The Effect of Emerging Artificial Intelligence Techniques on the Ethical Role of Computer Scientists

A Thesis
in TCC 402

Presented to

The Faculty of the
School of Engineering and Applied Science
University of Virginia

In Partial Fulfillment
of the Requirements for the Degree

Bachelor of Science in Computer Science

by

Nicholas S Dunnuck

March 29, 2002

On my honor as a University student, on this assignment I have neither given nor received unauthorized aid as defined by the Honor Guidelines for Papers in TCC Courses.

_____

Approved: David Evans _____ (Technical Advisor)

Approved: Kathryn Neeley _____ (TCC Advisor)

Abstract

Emerging technologies and programming techniques increase our ability to create intelligent software programs. With the advent of viable neural networking solutions, we have come even closer to building artificially intelligent machines. This project outlines the impact of neural networking on the development of artificial intelligence (AI) systems, explores the impact of AI systems on society, and proposes enhanced ethical and professional roles for artificial intelligence developers, with an emphasis on interpersonal communication and impact awareness.

The projections discussed here are provided both by technology experts and concerned non-experts. Computer systems will continue to get more powerful, and will become increasingly ubiquitous in the future, making the standards of development of artificial intelligence a salient topic in modern engineering. Despite a socially ingrained fear of intelligent machines, there is no governing body to oversee the continued development of AI systems.

Development of a strong artificial intelligence would surely call into question (for some) that which we define as "alive." It is yet unclear whether an electronic entity would be entitled to legal and civil rights. Furthermore, we do not know whether such an entity or race of entities would be dangerous to society. These problems indicate a strong ethical component in the development of intelligent software. This paper argues that intelligent machines will be intertwined in our future society, and addresses the lack of a concrete body to govern the development of computer software. The accompanying research further establishes that engineers will have increased ethical and political responsibilities in the development of artificial intelligence systems in the future.

Preface

I knew from the start that this would be a somewhat unorthodox engineering thesis. I was so fascinated by my studies in philosophy here at the University that they consumed even my primary education in engineering. My sole initial goal in developing this project was to combine my philosophical undertakings with my engineering background. At times it seemed bleak, because I was never satisfied with the amount of philosophy and the amount of engineering going into this paper. It seems to have turned out, though, just right. The engineering background in my research fueled what was to become a grand exercise in philosophy. Naturally, I spent uncountable hours revising my topic and reworking my paper. But I was myself surprised to find that after hours, months, days of just sitting and thinking, my final conclusions became clear just days before the final revision of this paper was completed. It was most rewarding to suddenly realize that I thought engineers should be politicians, even if no one else would ever think the same.

Acknowledgements are in order for Dave Evans, my technical advisor, who was willing to help me turn an amorphous mass of maybes and what-ifs into a bona fide thesis project. Without his help I wonder if I could ever have nailed down the scope of this paper, and without his persistent questioning I might not have said a credible word in it. Additionally, I had tremendous help from my TCC advisor, Kay Neeley, whose spirited discussion and commentary sometimes made me think she was more interested in this project than I was. She also guided me through 6 credit hours of critical thinking about engineers in society, 6 credit hours that profoundly affected the shape and success of this project.

Finally, special recognition is in order for Jessie Kokrda, my best friend, who offered me advice and argument, support and solace. She has been the greatest single positive force sustaining my sanity. No part of this paper would be as it is without her. Her intelligence, grace, beauty, comedy, and even her naïveté have helped me find a better, smarter, more open and more loving person inside myself.

Glossary

AI – An abbreviation for Artificial Intelligence.

Artificial Intelligence – The design and study of computer programs that react flexibly and intelligently to a wide range of situations.

Chinese Room – Classic philosophical version of the Turing test. The Chinese Room example argues that intelligence cannot come from mechanical computation, and thus computers could never be intelligent.

Distributed Computing – Use of a network of distributed computers (on a network) to perform cooperative parallel processing tasks.

Luddite – A person who believes that technology, in and of itself, is bad.

Moore's Law – Long-standing observation made in 1965 by Gordon Moore (co-founder of Intel). The law states that the number of transistors per square inch on a given integrated circuit will double every eighteen months. Roughly equivalent to saying that integrated circuits will double in speed and halve in size every eighteen months.

Neural Network – A computerized simulation of mathematical models that represent and act like neurons in the mammalian nervous system (the brain).

