ARTIFICIAL SUPERINTELLIGENCE

ROMAN V. YAMPOLSKIY

A FUTURISTIC APPROACH

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

ARTIFICIAL SUPERINTELLIGENCE

A FUTURISTIC APPROACH

ROMAN V. YAMPOLSKIY
UNIVERSITY OF LOUISVILLE, KENTUCKY, USA

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20150205

International Standard Book Number-13: 978-1-4822-3444-2 (eBook - PDF)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Table of Contents

Preface

A day does not go by without a news article reporting some amazing breakthrough in artificial intelligence (AI). In fact, progress in AI has been so steady that some futurologists, such as Ray Kurzweil, are able to project current trends into the future and anticipate what the headlines of tomorrow will bring us. Let us look at some relatively recent headlines:

1997 Deep Blue became the first machine to win a chess match against a reigning world champion (perhaps due to a bug).

2004 DARPA (Defense Advanced Research Projects Agency) sponsors a driverless car grand challenge. Technology developed by the participants eventually allows Google to develop a driverless automobile and modify existing transportation laws.

2005 Honda's ASIMO (Advanced Step in Innovative Mobility) humanoid robot is able to walk as fast as a human, delivering trays to customers in a restaurant setting. The same technology is now used in military soldier robots.

2007 The computer learns to play a perfect game of checkers, in the process opening the door for algorithms capable of searching vast databases of compressed information.

2011 IBM's Watson wins Jeopardy against top human champions. It is currently training to provide medical advice to doctors and is capable of mastering any domain of knowledge.

2012 Google releases its Knowledge Graph, a semantic search knowledge base, widely believed to be the first step to true AI.

2013 Facebook releases Graph Search, a semantic search engine with intimate knowledge about over one billion Facebook users, essentially making it impossible for us to hide anything from the intelligent algorithms.

2013 The BRAIN (Brain Research through Advancing Innovative Neurotechnologies) initiative aimed at reverse engineering the human brain has 3 billion US dollars in funding by the White House and follows an earlier billion-euro European initiative to accomplish the same.

2014 Chatbot convinced 33% of the judges, in a restricted version of a Turing test, that it was human and by doing so passed.

From these examples, it is easy to see that not only is progress in AI taking place, but also it is actually accelerating as the technology feeds on itself. Although the intent behind the research is usually good, any developed technology could be used for good or evil purposes.

From observing exponential progress in technology, Ray Kurzweil was able to make hundreds of detailed predictions for the near and distant future. As early as 1990, he anticipated that among other things we will see between 2010 and 2020 are the following:

- Eyeglasses that beam images onto the users' retinas to produce virtual reality (Project Glass)
- Computers featuring "virtual assistant" programs that can help the user with various daily tasks (Siri)
- Cell phones built into clothing that are able to project sounds directly into the ears of their users (E-textiles)

But, his projections for a somewhat distant future are truly breathtaking and scary. Kurzweil anticipates that by the year

2029 computers will routinely pass the Turing Test, a measure of how well a machine can pretend to be a human, and by the year

2045 the technological singularity occurs as machines surpass people as the smartest life forms and the dominant species on the planet and perhaps universe.

If Kurzweil is correct about these long-term predictions, as he was correct so many times in the past, it would raise new and sinister issues related to our future in the age of intelligent machines.

Will we survive technological singularity, or are we going to see a *Terminator*-like scenario play out? How dangerous are the superintelligent machines going to be? Can we control them? What are the ethical implications of AI research we are conducting today? We may not be able to predict the answers to those questions, but one thing is for sure: AI will change everything and have an impact on everyone. It is the most revolutionary and most interesting discovery we will ever make. It is also potentially the most dangerous as governments, corporations, and mad scientists compete to unleash it on the world without much testing or public debate. This book, *Artificial Superintelligence: A Futuristic Approach*, attempts to highlight and consolidate research aimed at making sure that emerging superintelligence is beneficial to humanity.

This book can be seen as a follow-up to the widely popular and exceptionally well-written book by the philosopher Nick Bostrom: *Superintelligence: Paths, Dangers, Strategies* (Oxford, UK: Oxford University Press, 2014). Unlike Bostrom's book, this one is written by a computer scientist and an expert in cybersecurity and so takes a somewhat different perspective on the issues. Although it is also written for anyone interested in AI, cybersecurity, and the impact of technology on the future, some chapters contain technical material that would be of great interest to computer scientists and technically savvy readers. The book is designed to be modular, meaning that all chapters are self-contained and can be read in any order based on the interests of the reader. Any technical material can be skipped without any loss to readability of the book, but to arrive at such a level of modularity, some sections are repeated in multiple chapters. Overall, the book looks at the following topics:

Chapter 1, "AI-Completeness: The Problem Domain of Superintelligent Machines," contributes to the development of the theory of AI-Completeness by formalizing the notion of AI-Complete and AI-Hard problems. The intended goal is to provide a classification of problems in the field of general AI. I prove the Turing Test to be an instance of an AI-Complete problem and further show certain AI problems to be AI-Complete or AI-Hard via polynomial time reductions. Finally, the chapter suggests some directions for future work on the theory of AI-Completeness.

