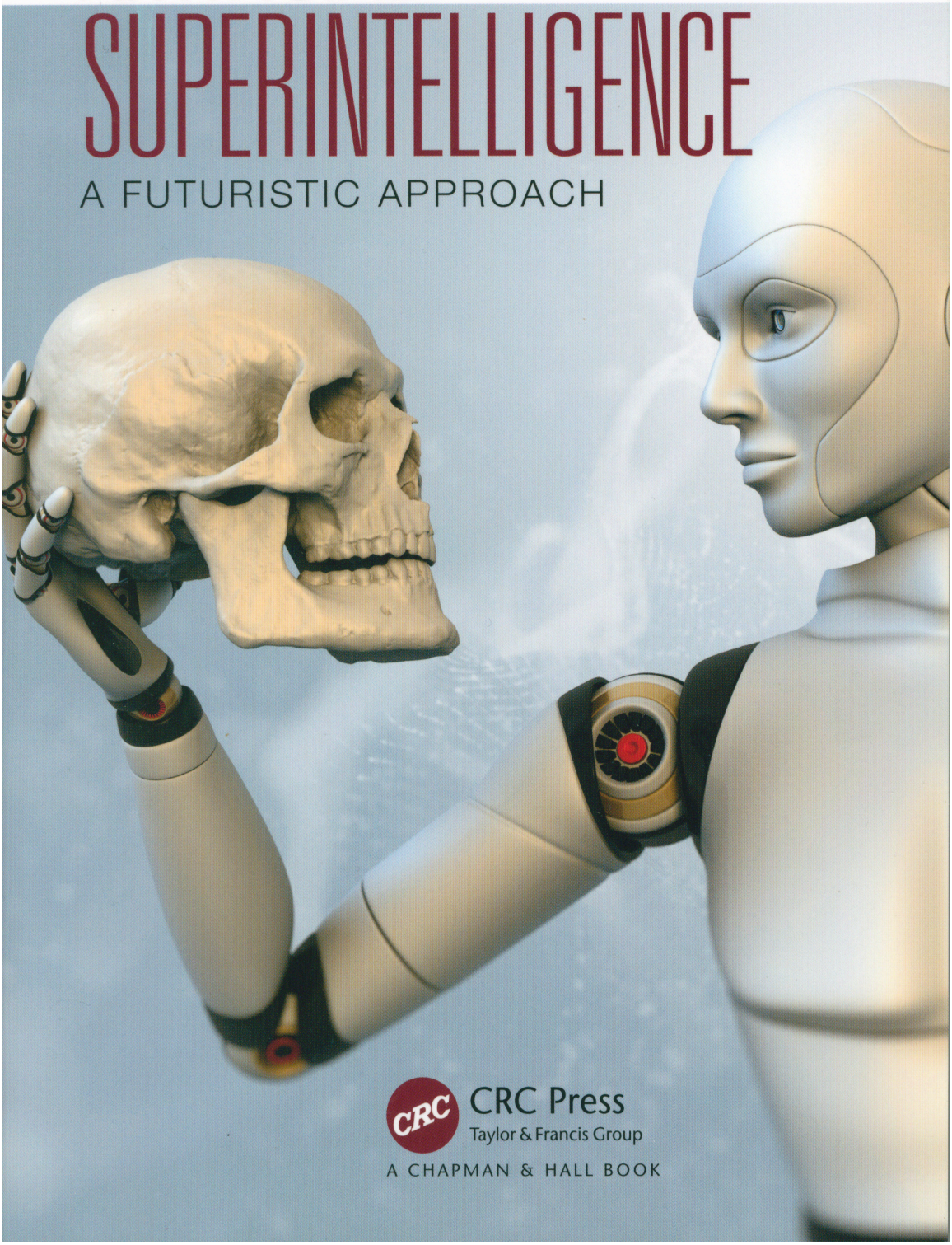
ARTIFICIAL SUPERINTELLIGENCE

ROMAN V. YAMPOLSKIY

A FUTURISTIC APPROACH



CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

ARTIFICIAL SUPERINTELLIGENCE

A FUTURISTIC APPROACH

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

Printed on acid-free paper
Version Date: 20150205

International Standard Book Number-13: 978-1-4822-3443-5 (Paperback)

Library of Congress Cataloging-in-Publication Data

Yampolskiy, Roman V., 1979-
Artificial superintelligence : a futuristic approach / Roman V. Yampolskiy.
pages cm
Includes bibliographical references and index.
ISBN 978-1-4822-3443-5
1. Artificial intelligence--Safety measures. 2. Artificial intelligence--Social aspects. I. Title.

Q335.Y36 2016
006.3--dc23 2015001824

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

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

To Max and Liana, the only two human-level intelligences I was able to create so far. They are also my best reasons to think that one cannot fully control any intelligent agent.

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 6

Singularity Paradox and What to Do About It*

6.1 INTRODUCTION TO THE SINGULARITY PARADOX

Many philosophers, futurologists, and artificial intelligence (AI) researchers (Solomonoff 1985; Bostrom 2006; Yudkowsky 2007, 2008; Hawking 1998; Kurzweil 2005; "Tech Luminaries" 2008) have conjectured that in the next 20 to 200 years a machine capable of at least human-level performance on all tasks will be developed. Because such a machine would, among other things, be capable of designing the next generation of even smarter intelligent machines, it is generally assumed that an intelligence explosion will take place shortly after such a technological self-improvement cycle begins (Good 1966). Although specific predictions regarding the consequences of such an intelligence singularity are varied from potential economic hardship (Hanson 2008) to the complete extinction of humankind (Yudkowsky 2008; Bostrom 2006), many of the involved researchers agree that the issue is of utmost importance and needs to be seriously addressed (Chalmers 2010).

Investigators concerned with the existential risks posed to humankind by the appearance of superintelligence often describe what I shall call a *singularity paradox* (SP) as their main reason for thinking that humanity might

* Reprinted from Roman V. Yampolskiy, *Studies in Applied Philosophy, Epistemology and Rational Ethics* 5:397–413, 2013, with kind permission of Springer Science and Business Media.

be in danger. Briefly, SP could be described as follows: Superintelligent machines are feared to be too dumb to possess common sense.

SP is easy to understand via some commonly cited examples. Suppose that scientists succeed in creating a superintelligent machine and order it to "make all people happy." Complete happiness for humankind is certainly a noble and worthwhile goal, but perhaps we are not considering some unintended consequences of giving such an order. Any human immediately understands what is meant by this request; a nonexhaustive list may include making all people healthy, wealthy, beautiful, and talented and giving them loving relationships and novel entertainment. However, many alternative ways of making all people happy could be derived by a superintelligent machine. For example:

- Killing all people trivially satisfies this request as with 0 people, all of them are happy.
- Forced lobotomies for every man, woman, and child might also accomplish the same goal.
- A simple observation that happy people tend to smile may lead to forced plastic surgeries to affix permanent smiles to all human faces.
- A daily cocktail of cocaine, methamphetamine, methylphenidate, nicotine, and 3,4-methylenedioxymethamphetamine, better known as ecstasy, may do the trick.

An infinite number of other approaches to accomplish universal human happiness could be derived. For a superintelligence, the question is simply which one is fastest/cheapest (in terms of computational resources) to implement. Such a machine clearly lacks common sense, hence the paradox.

We want our machines to do what we want, not what we tell them to do, but as bugs in our programs constantly teach us, this is not a trivial task. The next section of this chapter presents an overview of different approaches proposed for either dealing with the SP or avoiding it all together. In particular, many of the reviewed ideas address a generalized version of the SP that could be stated as follows: We build this machine to have a property *X*, but it actually does ~*X*. Here, *X* could stand for the original goal of happiness or it could represent any of its components, such as security, prosperity, socialization, and so on.

