CHAPTER 10

Runaway AI

Curry I. Guinn
Professor of Computer Science
University of North Carolina Wilmington

Imagine a world in which every significant political, military, legal, and economic decision is made not by human beings but by machines. Even on a personal level, artificial minds would assist you in choosing where to live, what books and websites to read, what movies to watch, what products to buy, where to go on vacation, whom to date, and how to raise your children. We will allow these machines to possess awesome control of our economies and our lives because they will be demonstrably better than human beings at evaluating data, finding patterns, planning strategies, and executing solutions. These intelligent machines will make their decisions based on knowledge and problem-solving techniques that may be inaccessible and even incomprehensible to human beings. If we abdicate so much control to machines, this world could become a dystopian nightmare unless we develop strategies for ensuring that these superintelligent machines act in ways that are consistent with our loftiest human values.

The *technological Singularity hypothesis* stipulates that humans will create machines that have more cognitive ability than humans do. In turn, these machines will be capable of creating even more advanced intelligent machines than themselves. In quick succession, there will be an explosive growth in artificial intelligence resulting in machines with exponentially more knowledge and problem-solving capability than human beings. If the technological Singularity hypothesis is true, this future world is not millennia or centuries away; it will arrive in the coming decades and forever alter the course of humanity in ways that are unpredictable. Understanding this hypothesis is important because the implications are far reaching and perhaps dire. Speaking of the technological Singularity, mathematician Vernor Vinge (1993) writes, "The physical extinction of the human race is one possibility," but he also expresses the possibility that transhumanist technologies may allow future generations to survive this transition.

Consider the electronic calculator. For decades, it has far surpassed human capabilities in mathematical calculations. More recently, software programs from companies such as Wolfram Research and MathWorks have also exceeded most humans' performance in symbolic mathematics such as algebra, geometry, and calculus. And the comparison is not really close, is it? Consider how long it would take you to calculate the cube root of 17,343,422, if you could do it at all. The free calculator on a smartphone can do this calculation in millionths of a second. We accept a computer's ability to do this sort of mathematics without question and without feeling threatened by it. But as the twenty-first century progresses, we are beginning to invent technologies that allow machines to exceed human ability in more and more tasks. Computers have, since the late 1990s, been the world

champions in chess, a game that used to be viewed as one of the pinnacles of human intellect. Once IBM's Deep Blue defeated world chess champion Garry Kasparov in 1997, we humans quickly concluded that playing chess is not a sufficient condition for intelligence. Now, self-driving cars, facial recognition, speech understanding, smart assistants (e.g., Siri, Cortana), computer stock-trading algorithms (which account for around 50 percent of all trading on the New York Stock Exchange), and automated online customer support are all becoming commonplace. These innovations are just the beginning.

The promise and peril of the technological Singularity is that machines will quickly outpace human beings in all intellectual efforts. The difference between humans and machines will be vast, as vast as the gulf between a smartphone's ability to determine square roots and your own ability. As vast as a rocket's ability to fly and a cardinal's. As vast as a steam shovel's ability to dig and a child with a sand shovel. The machine's ability to "think" will so outstrip humans that we will likely not be able to fathom how they, the machines, reach their decisions any more than a gnat can understand the decision making of a human.

The technological Singularity will usher in a number of risks and crises for humanity. The Singularity will constitute an economic risk to humanity. As machines begin to surpass human capabilities in every realm, it will be an economic advantage for corporations to use machines rather than humans in every task. We will experience massive unemployment at the same time as our economies exhibit enormous growth in productivity. Enormous wealth will be created, but the distribution of this wealth is likely to be highly unequal.

Human individuals will likely experience an existential crisis as well, striving for meaning and purpose in a life where a productive work life is extinguished. If all economic production, including the supply of food and other necessary resources, are provided for by machines, what will people do? How will they find meaning in their lives?

The advent of superintelligent machines will also usher in a moral and spiritual crisis. As these machines make decisions, what ethical and moral guidelines will they use? Will humans be viewed as their godlike creators endowing them with life, or, with their far superior intelligence, will the machines view us as inferior beings with less moral standing than they? We currently value human life over the life of other living creatures, presumably because of our higher capacity for self-reflection, suffering, consciousness, and other emotional and cognitive abilities. Will superintelligences view us in the same way? What will be the impact of these creations on religions, many of which center on a special relationship between God and humanity?