Strong Artificial Intelligence – Artificial intelligence programming designed to act as a self-contained intelligence. A computer program capable of thinking for itself.

SuperComputer – An electronic computing machine capable of performing one billion or more operations per second.

Swarm – A collection of tiny independent computers that communicate via a wireless network. They have little power individually, but are designed to work cooperatively in parallel.

Tractability – The capability to turn theory into reality. The availability of resources that allow a theoretical solution to be physically manifested.

Weak Artificial Intelligence – Any of a number of programming techniques that allow deterministic computer programs to respond appropriately to a wide variety of situations.

Chapter One: Introduction

Computer scientists continue to gain influence in our society. Greater influence means that large corporations and government bodies are funding and supporting computer engineers for development of the world's newest technologies. Computers manage increasingly many aspects of our lives, and we still have not tapped their full potential. Still, there is no specific body, and few rules in place to assure that computer program technologies will be safe and beneficial to the general public. Artificial intelligence (AI) is now becoming a reality, and no one knows for sure what direction it will take. In light of new developments in intelligent programming technologies like neural networking, this paper will argue that truly intelligent machines may be in our future. More importantly, it will establish that computer scientists have considerable ethical and political responsibilities to the public.

What is Artificial Intelligence?

Artificial intelligence is the design and study of computer programs that behave intelligently [Dean 1]. It is in many ways the ultimate goal of computer programming. There is an ongoing effort to make more intelligent computer programs that are easier to use, even at the expense of simplicity and efficiency. Programs, after all, are designed to solve problems. That they should do so intelligently is a logical objective. This chapter will explain what it means for a computer program to behave intelligently and outline some uses for intelligent programs.

Recognizing Artificial Intelligence

It is difficult to define exactly what we mean when saying that a computer program should behave intelligently. Most people can give an abstract definition of intelligence, and anyone can look it up in a dictionary. However, conventional definitions of intelligence, like many commonly used expressions, are too ambiguous to be directly and usefully applied to computers. It is impossible to describe artificial intelligence, or to gauge our progress in that field, without knowing how intelligence applies to computers.

In a paper in 1950, Alan Turing proposed a test to measure the intelligence of computer programs [Turing 50]. Turing refers to this test as the 'imitation game' (though it has since been dubbed simply the Turing test). In the imitation game, a human judge uses a Teletype or some other simple interface to interrogate both a man (A) and a woman (B). The interrogator does not know in advance whether A is male and B is female, or vice versa. It is A's job to convince the interrogator that A is actually a woman. If asked, for example, the length of his hair, A might indicate that it is straight and layered, with the longest strands being several inches. It is B's job to help the interrogator figure out which interrogatee is male and which is female. B might type things like, "I am the woman! Trust me!" Such statements, however, would be of limited value, since A could easily type the same.

Roughly half of the time, the interrogator might be fooled into believing that A is actually the woman. Suppose, however, that A were a computer rather than a man. If that computer could win the imitation game, i.e. fool the human interrogator, with the

same frequency as a man, then the computer is said to have passed the Turing test. In terms of Turing's original paper, the computer might be judged capable of thinking.

While passing of the Turing test implies some definition of artificial intelligence, it is insufficient for describing modern AI systems. As computer science has begun to mature, we have developed new goals and uses for artificial intelligence, as well as new technologies for achieving those goals. Intelligent systems need not be designed to fool a human judge. Nor is such a facade necessarily desirable. A human working in a factory, for example, would require rest, supervision, and incentive to continue working. These are not characteristics we choose to emulate in computer programs. Yet there seems to be something intelligent about a robotic system that can, for example, build or design cars.

It is perhaps better to think of artificial intelligence as the study and design of computer programs that respond flexibly in unanticipated situations [Dean 1]. A computer program can give the illusion of intelligence if it is designed to react sensibly to a large number of likely and unlikely situations. This is similar to the way we might judge human intelligence, by a person's ability to solve problems and cope effectively with a wide variety of situations [Dean]. In this case, it is not necessary for an intelligent program (or person) to develop an original solution to a problem.

Still, to say that a computer program should react sensibly to situations is analogous to saying that it should react intelligently. In other words, the meaning of intelligence in terms of computers remains elusive. For the purposes of this paper, we will say that artificial intelligence is defined by two major methodologies and their purposes. Weak artificial intelligence is the design of computer programs with the

intention of adding functionality while decreasing user intervention. Many modern word processors are designed to indicate misspelled words without being asked to do so by the user. Some programs will even correct misspellings automatically. This is an example of weak artificial intelligence.