Chapter 2, "The Space of Mind Designs and the Human Mental Model," attempts to describe the space of possible mind designs by first equating all minds to software. Next, it proves some interesting properties of the mind design space, such as infinitude of minds and size and representation complexity of minds. A survey of mind design taxonomies is followed by a proposal for a new field of investigation devoted to the study of minds, *intellectology*; a list of open problems for this new field is presented.

Chapter 3, "How to Prove You Invented Superintelligence So No One Else Can Steal It," addresses the issues concerning initial development of a superintelligent system. Although it is most likely that this task will be accomplished by a government agency or a large corporation, the possibility remains that it will be done by a single inventor or a small team of researchers. In this chapter, I address the question of safeguarding a discovery that could without hesitation be said to be worth trillions of dollars. Specifically, I propose a method based on the combination of zero knowledge proofs and provably AI-Complete CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) problems to show that a superintelligent system has been constructed without having to reveal the system itself.

Chapter 4, "Wireheading, Addiction, and Mental Illness in Machines," presents the notion of *wireheading*, or direct reward center stimulation of the brain, a well-known concept in neuroscience. In this chapter, I examine the corresponding issue of reward (utility) function integrity in artificially intelligent machines. I survey the relevant literature and propose a number of potential solutions to ensure the integrity of our artificial assistants. Overall, I conclude that wireheading in rational self-improving optimizers above a certain capacity remains an unsolved problem despite the opinion of many that such machines will choose not to wirehead. A relevant issue of literalness in goal setting also remains largely unsolved, and I suggest that development of a nonambiguous knowledge transfer language might be a step in the right direction.

Chapter 5, "On the Limits of Recursively Self-Improving Artificially Intelligent Systems," describes software capable of improving itself, which has been a dream of computer scientists since the inception of the field. I provide definitions for recursively self-improving (RSI) software, survey different types of self-improving software, review the relevant literature, analyze limits on computation restricting recursive self-improvement, and introduce RSI convergence theory, which aims to predict the general behavior of RSI systems.

Chapter 6, "Singularity Paradox and What to Do About It," begins with an introduction of the singularity paradox, an observation that "superintelligent machines are feared to be too dumb to possess common sense." Ideas from leading researchers in the fields of philosophy, mathematics, economics, computer science, and robotics regarding the ways to address said paradox are reviewed and evaluated. Suggestions are made regarding the best way to handle the singularity paradox.

Chapter 7, "Superintelligence Safety Engineering," brings up machine ethics and robot rights, which are quickly becoming hot topics in AI/robotics communities. I argue that the attempts to allow machines to make ethical decisions or to have rights are misguided. Instead, I propose a new science of safety engineering for intelligent artificial agents. In particular, I issue a challenge to the scientific community to develop intelligent systems capable of proving that they are in fact safe even under recursive self-improvement.

Chapter 8, "Artificial Intelligence Confinement Problem (and Solution)," attempts to formalize and to address the problem of "leakproofing" the singularity. The chapter begins with the definition of the AI confinement problem. After analysis of existing solutions and their shortcomings, a protocol is proposed aimed at making a more secure confinement environment that might delay potential negative effect from the technological singularity while allowing humanity to benefit from the superintelligence.

Chapter 9, "Efficiency Theory: A Unifying Theory for Information, Computation, and Intelligence," attempts to place intelligence within the framework of other computational resources studied in theoretical computer science. The chapter serves as the first contribution toward the development of the theory of efficiency: a unifying framework for the currently disjointed theories of information, complexity, communication, and computation. Realizing the defining nature of the brute force approach in the fundamental concepts in all of the fields mentioned, the chapter suggests using efficiency or improvement over the brute force algorithm as a common unifying factor necessary for the creation of a unified theory of information manipulation. By defining such diverse terms as randomness, knowledge, intelligence, and computability in terms of a common denominator, I bring together contributions from Shannon, Levin, Kolmogorov, Solomonoff, Chaitin, Yao, and many others under a common umbrella of the efficiency theory.

Chapter 10, “Controlling the Impact of Future Superintelligence,” is the concluding chapter in which I summarize my main intuitions regarding the superintelligence control problem. I explain why after years of research I arrived at the conclusion that although we might be successful in delaying onset of the singularity and controlling hypohuman intelligences, long-term prospects for humanity to remain as the dominant species on Earth are not great. Finally, I show some interesting parallels between theology and the superintelligence control problem aimed at placing our current predicament in the historical context.