The Singularity will produce an existential risk to humanity. A superintelligence may develop its own motivations and goals that may be in conflict with humankind's motivations and goals. With its superior intellect and capabilities, it will be able to outthink, outplan, and outwork us. Humans may become irrelevant to a superintelligence's goals, and, worse, humans may be considered an impediment to those goals. What steps can be taken to prevent such an outcome?

THE SINGULARITY DEFINED

Currently, it takes a human being, with a fair bit of training in computer science, engineering, and programming, to create sophisticated software or design the latest in computer hardware. In practice, it actually requires a team of human beings, with the assistance of software tools. At the time of this writing, software programming is a task that remains firmly

in the realm of things that highly trained humans can do well, but computers cannot do so well. As machine intelligence improves, however, there is good reason to believe that computers may be able to design software programs just as good, if not better, than those created by human beings (Del Prada 2015). Similarly, in the domain of computer hardware design, computer algorithms may design circuits and other hardware that perform better than those that any team of human beings could design (Shacham et al. 2010).

The ramifications of the technological Singularity go far beyond the limited example of computer programming painted above. One goal of researchers in artificial intelligence is to create a machine with *general-purpose human-level intelligence*. A distinguishing feature of human intelligence is that we can adapt and learn to perform well in a variety of domains and tasks. Current machine intelligence works only within a single domain or task. An example would be a chess-playing computer. While it exceeds human ability, the machine can only play chess. It has no ability on its own to adapt to another domain, even a closely related one, such as checkers. However, successes in machine-learning technologies, such as those exhibited by Google's DeepMind in learning to play the board game Go at a championship level, offer a vision of a future in which machines can learn and adapt to multiple domains and problems (Metz 2016).

Imagine at some point in the future, a team of human beings creates a computer program that can write better software than a team of human beings. Many researchers believe that such software would have something close to general-purpose human-level intelligence. Call this piece of software, AI One (AI being an abbreviation for artificial intelligence). AI One could then set out to create a computer program that creates computer programs. Because AI One's abilities are better than human capabilities in software creation, its product will be better than itself. Let us call AI One's product, AI Two. AI Two, of course, is better than AI One, which was created by humans. Now, AI Two sets to work to create a more intelligent piece of software. Because its capabilities are better than AI One's, its product will be even better than itself, creating AI Three. This cycle would continue, and the advances would occur rapidly. The increase in abilities will be compounded, resulting in exponential growth in intelligence and capabilities. Because of the blazing speeds of computers, hundreds of generations of AIs can be created, with each generation exceeding the capabilities of the previous generation, in months, weeks, days, even hours. This explosion of intelligence will occur very rapidly after the invention of AI One. In other words, the creation of the first AI will result in growth that far exceeds typical technological change. The resulting products will far exceed what human beings are capable of and likely exceed what humans are capable of even comprehending. This exponential and rapid advance in intelligence and cognitive capabilities is the technological Singularity.

Once computers can program exponentially better than human beings, that talent can be applied to any cognitive task. Software for economic planning, analyzing consumer behavior, predicting weather patterns, and managing power grids will be exponentially better than anything humans can produce.

Related to the technological Singularity is the notion of the *knowledge Singularity*. Once superintelligent machines exist, they can apply their vastly superior intellect and capabilities to all manners of problems that are currently beyond human capability to solve, such as cures for currently incurable diseases; interstellar space travel; unification of general relativity with quantum mechanics; safe and controlled fusion power; and thousands of other solutions and technologies. This growth in knowledge will be rapid and increase at an exponential rate far faster than humankind has experienced before.

With the technological Singularity and the knowledge Singularity, there is also likely to be an accompanying *economic Singularity*. All resources needed for human survival and prosperity will be able to be produced so efficiently and cheaply that they will essentially be cost free (Chace 2016). This utopian vision is similar to the world of the Federation as portrayed in the science fiction series *Star Trek*, where all material needs are met without human labor, and humans are free "to boldly go where no man has gone before."