Strong artificial intelligence is the design of a computer program that may be considered a self-contained intelligence (or intelligent entity). The intelligence of these programs is defined more in terms of human thought. They are designed to think in the same way that people think. Passage of the Turing test, for example, might be one criterion for development of a strong AI system. The ethical issues in this paper deal largely with the strong AI methodology. However, the bulk of useful artificial intelligence applications lie in the realm of weak AI.

Applications of Artificial Intelligence

Artificial intelligence is useful in many domains, and its reach is constantly growing. The first step in artificial intelligence programming is automated reasoning. Automated reasoning is a computation that takes some encoded knowledge about the world as input and provides inferred conclusions based on that knowledge as output [Dean 12]. In the beginning, this automated reasoning programming was merely academic. Today, automated reasoning is used in video games, air traffic control systems, and the Mars rover.

Clearly, some of these programs require skills that we might normally associate with natural intelligence. Some things seem quite simple, like driving around a sandy planet surface. Nonetheless they are all useful, and they represent what we consider to be intelligent computer programs. There seem to be, however, many more things that we

want computers to do for us. Research continues as we stretch the limitations of computing devices, challenge each other to create greater intelligence systems, and struggle to define computer intelligence.

Limitations of Artificial Intelligence

Computer power continues to increase exponentially. This refers both to the speed of computing devices and the influence they have over our lives. Moore's Law predicts that computers will double in speed and halve in size every eighteen months [Moore 65]. While this law has held for over 35 years, current trends suggest that engineers will someday be limited by the sizes of molecules used in the construction of integrated circuits. Growth of computer power is linked growth of artificial intelligence systems. What, then, are the limitations of AI?

There are some computer programs that create original paintings from brush-stroking rules and stored images of objects. There are others that generate sensible haiku poetry from lists of related words (Kurzweil 163-9). These programs might be said to have passed a simplified and artistic variation of the Turing test. They demand some consideration when talking about artificial intelligence, but, given that they simply follow expansive sets of rules, how intelligent are they? There is some question, for example, as to whether a computer-generated poem can really be art. In these cases, and most cases of AI development, intelligent behavior boils down to a search over some set of possible actions and outcomes.

There have been lengthy debates over the limits of computer intelligence. Surely we can rely on the fact that computers will continue to get more powerful, and thus able

to perform more tasks in less time. Still, there may be things that a computer could never do. Specifically, it is unclear whether any product of the strong artificial intelligence methodology will ever succeed. That is, it is unclear whether a computer program will ever constitute a mind.

The Chinese Room

During the mid 1960's and through the 1970's, academic institutions and individuals put a great deal of effort into the research and development of strong artificial intelligence. One such project, developed at Yale University by Roger Schank and his colleagues, was designed to answer various questions about predetermined material [Searle 509-10]. (Today these programs are called expert systems.) Some argued that these projects were the beginnings of strong AI. In response to such claims, Philosopher John Searle wrote a classic proposition commonly called the "Chinese Room" example to refute this claim [Searle]. The example was not only meant to demonstrate that strong AI was not a reality, but that no Turing machine could produce a strong artificial intelligence.

In Searle's example, he is locked in a room with a stack of Chinese symbols. Searle speaks no Chinese, and could just as well be locked in a room with a stack of meaningless squiggles. A second stack of squiggles is introduced, along with instructions showing how to correlate the first stack of symbols with the second stack. The instructions are in English, which Searle understands perfectly, and they allow him to correlate the Chinese symbols entirely by their shapes. That is to say that the semantics of the symbols remain unknown to Searle. People outside the room are allowed to send more stacks of Chinese symbols under the door. When Searle receives these stacks, he

reads in his instructions how to correlate the old symbols with the new ones. He is then able to send back an appropriate stack of squiggles according to his instructions.

Searle argues that with a large enough set of instructions, he could fool anyone into thinking that he knew Chinese. Yet he clearly does not know Chinese. Similarly, computers that answer questions about predetermined material do not *understand* that material. They simply manipulate sets of formal symbols according to instructions in their native language. Searle's example is immensely more complex, however, in the sense that the Chinese room is designed to handle any reasonable domain of knowledge that a Chinese person might ask about.