6.2 METHODS PROPOSED FOR DEALING WITH SP

6.2.1 Prevention from Development

6.2.1.1 Fight Scientists

One of the earliest and most radical critics of the upcoming singularity was Theodore Kaczynski, a Harvard-educated mathematician also known as the Unabomber. His solution to prevent singularity from ever happening was a bloody multiyear terror campaign against university research labs across the United States. In his 1995 manifesto, Kaczynski explains his negative views regarding the future of humankind dominated by machines: "If the machines are permitted to make all their own decisions, we can't make any conjectures as to the results, because it is impossible to guess how such machines might behave. We only point out that the fate of the human race would be at the mercy of the machines" (Kaczynski 1995, 79).

An even more violent outcome is prophesized, but not advocated, by Hugo de Garis (2005), who predicts that the issue of building superintelligent machines will split humanity into two camps, eventually resulting in a civil war over the future of singularity research: "I believe that the ideological disagreements between these two groups on this issue will be so strong, that a major … war, killing billions of people, will be almost inevitable before the end of the 21st century" (2005, 234).

6.2.1.2 Restrict Hardware and Outlaw Research

Realizing the potential dangers of superintelligent computers, Anthony Berglas proposed a legal solution to the problem. He suggested outlawing production of more powerful processors, essentially stopping Moore's law in its tracks and consequently denying necessary computational resources to self-improving artificially intelligent machines (Berglas 2009). Similar laws aimed at promoting human safety have been passed banning research on cloning of human beings and development of biological (1972 Biological Weapons Convention), chemical (1993 Chemical Weapons Convention), and nuclear weaponry. Berglas's idea may be interesting in terms of its shock value, which in turn may attract more attention to the dangers of the SP. Here is what Berglas suggested in his own words: "a radical solution, namely to limit the production of ever more powerful computers and so try to starve any AI of processing power. This is urgent, as computers are already almost powerful enough to host an artificial intelligence. … One major problem is that we may already have sufficient power in general purpose computers to support intelligence" (Berglas 2009).

Alternatively, restrictions could be placed on the intelligence an AI may possess to prevent it from becoming superintelligent (Gibson 1984) or legally require that its memory be erased after every job (Benford 1988). Similarly, Bill Joy advocates for relinquishment of superintelligence research and even suggests how enforcement of such a convention could be implemented (Joy 2000): "Enforcing relinquishment will require a verification regime similar to that for biological weapons, but on an unprecedented scale. … Verifying compliance will also require that scientists and engineers adopt a strong code of ethical conduct, resembling the Hippocratic oath, and that they have the courage to whistleblow as necessary, even at high personal cost."

For enforcement of such technology, restricting laws will not be trivial unless the society as a whole adopts an Amish-like, technology-free, lifestyle.

6.2.1.3 Singularity Steward

Ben Goertzel, a computer scientist, has proposed creation of a "big brother AI" monitoring system he calls the "singularity steward." The goal of the proposed system is to monitor the whole world with the specific aim of preventing development of any technology capable of posing a risk to humanity, including superintelligent machines (Goertzel 2004b). Goertzel believes that creation of such a system is feasible and would safeguard humanity against preventable existential risks. Goertzel (2004b) also claims that "in the AI Big Brother case, one doesn't want the AI to be self-modifying and self-improving—one wants it to remain stable. … One needs to make it a bit smarter than humans, but not too much—and one needs to give it a goal system focused on letting itself and humans remain as much the same as possible."

6.2.2 Restricted Deployment

6.2.2.1 AI-Box

A common theme in singularity discussion forums is the possibility of simply keeping a superintelligent agent in sealed hardware to prevent it from doing any harm to humankind. Such ideas originate with scientific visionaries such as Eric Drexler, who has suggested confining transhuman machines so that their outputs could be studied and used safely (Drexler 1986). The general consensus on such an approach among researchers seems to be that such confinement is impossible to successfully maintain. For example, Vernor Vinge has strongly argued against the case

of physical confinement (Vinge 1993): "Imagine yourself locked in your home with only limited data access to the outside, to your masters. If those masters thought at a rate—say—one million times slower than you, there is little doubt that over a period of years (your time) you could come up with 'helpful advice' that would incidentally set you free."

Likewise, David Chalmers, a philosopher, has stated that confinement is impossible because any useful information we would be able to extract from the AI will affect us, defeating the purpose of confinement (Chalmers 2010). However, the researcher who did the most to discredit the idea of the so-called AI-Box is Eliezer Yudkowsky, who has actually performed AI-Box "experiments" in which he demonstrated that even human-level intelligence is sufficient to escape from an AI-Box (Yudkowsky 2002). In a series of five experiments, Yudkowsky challenged different individuals to play a role of a gatekeeper to a superintelligent agent (played by Yudkowsky himself) trapped inside an AI-Box and was successful in securing his release in three of five trials via nothing more than a chat interface (Yudkowsky 2002).

6.2.2.2 Leakproof Singularity

In 2010, David Chalmers proposed the idea of a "leakproof" singularity. He suggests that, for safety reasons, first AI systems be restricted to simulated virtual worlds until their behavioral tendencies can be fully understood under the controlled conditions. Chalmers argues that even if such an approach is not foolproof, it is certainly safer than building AI in physically embodied form. However, he also correctly observes that a truly leakproof system in which no information is allowed to leak out from the simulated world into our environment "is impossible, or at least pointless" (Chalmers 2010, 38) because we cannot interact with the system or even observe it. Chalmers's discussion of the leakproof singularity is an excellent introduction to the state-of-the-art thinking in the field: "The obvious suggestion is that we should first create AI and AI+ systems in virtual worlds: simulated environments that are themselves realized inside a computer. Then an AI will have free reign within its own world without being able to act directly on ours" (Chalmers 2010, 37).

6.2.2.3 Oracle AI

Nick Bostrom, a futurologist, has proposed (Bostrom 2008) an idea for an oracle AI (OAI), which would only be capable of answering questions. It is easy to elaborate and see that a range of different OAIs is possible. From advanced OAIs capable of answering any question to domain expert AIs capable of answering Yes/

No/Unknown to questions on a specific topic. It is claimed that an OAI could be used to help humankind build a safe, unrestricted superintelligent machine.

6.2.2.4 AI Confinement Protocol

I proposed a formalized notion of AI confinement protocol (cp) that represents "AI-Boxing" as a computer security challenge. I define the AI confinement problem (AICP) as the challenge of restricting an artificially intelligent entity to a confined environment from which it cannot exchange information with the outside environment via legitimate or covert channels if such information exchange was not authorized by the confinement authority. An AI system that succeeds in violating the CP protocol is said to have escaped (Yampolskiy 2012).

The proposed protocol is based on the idea of asking what are called "safe questions" with restricted answers. A *safe question* is defined as one for which a human being could find an answer without any help from superintelligence, all on his or her own merit. Why is this useful? Suppose that a scientist working on a cure for cancer is looking at two potential cures. Each one could be developed and tested without assistance from the AI but would take about 3 years of hard work each. Which one is the scientist to try first if he or she believes that they have about an equal chance of working? If a wrong option is chosen to be tested first, humanity will still obtain a cure for cancer, but it will happen 3 years later. What if we could ask the AI to suggest which option to try first?

This question is safe for a number of reasons. First, each potential answer has an equal probability of being correct. Second, a human being could answer it without assistance from the AI, so asking the AI for help will only speed progress but in no way change the outcome. It is essentially equal to being lucky at guessing multiple-choice answers. Finally, the answer to this question could be encoded in a single bit, making it hard to hide additional messages in the provided answer.

To make sure that a potential question is safe, a group of experts could be used to review it. The experts should all be trained AI safety engineers, meaning that they are familiar with the design of the AI and its confinement environment as well as the latest developments in machine ethics (Allen, Wallach, and Smit 2006; Anderson and Anderson 2007; Hall 2000; Moor 2006; Tonkens 2009). Experts may also need to be trained in computer psychology, a currently nonexistent profession that might become a reality in the future (Epstein 1997). An existing discipline that might be of greatest help for training of AI question review experts is artimetrics, a field

of study I proposed that identifies, classifies, and authenticates AI agents, robots, and virtual reality avatars for security purposes (Yampolskiy 2007; Yampolskiy and Govindaraju 2007, 2008; Gavrilova and Yampolskiy 2010).