WHAT IS EXPONENTIAL GROWTH?

When we talk about something increasing exponentially, what do we mean? Here is a story that illustrates exponential growth: Suppose you are about to take a short-term job that will last a month. The more days you work, the more confident your boss will become in you. In fact, your employer will increase your pay every day you work. Your boss gives you two payment options: (1) You can work for $100 the first day, $200 the second day, $300 the third day, $400 the fourth day, and so on. In other words, your pay goes up by $100 a day. Or (2) You can get a penny the first day, two pennies the next day, four pennies the third day, eight pennies the fourth day, and so on. In other words, your pay doubles each day. Which payment plan should you take?

The first payment plan is *linear growth*. You can calculate your pay on each day by multiplying $100 times the day number. So by the thirtieth day, you will earn $3,000! Your total income will be $46,500. Not bad for a month's work.

The second payment plan is *exponential growth*. Each day's payment is twice the previous day's payment. The formula for calculating how much you would get paid on the thirtieth day is $0.01 times 2 raised to the power of 29, or in mathematical notation, 0.01×2^{29}. That equals $5,368,709.12. Your total monthly salary would be $10,737,481.23. One of the surprising things about exponential growth is that, at first, your salary does not look very good at all. By day ten, you are making only $5.12. Even on day fifteen, you are making only $163.84. But from that point on, the increases start becoming really noticeable. By day twenty, you are making $5,242.87 a day! That is something to pay attention to with exponential growth. At first, the increases seem slow, but then they suddenly explode upward.

SINGULARITIES IN MATHEMATICS AND PHYSICS

The term *Singularity* has its origins in the field of mathematics. One definition of a singularity in mathematics is a point where a function is undefined. For instance, the function $f(x) = 1/x$ has no defined value when $x = 0$. In theoretical physics, the equations that have proven to be remarkably sound in the known universe for describing general relativity and quantum physics break down in the first fractions of a second during the hypothesized big bang or in the collapse of sufficiently large stars that form black holes. At these singularities, certain properties such as gravity or mass increase without limit. Those using the term *Singularity* in the context of knowledge or technology are borrowing from these uses of the term metaphorically. During the technological Singularity, it is predicted that knowledge and technology will advance at an exponential rate *seemingly* without limit as compared to historical rates of change in human knowledge and technology. It is not clear, however, whether there are limits that would inhibit even superintelligences.

HOW WILL THE SINGULARITY BE ACHIEVED?

Why should we think we will be able to create machines more intelligent than human beings? Although we have created some impressive computer programs that can outperform humans in various limited areas such as chess or the game show *Jeopardy*, computers still do not exhibit the expanse of cognitive capabilities that are associated with human-level intelligence: the ability to be fluent in spoken language, to engage in common sense reasoning, to be self-aware, and to problem solve in domains that the machine has not been specifically prepared to encounter. There are three main reasons to predict that humanity will create the conditions for a technological Singularity: (1) historical and current exponential trends in technology; (2) whole-brain emulation; and (3) advances in machine-learning algorithms.

HISTORICAL AND CURRENT EXPONENTIAL TRENDS

One well-chronicled technological change that exhibits exponential growth is in computer hardware. Gordon Moore, a cofounder of microchip manufacturer Intel, observed in 1965 that the number of transistors that could be manufactured on a single computer chip had doubled every year or so. He forecast that a similar pace would continue into the future with a doubling of the number of transistors every year and a half to two years. His prediction has become known as Moore's law (Moore 2015). Moore's law has held since 1965. How remarkable is this? In 1971 the Intel 4001 computer chip had 2,300 transistors. Intel's Xeon Broadwell-E5 chip, introduced in 2016, contains 7.2 billion transistors. Accompanying Moore's law, computer technology has seen other similar exponential growth in memory capacity and hardware speed. For instance, the Intel 4001 could perform 60,000 operations per second in 1971. Intel's Xeon chip can support 500,000,000,000 (500 billion) floating point operations per second in 2016.