The Chinese room is really only a philosophical version of the Turing test, passage of which is likely to be insufficient for defining strong artificial intelligence. Moreover, Searle points out that he could fool anyone into thinking that he was Chinese. This means that Searle must have an uncountable number of symbol correlations, because he must account for previous conversations and answer compounding inquiries appropriately. Regardless of whether Searle knows what he is doing, something about the Chinese room must understand Chinese. To say that Searle does not understand Chinese is analogous to saying that my mouth does not understand English. The understanding is contained in the set of instructions. While there may be no metaphysical understanding, it seems that if Searle and the room can correlate intelligible symbols about any subject, I might say that they (they as an entity) were intelligent. Moreover, the constant addition of new instructions is directly analogous to learning, another seemingly intelligent trait.

Searle's example does raise, however indirectly, the question of tractability. The fact that something is possible in theory does not make it a reasonable undertaking. To build a computer program like the Chinese room (in the same way that Searle describes it) one would require almost infinite memory and constantly increasing computational power. The Chinese Room example shows, in part, why strong AI is so elusive. Imagine the sheer size of the translator's book if he could handle all possible sensible combinations of Chinese symbols. The human brain is constantly bombarded with input from the five senses and somehow manages it all. We simply lack the technology and understanding to replicate that type of behavior at present. Moreover, the human mind may be more than just the sum of its parts. This idea, called dualism or the mind-body problem, is another roadblock to the understanding of computer-based intelligence.

Ethical Issues

On a philosophical level, there are important moral issues facing the developers of strong AI systems. Given that the goal is to develop an independently intelligent computer program, we should consider briefly how to classify such an entity. A strong artificial intelligence would surely call into question (for some) that which we define as "alive." It is yet unclear whether an intelligent electronic entity would be alive and legally entitled to certain rights.

There is no evidence that intelligent life, as it applies to human-like intelligence, is sustainable without a soul. Nor is there evidence that a soul is necessary. In fact, there is no complete definition of a soul at all. For some it is a vehicle by which we relate to a higher power, and for others it is nothing but nonsense. We must therefore consider

questions pertaining to life and intelligence notwithstanding the existence or nonexistence of a soul. In that case, it is impossible to say whether an entity inside a computer would be alive. However, there is more than enough uncertainty to say that such consideration must be given. There is no accounting for science, and it is impossible to tell exactly what questions future science will answer. In science, therefore, the case of moral justification must not be taken in terms of what will happen, but in terms of what might happen [Neeley]. An intelligent entity within a machine would likely have a justifiable claim to legal and possibly even civil rights, and pulling the plug on that machine may well constitute negligent or malicious killing.

With regard to the metaphysical problem of a soul, many people in the world believe that souls exist, and that all intelligent creatures have souls. In Kenneth Branagh's 1994 cinematic adaptation, *Mary Shelley's Frankenstein*, Frankenstein's fiend asks of his creator, "What of my soul? Do I have one?" A reasonably intelligent computer entity may be compelled to ask the same questions. An independently thinking entity certainly might have rights to those answers. How would the AI programmers respond to such inquiries? For some it is not simply a question of whether computer programs can have souls, but a question of who would be willing to take responsibility for those souls.

The remainder of this paper introduces arguments that strong artificial intelligence systems may be in our future. As the introduction of intelligent systems would almost certainly change our world in a dramatic way, the paper will then discuss several possible futures involving artificially intelligent machines. At its conclusion, the paper will show that computer scientists have the ultimate responsibility in making their products as safe

as possible. The lack of a strictly enforced regulatory standard on software development means that computer scientists must exercise independent self-governance when developing controversial and unpredictable technologies such as artificial intelligence networks. This responsibility, however, should not fall solely on programmers. The paper will bring to light the necessity for trained engineers to be more intimately involved in managerial and political positions.

Chapter Two: Neural Networking Technology

Neural networking is a technique that mimics biological intelligence in order to create artificially intelligent systems. Development of this technology has given the computer science community great leverage in helping machines to emulate human beings and intelligent animals. While the underlying concepts of neural networking have been around for some time, modern refinements to the technology and its application have spurred renewed interest in advancing the science. The combination of neural networks with modern equipment and techniques paves the way for machines that truly utilize strong artificial intelligence. This chapter will discuss what new technologies facilitate neural networks, and what neural networks mean for the development of artificial intelligence systems.

What is Neural Networking?

Neural networks are collections of mathematical models designed to work together to emulate the known properties of biological nervous systems [PNNL n.p.]. The mammalian brain contains billions of neurons. Thus, although the concept of neural networking has been around since the 1950s, only recently has the computing power become available to begin developing true, usable neural networks.