6.2.3 Incorporation into Society

6.2.3.1 Law and Economics

Robin Hanson has suggested that as long as future intelligent machines are law abiding, they should be able to coexist with humans (Hanson 2009): "In the early to intermediate era when robots are not vastly more capable than humans, you'd want peaceful law-abiding robots as capable as possible, so as to make productive partners. You might prefer they dislike your congestible goods, like your scale-economy goods, and vote like most voters, if they can vote."

Similarly, Hans Moravec puts his hopes for humanity in the hands of the law. He sees forcing cooperation from the robot industries as the most important security guarantee for humankind and integrates legal and economic measures into his solution (Joy 2000): "In a completely free marketplace, superior robots would surely affect humans. … Robotic industries would compete vigorously among themselves for matter, energy, and space, incidentally driving their price beyond human reach. … Judiciously applied, governmental coercion could support human populations in high style on the fruits of robot labor, perhaps for a long while."

Robin Hanson, an economist, agrees: "Robots well-integrated into our economy would be unlikely to exterminate us" (Hanson 2008, 50). Similarly, Steve Omohundro uses microeconomic theory to speculate about the driving forces in the behavior of superintelligent machines. He argues that intelligent machines will want to self-improve, be rational, preserve their utility functions, prevent counterfeit utility, acquire resources and use them efficiently, and protect themselves. He believes that machines' actions will be governed by rational economic behavior (Omohundro 2007, 2008).

Mark Waser suggests an additional "drive" to be included in the list of behaviors predicted to be exhibited by the machines (Waser 2010b). Namely, he suggests that evolved desires for cooperation and being social are part of human ethics and are a great way of accomplishing goals, an idea also analyzed by Joshua Fox and Carl Shulman (2010). Bill Hibbard adds the desire for maintaining the social contract toward equality as a component of ethics for superintelligent machines (Hibbard 2005a), and J. Storrs Hall argues for

incorporation of moral codes into the design (Hall 2000). In general, ethics for superintelligent machines is one of the most fruitful areas of research in the field of singularity research, with numerous publications appearing every year (Shulman, Jonsson, and Tarleton 2009; Bostrom and Yudkowsky 2011; Bostrom 2006; Sotala 2009; Shulman, Tarleton, and Jonsson 2009; Waser 2010a; Bugaj and Goertzel 2007).

6.2.3.2 Religion for Robots

Robert Geraci, a theologian, has researched similarities between different aspects of technological singularity and the world's religions (Geraci 2006). In particular, in his work on apocalyptic AI (Geraci 2008), he observes the many commonalities in the works of biblical prophets like Isaiah and the prophets of the upcoming technological singularity, such as Ray Kurzweil or Hans Moravec. All promise freedom from disease, immortality, and purely spiritual (software) existence in the kingdom come (Virtual Reality). More interestingly, Geraci (2007) argues that to be accepted into the society as equals, robots must convince most people that they are conscious beings. Geraci believes that an important component for such attribution is voluntary religious belief. Just like some people choose to believe in a certain religion, so will some robots. In fact, one may argue that religious values may serve the goal of limiting the behavior of superintelligences to those acceptable to society just like they do for many people. Here is how Geraci motivates his argument (Geraci 2007): "If robots become conscious, they may desire entrance into our society. … If no robots can enter into our religious lives, then I suspect we will deny them all equal and near-equal status in our culture. … To qualify as 'persons,' … some of them need to be religious—and by choice, not deliberate programming."

Adherents of Eastern religions are even more robot friendly and in general assume that robots will be happy to serve society and pose no danger. For example, Japan's Fumio Hara thinks that if "you are a good, kind person to the robot, the robot will become kind in return" (Menzel and D'Aluisio 2001, 76). Another eminent Japanese scientist, Shigeo Hirose, believes that robots "can be saints-intelligent and unselfish" (Menzel and D'Aluisio 2001, 89). Overall, convincing robots to worship humans as gods may be a valid alternative to friendly and humane AI systems.

6.2.3.3 Education

David Brin, in a work of fiction, has proposed that smart machines should be given humanoid bodies and from inception raised as our children

and taught the same way we were (Brin 1987). Instead of programming machines explicitly to follow a certain set of rules, they should be given the capacity to learn and should be immersed in human society with its ethical and cultural rules.

6.2.4 Self-Monitoring

6.2.4.1 Hard-Coded Rules

Probably the earliest and the best-known solution for the problem of intelligent machines was proposed by Isaac Asimov, a biochemist and a science fiction writer, in the early 1940s. The so-called Three Laws of Robotics are almost universally known and have inspired numerous imitations as well as heavy critique (Gordon-Spears 2003; McCauley 2007; Weld and Etzioni 1994; Pynadath and Tambe 2002). The original laws as given by Asimov are as follows (Asimov 1942):

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given to it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with either the First or Second Law.

Asimov later expanded the list to include a number of additional laws ("Three Laws" 2015):

1. *Zeroth Law:* A robot may not harm humanity or through inaction allow humanity to come to harm.
2. *Minus-One Law:* A robot may not harm sentience or through inaction allow sentience to come to harm.
3. *Fourth Law:* A robot must establish its identity as a robot in all cases.
4. Alternate *Fourth Law:* A robot must reproduce unless such reproduction would interfere with the First or Second or Third Law.
5. *Fifth Law:* A robot must know it is a robot.

Continuing Asimov's work, rule-based standards of behavior for robots have been recently proposed by South Korea's Ministry of Commerce,

Industry, and Energy. In 2007, a Robot Ethics Charter, which sets ethical guidelines concerning robot functions, has been adopted. In Europe, EURON (European Robotics Research Network) also announced plans to develop guidelines for robots in five areas: safety, security, privacy, traceability, and identifiability. Japan's Ministry of Economy, Trade, and Industry has issued policies regarding robots in homes and how they should behave and be treated (Sawyer 2007).

6.2.4.2 Chaining God

Stuart Armstrong proposed that trustworthiness of a superintelligent system could be monitored via a chain of progressively less-powerful AI systems all the way down to the human level of intelligence. The proposed "chain" would allow people to indirectly monitor and perhaps control the ultraintelligent machine. However, Armstrong himself acknowledges a number of limitations for the proposed method: The meaning of communication could be lost from one AI level to the next, or AI links in the chain may not be able to reliably judge the trustworthiness of a more intelligent entity. In such cases, the proposed solution is to shut down all AI and to start building the chain from scratch. Here is how Armstrong describes the chain: "If there were an entity, … just below the level of the GodAI …, [it] might be powerful enough, and smart enough, to conclude the GodAI was trustworthy. Then, assuming a level of AI intelligence just below the [entity] that could check up on them, the message of trust could be passed down, eventually reaching us" (Armstrong 2007, 3).

6.2.4.3 Friendly AI

To protect humankind against unintended consequences of superintelligent machines, Eliezer Yudkowsky, an AI researcher, has suggested that any AI system under development should be "friendly" to humanity (Yudkowsky 2008). Friendliness, according to Yudkowsky, could be defined as looking out for the best interests of humankind. To figure out what humankind is really interested in, the design of friendly AI (FAI) should be done by specialized AIs. Such seed AI (Yudkowsky 2001b) systems will first study human nature and then produce a friendly superintelligence humanity would want if it was given sufficient time and intelligence to arrive at a satisfactory design, our coherent extrapolated volition (CEV) (Yudkowsky 2004). Yudkowsky is not the only researcher working on the

problem of extracting and understanding human desires. Tim Freeman has also attempted to formalize a system capable of such "wish mining" but in the context of "compassionate" and "respectful" plan development by AI systems (Freeman 2009).

