, COMPLEXITY, AND PROPERTIES OF MINDS 23

2.4 SPACE OF MIND DESIGNS 25

2.5 A SURVEY OF TAXONOMIES 30

2.6 MIND CLONING AND EQUIVALENCE TESTING ACROSS SUBSTRATES 34

2.7 CONCLUSIONS 35

REFERENCES 35

CHAPTER 3 ■ How to Prove You Invented Superintelligence So No One Else Can Steal It 41

3.1 INTRODUCTION AND MOTIVATION 41

3.2 ZERO KNOWLEDGE PROOF 43

3.3 CAPTCHA 43

3.4 AI-COMPLETENESS 45

3.5 SUPERCAPTCHA 48

3.6 CONCLUSIONS 50

REFERENCES 50

CHAPTER 4 ■ Wireheading, Addiction, and Mental Illness in Machines 57

4.1 INTRODUCTION 57

4.2 WIREHEADING IN MACHINES 60

4.2.1 Sensory Illusions: A Form of Indirect Wireheading 63

4.3 POTENTIAL SOLUTIONS TO THE WIREHEADING PROBLEM 64

4.4 PERVERSE INSTANTIATION 71

4.5 CONCLUSIONS AND FUTURE WORK 74

REFERENCES 76

CHAPTER 5 ■ On the Limits of Recursively Self-Improving Artificially Intelligent Systems 81

5.1 INTRODUCTION 81

5.2 TAXONOMY OF TYPES OF SELF-IMPROVEMENT 83

5.3 ON THE LIMITS OF RECURSIVELY SELF-IMPROVING ARTIFICIALLY INTELLIGENT SYSTEMS 89

5.4 ANALYSIS 93

5.5 RSI CONVERGENCE THEOREM 95
5.6 CONCLUSIONS 96
REFERENCES 98

CHAPTER 6 ▪ Singularity Paradox and What to Do About It 107

6.1 INTRODUCTION TO THE SINGULARITY PARADOX 107
6.2 METHODS PROPOSED FOR DEALING WITH SP 109
6.2.1 Prevention from Development 109
6.2.1.1 Fight Scientists 109
6.2.1.2 Restrict Hardware and Outlaw Research 109
6.2.1.3 Singularity Steward 110
6.2.2 Restricted Deployment 110
6.2.2.1 AI-Box 110
6.2.2.2 Leakproof Singularity 111
6.2.2.3 Oracle AI 111
6.2.2.4 AI Confinement Protocol 112
6.2.3 Incorporation into Society 113
6.2.3.1 Law and Economics 113
6.2.3.2 Religion for Robots 114
6.2.3.3 Education 114
6.2.4 Self-Monitoring 115
6.2.4.1 Hard-Coded Rules 115
6.2.4.2 Chaining God 116
6.2.4.3 Friendly AI 116
6.2.4.4 Humane AI 117
6.2.4.5 Emotions 117
6.2.5 Indirect Solutions 118
6.2.5.1 Why They May Need Us 118
6.2.5.2 Let Them Kill Us 120
6.2.5.3 War Against the Machines 120
6.2.5.4 If You Cannot Beat Them, Join Them 121
6.2.5.5 Other Approaches 121

6.3 ANALYSIS OF SOLUTIONS 122

6.4 FUTURE RESEARCH DIRECTIONS 128

6.5 CONCLUSIONS 128

REFERENCES 129

CHAPTER 7 ■ Superintelligence Safety Engineering 135

7.1 ETHICS AND INTELLIGENT SYSTEMS 135

7.2 ARTIFICIAL INTELLIGENCE SAFETY ENGINEERING 136

7.3 GRAND CHALLENGE 138

7.4 ARTIFICIAL GENERAL INTELLIGENCE RESEARCH IS UNETHICAL 138

7.5 ROBOT RIGHTS 140

7.6 CONCLUSIONS 140

REFERENCES 141

CHAPTER 8 ■ Artificial Intelligence Confinement Problem (and Solution) 145

8.1 INTRODUCTION 145

8.1.1 Artificial Intelligence Confinement Problem 146

8.2 HAZARDOUS SOFTWARE 147

8.3 CRITIQUE OF THE CONFINEMENT APPROACH 148

8.4 POSSIBLE ESCAPE PATHS 148

8.4.1 Social Engineering Attacks 149

8.4.2 System Resource Attacks 150

8.4.3 Beyond Current Physics Attacks 151

8.4.4 Pseudoscientific Attacks 152

8.4.5 External Causes of Escape 153

8.4.6 Information In-Leaking 153

8.5 CRITIQUE OF THE AI-BOXING CRITIQUE 154

8.6 COUNTERMEASURES AGAINST ESCAPE 154

8.6.1 Preventing Social Engineering Attacks 155

8.6.2 Preventing System Resource Attacks and Future Threats 155
8.6.3 Preventing External Causes of Escape 156
8.6.4 Preventing Information In-Leaking 156
8.7 AI COMMUNICATION SECURITY 157
8.8 HOW TO SAFELY COMMUNICATE WITH A SUPERINTELLIGENCE 159
8.9 CONCLUSIONS AND FUTURE WORK 161
REFERENCES 162

CHAPTER 9 ■ Efficiency Theory: A Unifying Theory for Information, Computation, and Intelligence 167

9.1 INTRODUCTION 167
9.2 EFFICIENCY THEORY 168
9.3 INFORMATION AND KNOWLEDGE 168
9.4 INTELLIGENCE AND COMPUTATION 172
9.5 TIME AND SPACE 174
9.6 COMPRESSIBILITY AND RANDOMNESS 175
9.7 ORACLES AND UNDECIDABILITY 176
9.8 INTRACTABLE AND TRACTABLE 177
9.9 CONCLUSIONS AND FUTURE DIRECTIONS 177
REFERENCES 180

CHAPTER 10 ■ Controlling the Impact of Future Superintelligence 185

10.1 WHY I WROTE THIS BOOK 185
10.2 MACHINE ETHICS IS A WRONG APPROACH 185
10.3 CAN THE PROBLEM BE AVOIDED? 187
REFERENCES 189

INDEX, 191