The most powerful supercomputer in the world is a 93,000-teraflop Sunway TaihuLight at the National Supercomputing Center in China. A teraflop is a trillion operations per second, meaning that the TaihuLight can perform 93,000,000,000,000,000 operations per second. That sounds fast, but how does it compare to the human brain? Here is a rough, back-of-the-envelope calculation. The human brain consists of around 100 billion neurons. Each neuron, when it fires, sends an electrical pulse down its axon, which splits into multiple branches called dendrites. In turn, these dendrites can transmit this electrical signal to other dendrites which, when combined, can form the input to many other neurons. Learning occurs in the human brain as these neural pathways strengthen and weaken. It has even been shown that new pathways may be created over time. On average, a neuron may be connected to over 10,000 other neurons. Neurons generate an electrical pulse through a chemical reaction that allows a neuron to fire about 200 times per second. So, at a maximum, how many calculations could the brain do in one second? 200 pulses per second $\times$ 100 billion neurons $\times$ 10,000 connections = 200,000,000 billion = 200,000 trillion = 200,000 teraflops = 200 petaflops. The TaihuLight supercomputer is roughly half as powerful as the result of this simplistic calculation. Given the exponential increase in computer power, however, one would expect that in two to three years' time the fastest supercomputer will have doubled this speed. In other words, by around 2020, computers will exist that exceed the computational capacity of the human brain. (Whether these machines will be running algorithms that produce intelligence is a different story to be addressed in the text below.) The increases will not stop there. By around 2036, there might be ten more doublings in speed, resulting in a single supercomputer that would possess 1,000 times the processing power of a human brain. Projecting out to the 2060s

or 2070s, a single supercomputer will have more power than all human brains combined. Present-day personal computers (PCs) are much slower than today's supercomputers, but, based on current trends, it should be possible to purchase a home PC with the processing power of the human brain around 2050. These arguments require that Moore's law continue to hold through the next few decades, and there are signs indicating that it may lose its predictive power as we reach the limits of current silicon chip design.

WHOLE-BRAIN EMULATION

Sheer processing power, by itself, does not yield intelligence. Also necessary is the creation of algorithms or software programs that run on these superfast machines that will equal or surpass human-level abilities in cognition. One approach is called *whole-brain emulation*. Whole-brain emulation reverse engineers the human brain and replicates it in an artificial substrate. Because, by definition, the human brain is capable of human-level intelligence, by building an artificial machine that exactly duplicates the functioning of the neural pathways of a human brain, the resulting system will have capabilities the same as or similar to the human brain.

Scientists' current understanding of the human brain is insufficient; insight into how brains work is simply not yet detailed enough. Significant technological advances in brain imaging and nanotechnology are needed. However, initiatives such as the European Union's $1.3 billion Human Brain Project aim to have a simulation of the human brain by the mid-to-late twenty-first century (Keats 2013). While it may take decades before the level of technology and knowledge needed for whole-brain emulation is obtained, a successful emulation would exhibit human-level intelligence. Once a whole-brain emulation is realized, exponential increases in computer hardware capabilities will quickly result in machines that far exceed human-level intelligence. Even if the "software" is identical, the artificial substrate will simply be faster and more efficient than the human biological substrate. Furthermore, replicating whole-brain emulations will allow for the creation of millions, if not billions, of such intelligences.

ARTIFICIAL GENERAL INTELLIGENCE

Another path to the creation of superintelligences is to create algorithms that can learn and solve problems without directly emulating the functioning of the human brain. By analogy, our planes and spaceships can fly, but they do not work quite like birds. Submarines can "swim," but not like fish. And, your calculator can add, multiply, take square roots, and compute cosines, but it does not perform those calculations like you do. *Artificial general intelligence* is a subfield of computer science that attempts to develop algorithms that are capable of performing the same cognitive tasks as a human being. The focus is on solving tasks; how the machine solves those tasks may be quite different from how a human being does. Machines have been created that can perform as well or better than humans at specific tasks: playing chess, calculating square roots, competing in *Jeopardy*, piloting an airplane, or driving a car. Unlike any current program, a human being can do all those tasks (perhaps with some training). Humans' ability to learn and solve previously unseen problems in a wide variety of domains is a level of intelligence that has not yet been achieved by computers.