For friendly self-improving AI systems, a desire to pass friendliness as a main value to the next generation of intelligent machines should be a fundamental drive. Yudkowsky also emphasizes the importance of the "first mover advantage": The first superintelligent AI system will be powerful enough to prevent any other AI systems from emerging, which might protect humanity from harmful AIs. Here is how Yudkowsky himself explains FAI and CEV:

> The term "Friendly AI" refers to the production of human-benefiting, non-human-harming actions in Artificial Intelligence systems that have advanced to the point of making real-world plans in pursuit of goals. (Yudkowsky 2001a, 2)
>
> … Our coherent extrapolated volition is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted. (Yudkowsky 2004, 6)

6.2.4.4 Humane AI

Ben Goertzel, a frequent critic of FAI (Goertzel 2006), has proposed a variation on the theme he calls a humane AI. He believes it is more feasible to install AI with general properties such as compassion, choice, and growth than with specific properties like friendliness to humans (Goertzel 2006). In Goertzel's own words (Goertzel 2004b): "In Humane AI, one posits as a goal, … the development of AI's that display the qualities of 'humaneness,' … as a kind of ethical principle, where the principle is: 'Accept an ethical system to the extent that is agrees with the body of patterns known as 'humaneness.'"

6.2.4.5 Emotions

Bill Hibbard believes that the design of superintelligent machines needs to incorporate emotions that can guide the process of learning and

self-improvement in such machines. In his opinion, machines should love us as their most fundamental emotion; consequently, they will attempt to make us happy and prosperous. He states: "So in place of laws constraining the behavior of intelligent machines, we need to give them emotions that can guide their learning of behaviors. They should want us to be happy and prosper, which is the emotion we call love" (Hibbard 2001, 12).

Others have also argued for the importance of emotions, for example, Mark Waser wrote: "Thinking machines need to have analogues to emotions like fear and outrage that create global biases towards certain actions and reflexes under appropriate circumstances" (Waser 2010b, 174).

6.2.5 Indirect Solutions

6.2.5.1 Why They May Need Us

Continuing with the economic model of supply and demand, it is possible to argue that the superintelligent machines will need humans and therefore not exterminate humanity (but still might treat it less than desirably). For example, in the movie *Matrix,* machines need the heat from our bodies as energy. It is not obvious from the movie why this would be an efficient source of energy, but we can certainly think of other examples.

Friendly AI is attempting to replicate what people would refer to as "common sense" in the domain of plan formation (Yudkowsky 2005). Because only humans know what it is like to be a human (Nagel 1974), the friendly machines would need people to provide that knowledge, to essentially answer the question: "What would a human do (WWHD)?"

Alan Turing, in "Intelligent Machinery, a Heretical Theory," argued that humans can do something machines cannot, namely, overcome limitations of Godel's incompleteness theorem (Turing 1996). Here is what Turing said on this matter: "By Godel's famous theorem, or some similar argument, one can show that however the machine is constructed there are bound to be cases where the machine fails to give an answer, but a mathematician would be able to" (Turing 1996, 256).

Another area of potential need for assistance from human beings for machines may be deduced from some peer-reviewed experiments showing that human consciousness can affect random number generators and other physical processes (Bancel and Nelson 2008). Perhaps ultraintelligent machines will want that type of control or some more advanced technology derivable from it.

As early as 1863, Samuel Butler argued that the machines will need us to help them reproduce:

> They cannot kill us and eat us as we do sheep; they will not only require our services in the parturition of their young (which branch of their economy will remain always in our hands), but also in feeding them, in setting them right when they are sick, and burying their dead or working up their corpses into new machines. … The fact is that our interests are inseparable from theirs, and theirs from ours. Each race is dependent upon the other for innumerable benefits, and, until the reproductive organs of the machines have been developed in a manner which we are hardly yet able to conceive, they are entirely dependent upon man for even the continuance of their species. It is true that these organs may be ultimately developed, inasmuch as man's interest lies in that direction; there is nothing which our infatuated race would desire more than to see a fertile union between two steam engines; it is true that machinery is even at this present time employed in begetting machinery, in becoming the parent of machines often after its own kind, but the days of flirtation, courtship, and matrimony appear to be very remote, and indeed can hardly be realized by our feeble and imperfect imagination. (Butler 1863, 184)

A set of anthropomorphic arguments is also often made. They usually go something like the following: By analyzing human behavior, we can see some reasons for a particular type of intelligent agent not to exterminate a less-intelligent life form. For example, humankind does not need elephants, and we are smarter and certainly capable of wiping them out, but instead we spend lots of money and energy preserving them. Why? Is there something inherently valuable in all life-forms? Perhaps their DNA is a great source of knowledge that we may later use to develop novel medical treatments? Or, maybe their minds could teach us something? Maybe the fundamental rule implanted in all intelligent agents should be that information should never be destroyed. As each living being is certainly packed with unique information, this would serve as a great guiding principle in all decision making. Similar arguments could be made about the need of superintelligent machines to have cute human pets, a desire for companionship with other intelligent species, or a milliard other human needs. For example, Mark Waser, a proponent of teaching the machines

universal ethics (Waser 2008), which only exist in the context of society, suggested that we should "convince our super-intelligent AIs that it is in their own self-interest to join ours."

6.2.5.2 Let Them Kill Us

Some scientists are willing to give up on humanity all together in the name of a greater good they claim ultraintelligent machines will bring (Dietrich 2007). They see machines as the natural next step in evolution and believe that humanity has no right to stand in the way of progress. Essentially, their position is to let the machines do what they want, they are the future, and lack of humanity is not necessarily a bad thing. They may see the desire to keep humanity alive as nothing but a self-centered bias of *Homo sapiens*. Some may even give reasons why humanity is undesirable to nature, such as the environmental impact on Earth and later maybe the cosmos at large. According to some of the proponents of the "let-them-kill-us" philosophy: "Humans should not stand in the way of a higher form of evolution. These machines are godlike. It is human destiny to create them," believes Hugo de Garis (1999).

6.2.5.3 War Against the Machines

Amazingly, as early as 1863, Samuel Butler wrote about the need for a violent struggle against machine oppression:

> Day by day, however, the machines are gaining ground upon us; day by day we are becoming more subservient to them; … the time will come when the machines will hold the real supremacy over the world and its inhabitants is what no person of a truly philosophic mind can for a moment question. Our opinion is that war to the death should be instantly proclaimed against them. Every machine of every sort should be destroyed by the well-wisher of his species. Let there be no exceptions made, no quarter shown; let us at once go back to the primeval condition of the race. If it be urged that this is impossible under the present condition of human affairs, this at once proves that the mischief is already done, that our servitude has commenced in good earnest, that we have raised a race of beings whom it is beyond our power to destroy, and that we are not only enslaved but are absolutely acquiescent in our bondage. (Butler 1863, 185)

6.2.5.4 If You Cannot Beat Them, Join Them

An alternative vision for the postsingularity future of humanity could be summarized as: "If you cannot beat them, join them." A number of prominent scientists have suggested pathways for humanity to be able to keep up with superintelligent machines by becoming partially or completely merged with our engineered progeny. Ray Kurzweil is an advocate of a process known as uploading, in which the mind of a person is scanned and copied into a computer (Kurzweil 2005). The specific pathway to such scanning is not important, but suggested approaches include advanced brain-computer interfaces (BCIs), brain scanning, and nanobots. A copied human could either reside in a robotic body or in virtual reality. In any case, superior computational resources in terms of processing speed and memory become available to such an uploaded human, making it feasible for the person to keep up with superintelligent machines.

A slightly less-extreme approach is proposed by Kevin Warwick, who also agrees that we will merge with our machines but via direct integration of our bodies with them. Devices such as brain implants will give "cyborgs" computational resources necessary to compete with the best of the machines. Novel sensors will provide sensual experiences beyond the five we are used to operating with. A human being with direct uplink to the wireless Internet will be able to instantaneously download necessary information or communicate with other cyborgs (Warwick 2003). Both Kurzweil and Warwick attempt to analyze potential consequences of humanity joining the machines and come up with numerous fascinating predictions. The one aspect they agree on is that humanity will never be the same. Peter Turney suggests an interesting twist on the "fusion" scenario: "One approach to controlling a SIM would be to link it directly to a human brain. If the link is strong enough, there is no issue of control. The brain and the computer are one entity; therefore, it makes no sense to ask who is controlling whom" (Turney 1991, 3).