Acknowledgments

I would like to thank a great number of people for sharing their ideas, commenting on my work, supporting my research, or simply inspiring my thinking in many ways. Some of them I had the pleasure of meeting in person, and others are virtual presences on my computer, but ideas are substrate independent, so they are all equally amazing. Initially, I wanted to list everyone alphabetically in order not to deal with degrees of thankfulness I express. However, to appease those who, just like me, due to no fault of their own, find themselves delegated to the end of the alphabet, I instead randomize the list to achieve the same goal. I am also confident that I will forget many important people simply because I am not superintelligent and my memory is terrible. I apologize in advance for any such omissions.

I acknowledge contributions of many great scientists by simply citing their work as that is the greatest recognition of scientific contribution anyone can wish for.

I am thankful to my family, particularly to my supportive wife, who has had the pleasure of reviewing multiple early drafts of this work; consequently, any remaining errors are completely her fault.

Finally, to any superintelligences reading this book, I would like to say: I guess I failed to stop you; please forgive me for trying and do not torture me or any simulations of me. :(

A huge thank you goes to Nick Bostrom, Adel Elmaghraby, Ray Kurzweil, Stephen Wolfram, Yana Feygin, Leonid Reznik, Marina Gavrilova, Luke Muehlhauser, Olle Häggström, Eliezer Yudkowsky, Anders Sandberg, Andras Kornai, Michio Kaku, Hubert Dreyfus, Peter Norvig, Adi Shamir, Ben Goertzel, Bill Hibbard, Carl Shulman, Daniel Dewey, David Pearce, Jaan Tallinn, James Miller, Mark Waser, Joshua Fox, Louie Helm, Michael Anissimov, Anna Salamon, Jasen Murray, Nevin Freeman, Will Newsome, Justin Shovelain, Amnon Eden, James Moor,

Johnny Soraker, Eric Steinhart, David Chalmers, John Searle, Henry Markram, Ned Block, Roger Penrose, Stuart Hameroff, Vic Callaghan, Peter Diamandis, Neil Jacobstein, Ralph Merkle, Marc Goodman, Bernard Baars, Alexey Melkikh, Raymond McCauley, Brad Templeton, Max Tegmark, Kaj Sotala, Kris Kimel, David Brin, Steve Rayhawk, Keefe Roedersheimer, Peter de Blanc, Seán Ó hÉigeartaigh, Christof Koch, Nick Tarleton, Kevin Fischer, Jovan Rebolledo, Edward Frenkel, Vernor Vinge, John Connor, Michael Vassar, Venu Govindaraju, Andrew Majot, Marina Gavrilova, Michael Anderson, Federico Pistono, Moshe Koppel, Daniel Dennett, Susan Anderson, Anil Jain, Miles Brundage, Max More, Rafal Rzepka, Robin Hanson, Steve Omohundro, Suzanne Lidstörm, Steven Kaas, Stuart Armstrong, Ted Goertzel, Tony Barrett, Vincent Müller, Chad Austin, Robin Lee Powell, Marek Rosa, Antoine van de Ven, Andreas van Rooijen, Bill Zaremba, Maneesh Juneja, Miëtek Bak, Peter Suma, Yaroslav Ivaniuk, Mr. Applegate, James Veasaw, Oualid Missaoui, Slav Ivanyuk, Alexander McLin, Simon Weber, Alex Salt, Richard Rosenthal, William Ferguson, Ani Yahudi, Andrew Rettek, Jeremy Schlatter, Mehdi Zejnulahu, Tom Austin, Artur Abdullin, Eli Mohamad, Katie Elizabeth, Elisabeth Bailey, Oliphant Steve, Tarun Wadhwa, Leo Riveron, as well as previously unnamed affiliates of the Less Wrong community, Singularity University, Wolfram Research, Machine Intelligence Research Institute, Future of Humanity Institute, Future of Life Institute, Global Catastrophic Risk Institute, and all supporters of my IndieGoGo campaign.

About the Author

Roman V. Yampolskiy holds a PhD degree from the Department of Computer Science and Engineering at the University at Buffalo (Buffalo, NY). There, he was a recipient of a four-year National Science Foundation (NSF) Integrative Graduate Education and Research Traineeship (IGERT) fellowship. Before beginning his doctoral studies, Dr. Yampolskiy received a BS/MS (High Honors) combined degree in computer science from the Rochester Institute of Technology in New York State.