Animal brains, including the human brain, comprise massive parallel systems. That is to say that they process multiple pieces of information at one time. Biological brains are composed of neurons, which are the interconnected but independent workhorses of biological nervous systems. Each neuron may be connected with as many as a thousand or more other neurons. This allows mammals to quickly perform tasks like

recognizing patterns and faces. For many years, the parallel processing concept kept computers from effectively emulating these brain functions in the same way. Modern computer architecture provides several solutions to this problem.

The Modern Supercomputer

Moore's Law has been amazingly accurate during the lifecycle of the modern computer. Supercomputers are capable of performing one billion or more operations per second. These computers in particular were key to the development of early functional neural networks.

The basic electronic computational method, sequential computing, involves the processing of one piece of information at a time. This piece of information could be as small as one bit, which has a value of either one or zero. In a black and white picture, it takes two bits to represent one pixel, or a small dot in the picture. The complexity of such a picture, called the resolution, is directly related to, and defined by, the number of pixels representing the picture. A printed page, for example, will often have a resolution of 300 dots per inch. This means that one square inch of print could be defined by as many as 90,000 dots, or bits. A simple computer must go through at least 180,000 operations just to process one square inch of printed paper.

There have been many tricks employed to help computers deal with this kind of processing, such as reducing the resolution of pictures being analyzed. Critical information can often be preserved even when resolution is decreased. But this does not reach the root of the problem, namely that human beings are somehow capable of processing printed pages in their native format, and at resolutions even greater than 300 dots per inch. People can quickly scan through entire pages to search for a pattern. Often

this pattern will appear to jump off the page and grab the attention of the reader, without their having to read the entire text. This is the advantage of parallelism, and this is where computers have traditionally fallen short.

Information is processed in some computers in terms bytes or vectors, which are collections of bits. A vector may contain 100 to 1000 bits or even more. These are still, however, only small scalar multiples in terms of processing power. Even if a computer processed 100 bits simultaneously, it would still require 168,300 operations just to read in a standard 8 ½ x 11 inch piece of paper. Whereas people might skip right over white space, a computer must analyze every square inch in order to make sure that it is actually white space. Actual processing of the information could also take thousands or tens of thousands of operations per bit to analyze its relationship with neighboring bits. Since human beings and other animals see much more than an 8 ½ x 11 inch window, it becomes clear that supercomputers are necessary to emulate neural processing in a sequential environment.

Parallel Computing and Distributed Networks

It is much faster to work with multiple pieces of information if they can all be processed at the same time. This is called parallel processing. The beginnings of parallelism in computing are represented by the vector-based computing architecture discussed above. Parallel processing used to be infeasible for general use because hardware was at such a premium. It is time that is at a premium today, particularly human time. The addition of a processor in a computer may cost only a few hundred dollars, and may bring an increase in speed of 85%. Assuming no human cost to

parallelize the task, one 50-hour job turned into a 30-hour job recovers the cost of hardware.

Parallel computing can be much more than a pair of processors, however. The University of Virginia has several projects that are advancing the technology of parallel computing. Two projects in particular, Legion and Centurion, represent great advances in building working neural networks.

The Centurion project features 384 individual processors connected together and working in parallel. These processors could combine for up to 240 billion operations per second, making short work of processing a printed page [Centurion]. The even more ambitious Legion project is a software system that aims to connect millions of computers together to work in parallel. The infrastructure that could make this goal a reality is already built and being refined in the form of the Internet. With tens of millions of computers simultaneously simulating small neural systems, we could be very close to fully simulating the estimated 100 billion neurons that comprise the human brain. Current technology would require a few hundred million computers working in parallel to accomplish this task.

Implications of Neural Nets on AI

Certainly, neural networking enhances the possibility of developing intelligent systems. It is, after all, a direct emulation of what AI programmers are trying to achieve. In its own way, it solves the previously discussed problem of trying to figure out what it means for machines to be intelligent. If we consider ourselves intelligent, and directly emulate our own brains, then the product should likewise be intelligent. The pursuit of

intelligent software is neither unsavory nor unethical. Indeed, many perceived shortcomings in today's software come in part from our inability to program sufficiently intelligent programs.