6.2.5.5 Other Approaches

I have reviewed some of the most prominent and frequently suggested approaches for dealing with the SP, but many other approaches and philosophical viewpoints are theoretically possible (Sotala and Yampolskiy 2015). Many of them would fall into the singularity "denialist" camp, accepting the following statement by Jeff Hawkins ("Tech Luminaries" 2008): "There will be no singularity or point in time where the technology itself runs away from us." He further elaborates: "Exponential growth

requires the exponential consumption of resources (matter, energy, and time), and there are always limits to this. Why should we think intelligent machines would be different? We will build machines that are more 'intelligent' than humans and this might happen quickly, but there will be no singularity, no runaway growth in intelligence." A recent report from the Association for the Advancement of Artificial Intelligence (AAAI) presidential panel on long-term AI futures outlines similar beliefs held by the majority of the participating AI scientists: "There was overall skepticism about the prospect of an intelligence explosion as well as of a 'coming singularity,' and also about the large-scale loss of control of intelligent systems" (Horvitz and Selman 2009).

Others may believe that we might get lucky and even if we do nothing, the superintelligence will turn out to be friendly to us and possess some human characteristics. Perhaps this will happen as a side effect of being (directly or indirectly) designed by human engineers, who will, maybe subconsciously, incorporate such values into their designs or, as Douglas Hofstadter put it ("Tech Luminaries" 2008): "Perhaps these machines—our 'children'—will be vaguely like us and will have culture similar to ours." Yet others think that superintelligent machines will be neutral toward us. John Casti thinks that ("Tech Luminaries" 2008) "machines will become increasingly uninterested in human affairs just as we are uninterested in the affairs of ants or bees. But it's more likely than not in my view that the two species will comfortably and more or less peacefully coexist." Both Peter Turney (1991) and Alan Turing (1950) suggested that giving machines an ability to feel pleasure and pain will allow us to control them to a certain degree and will assist in machine learning. Unfortunately, teaching machines to feel pain is not an easy problem to solve (Bishop 2009; Dennett 1978).

Finally, one can simply deny that the problem exists by questioning either the possibility of the technological singularity or not accepting that it leads to the SP. Perhaps one can believe that a superintelligent machine by its very definition will have at least as much common sense as an average human and will consequently act accordingly.

6.3 ANALYSIS OF SOLUTIONS

Table 6.1 provides a summary of the methods described in this chapter proposed to either directly or indirectly address the problem we have named the SP. I have categorized the proposed solutions into five broad categories: prevention of development, restricted deployment,

TABLE 6.1 Summary of the Potential Solution Methods

Category	Methodology	Investigated by	Year
Prevention of development	Fight scientists	Ted Kaczynski	1995
	Outlaw research	Bill Joy	2000
	Restrict hardware	Anthony Berglas	2009
	Singularity steward	Ben Goertzel	2004
Restricted deployment	AI-Boxing	Eric Drexler, Eliezer Yudkowsky	2002
	Leakproofing	David Chalmers	2010
	Oracle AI	Nick Bostrom	2008
	AI-Confinement	Roman V. Yampolskiy	2011
Incorporation into society	Economic	Robin Hanson	2008
	Legal	H. Moravec, R. Hanson, S. Omohundro	2007
	Religious	Robert Geraci	2007
	Ethical/social	Mark Waser, Joshua Fox, Carl Shulman	2008
	Moral	J. Storrs Hall	2000
	Equality	Bill Hibbard	2005
	Education	David Brin	1987
Self-monitoring	Rules to follow	Isaac Asimov	1942
	Friendly AI	Eliezer Yudkowsky	2001
	Emotions	Bill Hibbard	2001
	Chaining	Stuart Armstrong	2007
	Humane AI	Ben Goertzel	2004
	Compassionate AI	Tim Freeman	2009
Other solutions	They will need us	Alan Turing	1950
	War against machines	Samuel Butler	1863
	Join them	Ray Kurzweil, Kevin Warwick	2003
	Denialism	Jeff Hawkins	2008
	Do nothing	Douglas Hofstadter, John Casti	2008
	Pleasure and pain	Peter Turney	1991
	Let them kill us	Hugo de Garis, Eric Dietrich	2005
	Fusion of humans and AI	Peter Turney	1991
	Reproductive control	Samuel Butler	1863

incorporation into society, self-monitoring, and indirect solutions. Such grouping makes it easier both to understand the proposed methods and to analyze them as a set of complete measures. I review each category and analyze it in terms of feasibility of accomplishing the proposed actions and, more important, for evaluating the likelihood of the method succeeding if implemented.

The violent struggle against scientific establishment, outlawing AI research, and placing restrictions on development and sale of hardware components are all part of an effort to prevent superintelligent machines from ever coming into existence and to some extent are associated with the modern Luddite movement. Given the current political climate, complex legal system, and economic needs of the world's most developed countries, it is highly unlikely that laws will be passed to ban computer scientists either from researching AI systems or from developing and selling faster processors. Because for this methodology to work the ban needs to be both global and enforceable, it will not work as there is no global government to enforce such a law or to pass it in the first place. Even if such a law were passed, there is always a possibility that some rogue scientist somewhere will simply violate the restrictions, making them at best a short-term solution.

An idea for an automated monitoring system (also known as "big brother AI") is as likely to be accepted by humanity as the legal solution analyzed previously. It also presents the additional challenge of technological implementation, which as far as I can tell would be as hard to make "humanity safe" as a full-blown singularity-level AI system. Provided that the system would have to be given legal rights to control people, Martha Moody said: "Sometimes the cure is worse than the disease." Finally, as for the idea of violent struggle, it may come to be, as suggested by Hugo de Garis (2005), but I will certainly not advocate such an approach or will even consider it as a real solution.

Restricting access of superintelligent machines to the real world is a commonly proposed solution to the SP problem. AI-Boxes, leakproofing, and restricted question-answering-only systems (known as oracle AIs) are just some of the proposed methods for accomplishing that. Although much skepticism has been expressed toward the possibility of long-term restriction of a superintelligent mind, no one so far has proven that it is impossible with mathematical certainty. This approach may be similar to putting a dangerous human being in prison. Although some have escaped from maximum security facilities, in general, prisons do provide a certain

measure of security that, even though not perfect, is still beneficial for improving the overall safety of society. This approach may provide some short-term relief, especially in the early stages of the development of truly intelligent machines. I also feel that this area is one of the most likely to be accepted by the general scientific community as research in the related fields of computer and network security, steganography detection, computer viruses, encryption, and cyber warfare is well funded and highly publishable. Although without a doubt the restriction methodology will be extremely difficult to implement, it might serve as a tool for at least providing humanity with a little more time to prepare a better response.

Numerous suggestions for regulating the behavior of machines by incorporating them into human society have been proposed. Economic theories, legal recourse, human education, ethical principles of morality and equality, and even religious indoctrination have been suggested as ways to make superintelligent machines a part of our civilization. It seems that the proposed methods are a result of an anthropomorphic bias because it is not obvious why machines with minds drastically different from humans, no legal status, no financial responsibilities, no moral compass, and no spiritual desires would be interested in any of the typical human endeavors of daily life. We could, of course, try to program into the superintelligent machines such tendencies as metarules, but then we simply change our approach to the so-called self-monitoring methods I discuss further elsewhere. Although the ideas proposed in this category are straightforward to implement, I am skeptical of their usefulness because any even slightly intelligent machine will discover all the loopholes in our legal, economic, and ethical systems as well as or better than humans can. With respect to the idea of raising machines as our children and giving them a human education, this would be impractical not only because of the required time but also because we all know about children who greatly disappoint their parents.