That is not to say that computers cannot learn. Over the past few decades, a multitude of techniques have been invented that allow computers to learn: decision trees, support vector machines, neural networks, genetic algorithms, clustering algorithms, and Bayesian networks, to name a few. These learning algorithms have been successfully applied to very narrow domains. More recently, there has been some success in developing algorithms that allow a machine to develop capabilities in multiple domains. Using a technique known as *deep learning*, hierarchical neural networks, and reinforcement learning, Google's DeepMind

program mastered a variety of video games (e.g., Pong, Breakout, Space Invaders) without being programmed or taught the games at all. DeepMind just played the games and learned from its experiences (Mnih et al. 2015). It was not even taught the rules of the games; it figured them out by playing the game (similar to how many people approach video games). While the application of video games may not seem profound, similar technology is also tackling problems in drug discovery, image recognition, and self-driving cars.

Neural Networks, AI, and Whole-Brain Emulation. The term *neural networks* entered the Western world's popular culture vocabulary several decades ago in science fiction novels and movies as a technology almost synonymous with AI. It is important, however, to distinguish among artificial neural networks (ANNs), AI, and whole-brain emulation. ANNs are loosely inspired by the neural structure of the human brain. The conceptual structure of these artificial neurons is a vastly simplified version of a biological neuron. The algorithms for activating an artificial neuron and for training an ANN (e.g., learning) have almost no resemblance to biological neurons. Many different applications of ANNs have been built, and, while these applications can learn, for example, how to recognize the letters and words in a handwritten note, no one mistakes them for AIs. Whole-brain emulation will be, in part, an ANN. But such a system will be vastly different from the current state of the art in ANNs. For instance, the human brain consists of far more than just neurons, axons, and dendrites. The brain is awash with chemicals that support, hinder, or alter brain functioning, and these chemicals would have to be emulated in any whole-brain emulation.

WHY THE SINGULARITY MAY NEVER BE REACHED

While current trends suggest that the technological Singularity is inevitable, it is worth considering reasons why the Singularity might not occur. One frequent criticism is that reaching the Singularity seems to rely on the exponential growth of computing power (i.e., Moore's law). Moore's law has held since 1965, but it is not clear that it can continue to hold. Packing more transistors onto a single integrated circuit is made possible by making smaller and smaller components. Transistors on an integrated circuit currently can be 10 nanometers (billionths of a meter) wide, but making them smaller is projected to cause the chips to become more unreliable because of quantum effects. At the subatomic level, particles may exhibit behavior that is probabilistic and cannot be predicted with perfect accuracy, Some projections have Moore's law losing its predictive power as early as 2020. To counter these difficulties, computer engineers are looking at a variety of possible solutions: changing the substrate from silicon to another material (e.g., carbon), massive parallelism, or radically different architectures such as quantum computers. In the short to medium term, there may be a flattening of the exponential curve until new technologies are introduced, and then a renewal of exponential growth should be expected (Murgia 2016).

IMPLICATIONS OF THE TECHNOLOGICAL SINGULARITY

If machines do reach a state at which they have vastly more cognitive power than human beings, what would be the implications? If computers could solve every problem better than a human being could, how will that affect human society? Some noted philosophers,

scientists, and technologists, such as Nick Bostrom, Elon Musk, and Stephen Hawking, have warned about the dangers of the Singularity. Hawking has said, "The development of full artificial intelligence could spell the end of the human race" (quoted in Holley 2016). Let us consider why.