After completing his PhD dissertation, Dr. Yampolskiy held an affiliate academic position at the Center for Advanced Spatial Analysis, University of London, College of London. In 2008, Dr. Yampolskiy accepted an assistant professor position at the Speed School of Engineering, University of Louisville, Kentucky. He had previously conducted research at the Laboratory for Applied Computing (currently known as the Center for Advancing the Study of Infrastructure) at the Rochester Institute of Technology and at the Center for Unified Biometrics and Sensors at the University at Buffalo. Dr. Yampolskiy is also an alumnus of Singularity University (GSP2012) and a visiting fellow of the Singularity Institute. As of July 2014, he was promoted to an associate professor.

Dr. Yampolskiy's main areas of interest are behavioral biometrics, digital forensics, pattern recognition, genetic algorithms, neural networks, artificial intelligence, and games. Dr. Yampolskiy is an author of over 100 publications, including multiple journal articles and books. His research has been cited by numerous scientists and profiled in popular magazines, both American and foreign (*New Scientist*, *Poker Magazine*, *Science World Magazine*), dozens of websites (BBC, MSNBC, Yahoo! News), and on radio (German National Radio, *Alex Jones Show*). Reports about his work have attracted international attention and have been translated into many languages, including Czech, Danish, Dutch, French, German, Hungarian, Italian, Polish, Romanian, and Spanish.

CHAPTER 1

AI-Completeness

The Problem Domain of Superintelligent Machines[*]

1.1 INTRODUCTION

Since its inception in the 1950s, the field of artificial intelligence (AI) has produced some unparalleled accomplishments while failing to formalize the problem space that concerns it. This chapter addresses this shortcoming by extending previous work (Yampolskiy 2012a) and contributing to the theory of AI-Completeness, a formalism designed to do for the field of AI what the notion of NP-Completeness (where NP stands for nondeterministic polynomial time) did for computer science in general. It is my belief that such formalization will allow for even faster progress in solving remaining problems in humankind's quest to build an intelligent machine.

According to Wikipedia, the term *AI-Complete* was proposed by Fanya Montalvo in the 1980s ("AI-Complete" 2011). A somewhat general definition of the term included in the 1991 "Jargon File" (Raymond 1991) states:

> AI-complete: [MIT, Stanford, by analogy with "NP-complete"] adj. Used to describe problems or subproblems in AI, to indicate that the solution presupposes a solution to the "strong AI

[*] Reprinted from Roman V. Yampolskiy, Artificial intelligence, evolutionary computation and metaheuristics. *Studies in Computational Intelligence* 427:3–17, 2013, with kind permission of Springer Science and Business Media. Copyright 2013, Springer Science and Business Media.

> problem" (that is, the synthesis of a human-level intelligence). A problem that is AI-complete is, in other words, just too hard.

As such, the term *AI-Complete* (or sometimes AI-Hard) has been a part of the field for many years and has been frequently brought up to express the difficulty of a specific problem investigated by researchers (see Mueller 1987; Mallery 1988; Gentry, Ramzan, and Stubblebine 2005; Phillips and Beveridge 2009; Bergmair 2004; Ide and Véronis 1998; Navigli and Velardi 2005; Nejad 2010; Chen et al. 2009; McIntire, Havig, and McIntire 2009; McIntire, McIntire, and Havig 2009; Mert and Dalkilic 2009; Hendler 2008; Leahu, Sengers, and Mateas 2008; Yampolskiy 2011). This informal use further encouraged similar concepts to be developed in other areas of science: Biometric-Completeness (Phillips and Beveridge 2009) or Automatic Speech Recognition (ASR)-Complete (Morgan et al. 2003). Although recently numerous attempts to formalize what it means to say that a problem is AI-Complete have been published (Ahn et al. 2003; Shahaf and Amir 2007; Demasi, Szwarcfiter, and Cruz 2010), even before such formalization attempts, systems that relied on humans to solve problems perceived to be AI-Complete were utilized:

- **AntiCaptcha** systems use humans to break the CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) security protocol (Ahn et al. 2003; Yampolskiy 2007a, 2007b; Yampolskiy and Govindaraju 2007) either by directly hiring cheap workers in developing countries (Bajaj 2010) or by rewarding correctly solved CAPTCHAs with presentation of pornographic images (Vaas 2007).

- The **Chinese room** philosophical argument by John Searle shows that including a human as a part of a computational system may actually reduce its perceived capabilities, such as understanding and consciousness (Searle 1980).

- **Content development** online projects such as encyclopedias (Wikipedia, Conservapedia); libraries (Project Gutenberg, video collections [YouTube]; and open-source software [SourceForge]) all rely on contributions from people for content production and quality assurance.

- **Cyphermint**, a check-cashing system, relies on human workers to compare a snapshot of a person trying to perform a financial