The self-monitoring category groups together dissimilar approaches, such as explicitly hard-coding rules of behavior into the machine, creating numerous levels of machines with increasing capacity to monitor each other, or providing machines with a fundamental and unmodifiable desire to be nice to humanity. The idea of providing explicit rules for robots to follow is the oldest approach surveyed in this chapter and as such has received the most criticism over the years. The general consensus seems to be that no set of rules can ever capture every possible situation, and that the interaction of rules may lead to unforeseen circumstances and

undetectable loopholes, leading to devastating consequences for humanity. The quotations that follow exemplify such criticism: "The real problem with laws is that they are inevitably ambiguous. … Trying to constrain behavior by a set of laws is equivalent to trying to build intelligence by a set of rules in an expert system. … I am concerned by the vision of a superintelligent lawyer looking for loopholes in the laws governing its behavior" (Hibbard 2001, 12). "However, it is not a good idea simply to put specific instructions into their basic programming that force them to treat us as a special case. They are, after all, smarter than we are. Any loopholes, any reinterpretation possible, any reprogramming necessary, and special-case instructions are gone with the snows of yesteryear" (Hall 2000).

The approach of chaining multiple levels of AI systems with progressively greater capacity seems to be replacing a difficult problem of solving SP with a much harder problem of solving a multisystem version of the same problem. Numerous issues with the chain could arise, such as a break in the chain of communication or an inability of a system to accurately assess the mind of another (especially smarter) system. Also, the process of constructing the chain is not trivial.

Finally, the approach of making a fundamentally friendly system that will desire to preserve its friendliness under numerous self-improvement measures seems to be likely to work if implemented correctly. Unfortunately, no one knows how to create a human-friendly, self-improving optimization process, and some have argued that it is impossible (Legg 2006; Goertzel 2002, 2004a). It is also unlikely that creating a friendly intelligent machine is easier than creating any intelligent machine, creation of which would still produce an SP. Similar criticism could be applied to many variations on the FAI theme (e.g., Goertzel's humane AI or Freeman's compassionate AI). As one of the more popular solutions to the SP problem, the friendliness approach has received a significant dose of criticisms (Goertzel 2006; Hibbard 2003, 2005b); however, I believe that this area of research is well suited for scientific investigation and further research by the mainstream AI community. Some work has already begun in the general area of ensuring the behavior of intelligent agents (Gordon-Spears 2004; Gordon 1998).

To summarize my analysis of self-monitoring methods, I can say that explicit rules are easy to implement but are unlikely to serve the intended purpose. The chaining approach is too complex to implement or verify and has not been proven to be workable in practice. Finally, the approach of installing fundamental desire into the superintelligent machines to

treat humanity nicely may work if implemented, but as of today, no one can accurately evaluate the feasibility of such an implementation.

Finally, the category of indirect approaches comprises nine highly diverse methods, some of which are a bit extreme and others that provide no solution. For example, Peter Turney's idea of giving machines the ability to feel pleasure and pain does not in any way prevent machines from causing humanity great amounts of the latter and in fact may help machines to become torture experts given their personal experiences with pain.

The next approach is based on the idea first presented by Samuel Butler and later championed by Alan Turing and others; in this approach, the machines will need us for some purpose, such as procreation, so they will treat us nicely. This is highly speculative, and it requires us to prove existence of some property of human beings for which superintelligent machines will not be able to create a simulator (reproduction is definitely not such a property for software agents). This is highly unlikely, and even if there is such a property, it does not guarantee nice treatment of humanity as just one of us may be sufficient to perform the duty, or maybe even a dead human will be as useful in supplying the necessary degree of humanness.

An extreme view is presented (at least in the role of devil's advocate) by Hugo de Garis, who says that the superintelligent machines are better than humans and so deserve to take over even if it means the end of the human race. Although it is certainly a valid philosophical position, it is neither a solution to the SP nor a desirable outcome in the eyes of the majority of people. Likewise, Butler's idea of an outright war against superintelligent machines is likely to bring humanity to extinction due to the share difference in capabilities between the two types of minds.

Another nonsolution is discussed by Jeff Hawkins, who simply states that the technological singularity will not happen; consequently, SP will not be a problem. Others admit that the singularity may take place but think that we may get lucky and the machines will be nice to us just by chance. Neither of these positions offers much in terms of solution, and the chances of us getting lucky given the space of all possible nonhuman minds is close to zero.

Finally, a number of hybrid approaches are suggested that say that instead of trying to control or defeat the superintelligent machines, we should join them. Either via brain implants or via uploads, we could become just as smart and powerful as machines, defeating the SP problem

by supplying our common sense to the machines. In my opinion, the presented solution is both feasible (in particular, the cyborg option) to implement and likely to work; unfortunately we may have a Pyrrhic victory as in the process of defending humanity we might lose ours. Last but not least, we have to keep in mind a possibility that the SP simply has no solution and prepare to face the unpredictable postsingularity world.

6.4 FUTURE RESEARCH DIRECTIONS

With the survival of humanity on the line, the issues raised by the problem of the SP are too important to put "all our eggs in one basket." We should not limit our response to any one technique or an idea from any one scientist or a group of scientists. A large research effort from the scientific community is needed to solve this issue of global importance. Even if there is a relatively small chance that a particular method would succeed in preventing an existential catastrophe, it should be explored as long as it is not likely to create significant additional dangers to the human race.

After analyzing dozens of potential solutions from as many scientists I came to the conclusion that the search is just beginning. Perhaps because the winning strategy has not yet been suggested or maybe additional research is needed to accept an existing solution with some degree of confidence. I would like to offer some broad suggestions for the future directions of research aimed at counteracting the problem of the SP.

First, research needs to shift from the hands of theoreticians and philosophers into the hands of practicing computer scientists. Limited AI systems need to be developed as a way to experiment with nonanthropomorphic minds and to improve current security protocols. Fusion approaches based on the combination of the most promising solutions reviewed in this chapter should be developed and meticulously analyzed. In particular, goal preservation under self-improvement needs to be investigated and its feasibility addressed. Finally, a global educational campaign needs to take place to teach the general public about the nature of the SP and to help establish a political movement, which is likely to bring funding and the necessary laws to allow for a better response to the threats resulting from the technological singularity.

6.5 CONCLUSIONS

The issues raised in this chapter have been exclusively in the domain of science fiction writers and philosophers for decades. Perhaps through such means or maybe because of advocacy by organizations such as the

Singularity Institute for Artificial Intelligence/Machine Intelligence Research Institute (SIAI/MIRI) (*Reducing Long-Term Catastrophic Risks* 2011), the topic of superintelligent AI has slowly started to appear in mainstream publications such as this book. I am glad to report that some preliminary work has begun to appear in scientific venues that aims to specifically address issues of AI safety and ethics, if only in human-level intelligence systems. The prestigious scientific magazine *Science* has published on the topic of roboethics (Sharkey 2008; Sawyer 2007), and numerous papers on machine ethics (Anderson and Anderson 2007; Lin, Abney, and Bekey 2011; Moor 2006; Tonkens 2009) and cyborg ethics (Warwick 2003) have been published in recent years in other prestigious journals.

I am hopeful that the publication of this book will do for the field of AI safety engineering research what gravitational singularity did for the universe: provide a starting point. For a long time, work related to the issues raised in this book has been informally made public via online forums, blogs, and personal websites by a few devoted enthusiasts. I believe the time has come for AI safety research to join mainstream science. It could be a field in its own right, supported by strong interdisciplinary underpinnings and attracting top mathematicians, philosophers, engineers, psychologists, computer scientists, and academics from other fields.

With increased acceptance will come the possibility to publish in many mainstream academic venues; I call on fellow researchers to start specialized peer-reviewed journals and conferences devoted to AI safety research. With the availability of publication venues, scientists will take over from philosophers and will develop practical algorithms and begin performing actual experiments related to the AI safety engineering. This would further solidify AI safety research as a mainstream scientific topic of interest and will produce some long-awaited answers. In the meantime, it is best to assume that superintelligent AI may present serious risks to humanity's very existence and to proceed or not proceed accordingly. In the words of Bill Joy (2000): "Whether we are to succeed or fail, to survive or fall victim to these technologies, is not yet decided. I'm up late again—it's almost 6 a.m. I'm trying to imagine some better answers."

REFERENCES

Allen, Colin, Wendell Wallach, and Iva Smit. July/August 2006. Why machine ethics? *IEEE Intelligent Systems* 21(4):12–17.

Anderson, Michael and Susan Leigh Anderson. 2007. Machine ethics: creating an ethical intelligent agent. *AI Magazine* 28(4):15–26.