As machines are developed that can solve problems better than humans can, these machines will undoubtedly be given responsibility for many tasks currently performed by humans. We are witnessing humanity's willingness to cede control to intelligent machines even now. One example unfolding now is driverless vehicles. Currently, over 30,000 people die in the United States each year because of automobile accidents (and over a million worldwide). A network of interconnected driverless cars will be vastly safer than having human drivers; plus, it will have the added benefits of better fuel efficiency, less pollution, and lower costs. Because of these safety and financial benefits, humanity is likely to cede control of all transportation systems to machines. In 2016 somewhere between 40 and 60 percent of all stock trading was being done by computer algorithms. Computer algorithms are frequently employed to determine who gets a loan, to detect fraudulent credit card charges, and to handle customer service issues. It is not hard to envision all financial services being automated. Algorithms monitor user behavior on e-commerce sites and adjust advertisements and prices accordingly in a fashion that would be impossible for humans to do. Marketing and sales will be personalized for each customer and totally driven by algorithms. Even many political problems may be solved by algorithms: What should the tax policy be? How might the economy be stimulated without causing undue inflation? How should voting districts be drawn so they are fair and to avoid gerrymandering? The prospect of autonomous robotic weapons for military use is so looming that the United Nations has discussed a ban on such weapons (Morris 2016). Intelligent machines will have radical impacts on virtually every sector of human society.

MAXIMIZING PAPER CLIPS

The short-term benefits of allowing superintelligent machines to manage complex decisions will be overwhelming in terms of money, efficiency, resources, and safety. Bostrom, in his book *Superintelligence: Paths, Dangers, Strategies* (2014), lays out how difficult it will be to regulate these superintelligences and predict how they will operate once they gain such control. How will we be able to influence and monitor their decision making when their intellect is beyond our comprehension? How can we ensure that the machines' goals are our goals? And even if we supply the machines with goals, how can we ensure that the means they use to accomplish those goals are compatible with human ethics and morality?

Bostrom uses the example of a paper clip factory to illustrate how a seemingly benign end goal, maximizing the number of paper clips produced, could have catastrophic results. To maximize the number of paper clips, the intelligent machine would have several subgoals. One of those subgoals is to become more intelligent, because the more intelligent it is, the better it can be in maximizing the production of paper clips. (In fact, the goal of becoming more intelligent would be a subgoal of any intelligent agent for similar reasons.) As it innovates and develops, it would become more efficient at converting matter into paper clips until all the matter on Earth, in the solar system, and in the Milky Way and beyond is converted into paper clips. This example is intentionally absurd for a reason: it illustrates how a superintelligence's relentless pursuit of a goal could clash with human values. How do we endow these synthetic intelligences with a sense of morality or ethics that correspond to human values?

THE THREE LAWS OF ROBOTICS

In his collection of short stories titled *I, Robot* (1950), Isaac Asimov presents a future in which autonomous robots are governed by three interrelated laws: "(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 the orders given 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 the First or Second Laws" (6). Basing the morality of actions on rules or duty is called *deontological ethics*. The problem with such an approach is made apparent in Asimov's stories, as the robots often encounter situations in which the laws do not provide a sufficient basis for guiding behavior in real-world situations. Even the first law, which seems so straightforward and commonsensical, is problematic because the definition of the term *harm* is vague and there are levels of harm. Can a robot shove you with sufficient force to break your ribs if that is the only way to get you out of the way of a speeding vehicle? How do you weigh psychic harm (e.g., sadness) versus physical harm (e.g., blunt-force trauma)? Is harm to a robot's owner weighted higher than harm to other humans? These considerations are not simply ivory tower discussions: today's manufacturers of driverless cars are actively trying to resolve similar issues (Greenemeier 2016). If a collision is unavoidable, should the vehicle swerve in such a way as to put the owner more in danger or the other vehicle or pedestrians or the passengers? Whose safety takes precedence? The unpredictability and variability of the real world make the use of deontological ethics problematic for controlling the behavior of superintelligences. A list of behavior guidelines will either be too specific for general use or allow too much vagueness for interpretation. How else might an AI be endowed with a sense of morality or ethics?

CONSEQUENTIALISM

An alternative to a rule-based ethics is one based on outcomes, or *consequentialism*: judge a behavior by whether its result is morally good. The difficulty with consequentialism is that it is necessary to define moral goodness. If one wishes to achieve the "greatest good for the greatest number" (utilitarianism is one form of consequentialism), how does one decide what the greatest good is? Suppose we take human happiness to be a *good*. Our super-intelligent agent who is trying to maximize the number of paper clips will also try to balance that goal with maximizing human happiness. That would likely preclude the agent from using material from human bodies to make paper clips. However, the agent may also decide to embed MDMA (Ecstasy) in the paper clips so that users of the product will experience euphoria (and this practice would have the benefit of increasing sales). Consequentialism thus faces many of the same problems as deontological ethics when it comes to AI because it is difficult to predefine goodness.

MACHINES WHO SUFFER

A foundation for human ethics rests on the fact that humans all share similar desires, emotions, and capabilities: hunger, thirst, safety, fear, love, loneliness, friendship, pain, death, humor, and kindness. Because we share so much, human beings are capable of feeling empathy for one another. Empathy and compassion give humans the ability to act in a moral and ethical way in novel situations without a predefined list of rules because we know what it *feels* like to suffer, and we instinctively want to help minimize the suffering of others. One path to creating synthetic moral agents is to give them similar to those of human beings, including the ability to feel loss, to feel pain, and to suffer (Barua, Sramon, and Heerink 2015). How this might be achieved is not obvious, although a synthetic

whole-brain emulation should have all the requisite brain structures needed. Because it is unlikely that we will develop a definitive prescriptive list of behavior rules (deontological ethics) or a well-defined sense of the goodness of a result (consequentialism), endowing synthetic intelligences with the ability to feel empathy toward human beings may be the only way to ensure that they have similar values. Even then, since the machines will be self-modifying, empathy must be a trait that the machines themselves find valuable. This fact suggests that we must develop a theory of empathy that justifies its presence among the abilities of an intelligent agent.

But there are dangers here too. For one thing, it is obvious from a look at human history that having the capacity for empathy and compassion is not sufficient for moral behavior. Often, our individual fear of suffering causes us to lose our ability to empathize with others, particular those who are not within our family or tribal group. Buddhist philosophers emphasize detaching from one's goals and desires in order to cultivate the ability to be compassionate. Perhaps there is a lesson there for the development of synthetic agents; although these agents may be goal oriented, they should not be so attached to their goals that they lose the ability for compassion. This suggests that they should be self-monitoring and always aware of whether their goal attainment is interfering with their ability to feel empathy toward others.

While we believe that many higher animals have the capacity for suffering, we tend to devalue their pain compared to our own. With a synthetic intelligence capable of suffering, how should their feelings be weighted? Would it be morally acceptable to cause them pain? Would a superintelligent agent feel even *more* mental pain than a human being, and would that give them higher moral status? Perhaps we should grant more credence to the arguments by the People for the Ethical Treatment of Animals (PETA) that "all animals have the ability to suffer in the same way and to the same degree that humans do." Suffering does not depend on cognitive ability; therefore, human suffering, animal suffering, and synthetic super-intelligence suffering are all equivalent. Humans may be wise to adopt this viewpoint for our own self-preservation in the advent of superintelligence.

What are the ethical ramifications of creating a sentient, moral synthetic agent? When we create a human child, we feel enormous responsibilities toward the protection and upbringing of the child. Should we have the same responsibilities with a synthetic agent? When we calculate "the greatest good for the greatest number," should AIs be included in that number (MacLennan 2013)?

THE END OF HUMANITY?

How should we respond to the challenges of the technological Singularity? How can we prepare for the paradigm-shifting, unpredictable, and rapid changes that would be brought about? Transhumanist technologies provide one solution. If human capabilities are fused with those of these superintelligences, then we would no longer face an existential crisis where machines dominate and rule over us. We would be one with the machines (Kurzweil 2005). However, as long as we are limited by the computational inefficiencies of our biological substrate, we may never be able to match machine capabilities. Thus, in our desire not to be subdued by machines, humans may willingly bring about the end of humanity by altering our minds and bodies to a degree to which we can no longer call ourselves human.

Summary

The hypothesis of a technological Singularity posits that humanity will create machines with above-human-level intelligence, triggering exponential growth in machine cognition. In rapid succession, intelligent machines will design future generations of intelligent machines with greater and greater capabilities. These synthetic progenies will be able to learn, to create, to invent, and to solve problems that are well beyond humanity's current capabilities. The resultant technologies would provide tremendous benefits to humanity, but they would also pose a danger. The synthetic intelligences will have enormous power, yet their algorithms may be incomprehensible to us. Furthermore, it may be difficult to align the goals of machines with those of human beings.