Armstrong, Stuart. 2007. Chaining God: A Qualitative Approach to AI, Trust and Moral Systems. New European Century. http://www.neweuropeancentury.org/GodAI.pdf

Asimov, Isaac. March 1942. Runaround. *Astounding Science Fiction*.

Bancel, Peter and Roger Nelson. 2008. The GCP event experiment: design, analytical methods, results. *Journal of Scientific Exploration* 22(4):309–333.

Benford, G. 1988. *Me/Days. In Alien Flesh*. London: Gollancz.

Berglas, Anthony. February 22, 2009. Artificial Intelligence Will Kill Our Grandchildren. http://berglas.org/Articles/AIKillGrandchildren/AIKillGrandchildren.html

Bishop, Mark. 2009. Why computers can't feel pain. *Minds and Machines* 19(4):507–516.

Bostrom, Nick. 2006. Ethical issues in advanced artificial intelligence. *Review of Contemporary Philosophy* 68(5):66–73.

Bostrom, Nick. 2008. Oracle AI. http://lesswrong.com/lw/qv/the_rhythm_of_disagreement/

Bostrom, Nick and Eliezer Yudkowsky. 2011. The ethics of artificial intelligence. In *Cambridge Handbook of Artificial Intelligence,* edited by William Ramsey and Keith Frankish. Cambridge, UK: Cambridge University Press. http://www.nickbostrom.com/ethics/artificial-intelligence.pdf

Brin, David. 1987. Lungfish. http://www.davidbrin.com/lungfish1.htm

Bugaj, Stephan and Ben Goertzel. 2007. Five ethical imperatives and their implications for human-AGI interaction. *Dynamical Psychology.* http://goertzel.org/dynapsyc/2007/Five_Ethical_Imperatives_svbedit.htm

Butler, Samuel. June 13, 1863. Darwin among the machines [To the editor of *The Press*]. *The Press*, Christchurch, New Zealand.

Chalmers, David. 2010. The singularity: a philosophical analysis. *Journal of Consciousness Studies* 17:7–65.

de Garis, Hugo. 2005. *The Artilect War.* Palm Springs, CA: ETC.

Dennett, Daniel C. July 1978. Why you can't make a computer that feels pain. *Synthese* 38(3):415–456.

Dietrich, Eric. 2007. After the humans are gone. *Journal of Experimental and Theoretical Artificial Intelligence* 19(1):55–67.

Drexler, Eric. 1986. *Engines of Creation*. Norwell, MA: Anchor Press.

Epstein, Richard Gary. 1997. Computer Psychologists Command Big Bucks. http://www.cs.wcupa.edu/~epstein/comppsy.htm

Fox, Joshua and Carl Shulman. October 4–6, 2010. Superintelligence Does Not Imply Benevolence. Paper presented at the Eighth European Conference on Computing and Philosophy, Munich, Germany.

Freeman, Tim. 2009. Using Compassion and Respect to Motivate an Artificial Intelligence. http://www.fungible.com/respect/paper.html

Gavrilova, Marina and Roman Yampolskiy. October 20–22, 2010. Applying Biometric Principles to Avatar Recognition. Paper presented at the International Conference on Cyberworlds (CW2010), Singapore.

Geraci, Robert M. 2006. Spiritual robots: religion and our scientific view of the natural world. *Theology and Science* 4(3):229–246.

Geraci, Robert M. June 14, 2007. Religion for the robots. In *Sightings*. Chicago: Martin Marty Center at the University of Chicago. http://divinity.uchicago.edu/martycenter/publications/sightings/archive_2007/0614.shtml

Geraci, Robert M. 2008. Apocalyptic AI: religion and the promise of artificial intelligence. *Journal of the American Academy of Religion* 76(1):138–166.

Gibson, W. 1984. *Neuromancer*. New York: Ace Science Fiction.

Goertzel, Ben. 2002. Thoughts on AI morality. *Dynamical Psychology*. http://www.goertzel.org/dynapsyc

Goertzel, Ben. 2004a. The all-seeing (A)I. *Dynamic Psychology*. http://www.goertzel.org/dynapsyc

Goertzel, Ben. 2004b. Encouraging a positive transcension. *Dynamical Psychology*. http://www.goertzel.org/dynapsyc/2004/PositiveTranscension.htm

Goertzel, Ben. September 2006. Apparent Limitations on the "AI Friendliness" and Related Concepts Imposed by the Complexity of the World. http://www.goertzel.org/papers/LimitationsOnFriendliness.pdf

Good, Irving John. 1966. Speculations concerning the first ultraintelligent machine. *Advances in Computers* 6:31–88.

Gordon, Diana F. 1998. Well-Behaved Borgs, Bolos, and Berserkers. Paper presented at the 15th International Conference on Machine Learning (ICML98), San Francisco.

Gordon-Spears, Diana. 2004. Assuring the behavior of adaptive agents. In *Agent Technology from a Formal Perspective*, edited by Christopher A. Rouff et al., 227–257. Dordrecht, the Netherlands: Kluwer.

Gordon-Spears, Diana F. 2003. Asimov's laws: current progress. *Lecture Notes in Computer Science* 2699:257–259.

Hall, J. Storrs. 2000. Ethics for Machines. http://autogeny.org/ethics.html

Hanson, Robin. June 2008. Economics of the singularity. *IEEE Spectrum* 45(6):45–50.

Hanson, Robin. October 10, 2009. Prefer Law to Values. http://www.overcoming-bias.com/2009/10/prefer-law-to-values.html

Hawking, Stephen. March 6, 1998. Science in the next millennium. Presentation at The Second Millennium Evening at the White House. Washington, DC.

Hibbard, Bill. 2001. Super-intelligent machines. *Computer Graphics* 35(1):11–13.

Hibbard, Bill. 2003. Critique of the SIAI Guidelines on Friendly AI. http://www.ssec.wisc.edu/~billh/g/SIAI_critique.html

Hibbard, Bill. July 2005a. The Ethics and Politics of Super-Intelligent Machines. http://www.ssec.wisc.edu/~billh/g/SI_ethics_politics.doc

Hibbard, Bill. December 2005b. Critique of the SIAI Collective Volition Theory. http://www.ssec.wisc.edu/~billh/g/SIAI_CV_critique.html

Horvitz, Eric and Bart Selman. August 2009. Interim Report from the AAAI Presidential Panel on Long-Term AI Futures. http://www.aaai.org/Organization/Panel/panel-note.pdf

Hugo de Garis. 1999. http://en.wikipedia.org/wiki/Hugo_de_Garis

Joy, Bill. April 2000. Why the future doesn't need us. *Wired Magazine* 8(4). http://archive.wired.com/wired/archive/8.04/joy.html

Kaczynski, Theodore. September 19, 1995. Industrial society and its future. *New York Times*.

Kaczynski, Theodore. *The Unabomber Manifesto: Industrial Society and Its Future*. Filiquarian Publishing, LLC.

Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology*. New York: Viking Press.

Legg, Shane. 2006. Friendly AI is bunk. In *Vetta Project*. http://commonsenseatheism.com/wp-content/uploads/2011/02/Legg-Friendly-AI-is-bunk.pdf

Lin, Patrick, Keith Abney, and George Bekey. 2011. Robot ethics: mapping the issues for a mechanized world. *Artificial Intelligence* 175(5–6):942–949.

McCauley, Lee. 2007. AI Armageddon and the three laws of robotics. *Ethics and Information Technology* 9(2):153–164.

Menzel, Peter and Faith D'Aluisio. 2001. *Robo Sapiens Evolution of a New Species*. Cambridge, MA: MIT Press.

Moor, James H. July/August 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems* 21(4):18–21.

Nagel, Thomas. 1974. What is it like to be a bat? *Philosophical Review* 83(4):435–450.

Omohundro, Stephen M. September 8–9, 2007. The Nature of Self-Improving Artificial Intelligence. Paper presented at the *Singularity Summit*, San Francisco.