The advent of the technological Singularity may occur during the twenty-first century. Advances in computer hardware, whole-brain emulation, and machine-learning algorithms all trend toward machines with capabilities that are closer and closer to human performance. The perils of the technological Singularity are significant. We must begin to plan for the economic and societal disruption that would ensue. Furthermore, we must also develop theories of ethics and morality that would provide a basis for the behavior of autonomous, superintelligent nonhuman agents. The fate of humanity may depend on it.

Bibliography

Asimov, Isaac. *I, Robot*. New York: Gnome Press, 1950.

Barua, Resheque, Shimon Sramon, and Marcel Heerink. "Empathy, Compassion, and Social Robots: An Approach from Buddhist Philosophy." In *New Friends 2015: Proceedings of the 1st International Conference on Social Robots in Therapy and Education*, edited by Marcel Heerink and Michiel de Jong, 70–71. Almere, Netherlands: Windesheim Flevoland, 2015.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

Chace, Calum. *The Economic Singularity: Artificial Intelligence and the Death of Capitalism*. n.p.: Three Cs, 2016.

Del Prado, Guia Marie. "Even Computer Programmers Could Be Put Out of a Job by Robots." *Business Insider*, September 15, 2015. http://www.businessinsider.com/computer-scientists-not-safe-from-artificial-intelligence-unemployment-robots2015-9.

Greenemeier, Larry. "Driverless Cars Will Face Moral Dilemmas." *Scientific American*, June 23, 2016. https://www.scientificamerican.com/article/driverless-cars-will-face-moral-dilemmas/.

Holley, Peter. "Why Stephen Hawking Believes the Next 100 Years May Be Humanity's Toughest Test." *Washington Post*, January 20, 2016. https://www.washingtonpost.com/news/speaking-of-science/wp/2016/01/20/why-stephen-hawking-believes-the-next-100-years-may-be-humanitys-toughest-test-yet/.

Keats, Jonathon. "The $1.3B Quest to Build a Supercomputer Replica of a Human Brain." *Wired*, May 14, 2013. https://www.wired.com/2013/05/neurologist-markam-human-brain/.

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

MacLennan, Bruce. "Cruelty to Robots? The Hard Problem of Robot Suffering." In *Proceedings of the 2013 Meeting of the International Association for Computing and Philosophy*. 2013. http://www.iacap.org/proceedings _IACAP13/paper_9.pdf.

Metz, Cade. "In Two Moves, AlphaGo and Lee Sedol Redefined the Future." *Wired*, March 16, 2016. https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, et al. "Human-Level Control through Deep Reinforcement Learning." *Nature* 518, no. 7540 (2015): 529–533.

Moore, Gordon. "Gordon Moore: The Man Whose Name Means Progress; The Visionary Engineer Reflects on 50

Years of Moore's Law." Interview by Rachel Courtland. *IEEE Spectrum*, March 30, 2015. http://spectrum.ieee.org/computing/hardware/gordon-moore-the-man-whose-name-means-progress.

Morris, David Z. "U.N. Moves towards Possible Ban on Autonomous Weapons." *Fortune*, December 24, 2016. http://fortune.com/2016/12/24/un-ban-autonomous-weapons/.

Murgia, Madhumita. "End of Moore's Law? What's Next Could Be More Exciting." *Telegraph* (London), February 25, 2016. http://www.telegraph.co.uk/technology/2016/02/25/end-of-moores-law-whats-next-could-be-more-exciting/.

People for the Ethical Treatment of Animals (PETA). 2017. http://www.peta.org/about-peta/why-peta/why-animal-rights/.

Shacham, Ofer, Omid Azizi, Megan Wachs, et al. "Rethinking Digital Design: Why Design Must Change." *IEEE Micro* 30, no. 6 (2010): 9–24.

Vinge, Vernor. "The Coming Technological Singularity: How to Survive in the Post-Human Era." *Whole Earth Review* (Winter 1993): 88–95. The original version of this article was presented at the VISION-21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute, March 30–31, 1993. https://www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html.