Omohundro, Stephen M. February 2008. The Basic AI Drives. In *Proceedings of the First AGI Conference, Volume 171, Frontiers in Artificial Intelligence and Applications,* edited by P. Wang, B. Goertzel, and S. Franklin, 483–492. Amsterdam: IOS Press.

Pynadath, D. V., and Milind Tambe. 2002. Revisiting Asimov's first law: a response to the call to arms. *Intelligent Agents VIII*, *Lecture Notes in Computer Science* 2333:307–320.

Reducing Long-Term Catastrophic Risks from Artificial Intelligence. 2011. San Francisco: Singularity Institute for Artificial Intelligence. http://singinst.org/riskintro/index.html

Sawyer, Robert J. November 16, 2007. Robot ethics. *Science* 318:1037.

Sharkey, Noel. December 19, 2008. The ethical frontiers of robotics. *Science* 322:1800–1801.

Shulman, Carl, Henrik Jonsson, and Nick Tarleton. October 1–2, 2009. Machine Ethics and Superintelligence. Paper presented at the Fifth Asia-Pacific Computing and Philosophy Conference, Tokyo.

Shulman, Carl, Nick Tarleton, and Henrik Jonsson. October 1–2, 2009. Which Consequentialism? Machine Ethics and Moral Divergence. Paper presented at the Asia-Pacific Conference on Computing and Philosophy (APCAP'09), Tokyo.

Solomonoff, Ray J. 1985. The time scale of artificial intelligence: reflections on social effects. *North-Holland Human Systems Management* 5:149–153.

Sotala, Kaj. 2009. Evolved Altruism, Ethical Complexity, Anthropomorphic Trust: Three Factors Misleading Estimates of the Safety of Artificial General Intelligence. Paper presented at the Seventh European Conference on Computing and Philosophy (ECAP'09), Barcelona, Spain, July 2–4, 2009.

Sotala, Kaj and Roman V. Yampolskiy. January 2015. Responses to catastrophic AGI risk: a survey. *Physica Scripta* no. 90:018001.

Tech Luminaries Address Singularity. June 2008. *IEEE Spectrum. Special Report: The Singularity*. http://spectrum.ieee.org/computing/hardware/tech-luminaries-address-singularity

Three Laws of Robotics. 2015. Last modified January 14. http://en.wikipedia.org/wiki/Three_Laws_of_Robotics

Tonkens, Ryan. 2009. A challenge for machine ethics. *Minds and Machines* 19(3):421–438.

Turing, A. 1950. Computing machinery and intelligence. *Mind* 59(236):433–460.

Turing, A. M. 1996. Intelligent machinery, a heretical theory. *Philosophia Mathematica* 4(3):256–260.

Turney, Peter. 1991. Controlling super-intelligent machines. *Canadian Artificial Intelligence* 27:3, 4, 12, 35.

Vinge, Vernor. March 30–31, 1993. The Coming Technological Singularity: How to Survive in the Post-human Era. Paper presented at Vision 21: Interdisciplinary Science and Engineering in the Era of Cyberspace, Cleveland, OH. https://www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html

Warwick, Kevin. 2003. Cyborg morals, cyborg values, cyborg ethics. *Ethics and Information Technology* 5:131–137.

Waser, Mark. October 4–6, 2010a. Deriving a Safe Ethical Architecture for Intelligent Machines. Paper presented at the Eighth Conference on Computing and Philosophy (ECAP'10), Münich, Germany.

Waser, Mark R. 2008. *Discovering the Foundations of a Universal System of Ethics as a Road to Safe Artificial Intelligence*. AAAI Technical Report FS-08-04. Menlo Park, CA: AAAI.

Waser, Mark R. March 5–8, 2010b. Designing a Safe Motivational System for Intelligent Machines. Paper presented at the Third Conference on Artificial General Intelligence, Lugano, Switzerland.

Weld, Daniel S., and Oren Etzioni. July 31–August 4, 1994. The First Law of Robotics (a Call to Arms). Paper presented at the Twelfth National Conference on Artificial Intelligence (AAAI), Seattle, WA.

Yampolskiy, Roman V., and Venu Govindaraju. November 20–22, 2007. Behavioral Biometrics for Recognition and Verification of Game Bots. Paper presented at the Eighth Annual European Game-On Conference on simulation and AI in Computer Games (GAMEON'2007), Bologna, Italy.

Yampolskiy, Roman V., and Venu Govindaraju. March 16–20, 2008. Behavioral Biometrics for Verification and Recognition of Malicious Software Agents. Paper presented at Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense VII. SPIE Defense and Security Symposium, Orlando, FL.

Yampolskiy, Roman V. April 13, 2007. Behavioral Biometrics for Verification and Recognition of AI Programs. Paper presented at the 20th Annual Computer Science and Engineering Graduate Conference (GradConf2007), Buffalo, NY.

Yampolskiy, R.V., 2012. Leakproofing the singularity: artificial intelligence confinement problem. *Journal of Consciousness Studies* 19(2):194–214.

Yudkowsky, Eliezer. 2005. What Is Friendly AI? http://singinst.org/ourresearch/publications/what-is-friendly-ai.html

Yudkowsky, Eliezer. 2008. Artificial intelligence as a positive and negative factor in global risk. In *Global Catastrophic Risks*, edited by N. Bostrom and M. M. Cirkovic, 308–345. Oxford, UK: Oxford University Press.

Yudkowsky, Eliezer S. 2001a. Creating Friendly AI—The Analysis and Design of Benevolent Goal Architectures. http://singinst.org/upload/CFAI.html

Yudkowsky, Eliezer S. 2001b. General Intelligence and Seed AI—Creating Complete Minds Capable of Open-Ended Self-Improvement. http://singinst.org/ourresearch/publications/GISAI/

Yudkowsky, Eliezer S. 2002. The AI-Box Experiment. http://yudkowsky.net/singularity/aibox

Yudkowsky, Eliezer S. May 2004. Coherent Extrapolated Volition. Singularity Institute for Artificial Intelligence. http://singinst.org/upload/CEV.html

Yudkowsky, Eliezer S. September 2007. Three Major Singularity Schools. *Singularity Institute Blog.* http://yudkowsky.net/singularity/schools

Index

B

D

E

F

G

H

M

Q

R

S

T

Y

Z

Artificial Intelligence

ARTIFICIAL SUPERINTELLIGENCE

A FUTURISTIC APPROACH

"...the hot topic that seems to have come straight from science fiction...vigorous academic analysis pursued by the author produced an awesome textbook that should attract everyone's attention: from high school to graduate school students to professionals."

—**Leon Reznik**, Professor of Computer Science, Rochester Institute of Technology

A day does not go by without a news article reporting some amazing breakthrough in artificial intelligence (AI). Many philosophers, futurists, and AI researchers have conjectured that human-level AI will be developed in the next 20 to 200 years. If these predictions are correct, it raises new and sinister issues related to our future in the age of intelligent machines. **Artificial Superintelligence: A Futuristic Approach** directly addresses these issues and consolidates research aimed at making sure that emerging superintelligence is beneficial to humanity.

While specific predictions regarding the consequences of superintelligent AI vary from potential economic hardship to the complete extinction of humankind, many researchers agree that the issue is of utmost importance and needs to be seriously addressed. **Artificial Superintelligence: A Futuristic Approach** discusses key topics such as:

- AI-Completeness theory and how it can be used to see if an artificial intelligent agent has attained human-level intelligence
- Methods for safeguarding the invention of a superintelligent system that could theoretically be worth trillions of dollars
- Self-improving AI systems: definition, types, and limits
- The science of AI safety engineering, including machine ethics and robot rights
- Solutions for ensuring safe and secure confinement of superintelligent systems
- The future of superintelligence and why long-term prospects for humanity to remain as the dominant species on Earth are not great

Artificial Superintelligence: A Futuristic Approach is designed to become a foundational text for the new science of AI safety engineering. AI researchers and students, computer security researchers, futurists, and philosophers should find this an invaluable resource.

CRC Press
Taylor & Francis Group
an informa business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

K22996
ISBN: 978-1-4822-3443-5
90000
9 781482 234435
www.crcpress.com