Neural networks are already being used successfully in many commercial applications ranging from document processing to the food industry. Neural network systems are particularly good at pattern recognition, which has uses in odor analysis, handwriting recognition, credit analysis and many other tasks [PNNL]. Computers that are able to do these tasks are useful because, although people are very good at pattern recognition, we are not as good at the mundane tasks that follow. It is easy, for example, for a computer to track and analyze credit card use for thousands of people 24 hours a day. Computers can consistently analyze food odors and aromas in cases where human sensation may become numb, or in cases where the smell of bad food might make people sick.

Neural networks also bring us closer to developing strong artificial intelligence. By directly emulating mammalian brains, we should be getting closer to developing a program that has its own intelligence. If, as is commonly accepted, the entirety of human intelligence lies within the structure of the brain, it is possible that we need only simulate enough neurons to mimic that brain. In some ways, we are restricted by our limited knowledge of actual neural function, but we have substantial observational information regarding the function of individual neurons [Clabaugh]. With continued research, we may be able to develop an entire artificial brain.

The idea of neural networking again raises moral and philosophical complications. It reintroduces the idea of a living electronic entity in a way that is easier

to relate to as human beings. A neural network is a replica (albeit a small and grossly simplified replica, given current technology) of our own brain structure. Of particular note in considering moral implications is the fragility of electronic computing systems. Computers get turned on and off constantly, and in the future we mightn't need specialized computer hardware to build complex (human-like) neural networks. The accidental powering down of a personal computer containing a living entity could happen in an instant, and would no doubt be morally catastrophic.

The ongoing development of modern computing technologies enable programmers and biologists to simulate real biological systems with increasing accuracy. Super fast computers and those that process data in parallel help unlock the secrets of biological nervous systems. But such simulations could mean serious moral ramifications if they are successful in achieving their goal. Furthermore, the results of simulating biological life could be more than just theoretical. It is important for engineers to consider the very likely prospect of (desirable and undesirable) side-effects from the development of technologies discussed above.

Chapter Three: Technology Beyond Our Control

Artificial intelligence is in itself a useful tool for helping automated systems reach their maximum potential. By working intelligently, computers can do more work in less time and even consume less power. But there may be limits to the safety of intelligent systems. Some dystopian views of the future fear that intelligent machines will grow beyond our control and eventually take over the world. On the surface, these fears appear rooted in science fiction, but their basis may not be entirely unfounded. This chapter explores some of the less savory forecasts for the future of intelligent machines from science fiction to scientific prediction.

The Matrix: Intelligent Machines Succeed Humanity

The cinema blockbuster *The Matrix* is more than just a sequence of good special effects. It is a story that in many ways parallels Mary Shelley's classic *Frankenstein*. Both *The Matrix* and *Frankenstein* focus on the consequences of allowing science to get beyond our control. Similarly, both plots derive from the human desire to create life, and in particular, the fantasy of creating life from inanimate parts. The more modern story of *The Matrix* highlights the idea of strong AI, and makes it more real by painting everyone into a computerized world. The movie demonstrates an undesirable scenario that could occur from the creation of strong artificial intelligence. More generally, it supports a theory that sufficiently intelligent machines could replace humanity.

As previously discussed, one major goal of artificial intelligence is the development of more efficient computerized tools. It is only natural that, given a set of tools, we would seek to use them in the most efficient manner possible. Moreover, we as

human beings seem generally fascinated with life. The very idea of creating life-like programs may be what drives many to that pursuit. But it is possible that the unintended consequences of developing intelligent machines, particularly those that might be tantamount to a life form, could be grave for humanity.

The Threefold Danger

Based on the increasing power of computers, a strong artificial intelligence at some point in the future would likely be capable of thinking *at least* as well as a human being, particularly if it were based on a human-emulating neural network. The program could solve a variety of problems, communicate with others, learn, and even be creative. Of course this program doesn't currently exist, but it could do all these things if it did. A logical step would be to embody the intelligence within a machine such as a robot, in order that it may be mobile and sustain its own existence (since a strong AI seeks to be life-like). If many of these machines were built, they could be called a race of robots, and a race of human-like intelligences would likely be a competitor for natural resources.

It is in the nature of human beings to adapt the world to our liking. In general we consider ourselves to be the most important species on the planet. A race of robots might have different ideal living conditions, and, if they were programmed to think like people, robots would probably view themselves as the most important race [Moravec]. This kind of competition illustrates a clear conflict that could result from the development of a strong artificial intelligence.

Bill Joy, chief engineer at Sun Microsystems and author of the manifesto "Why the Future Doesn't Need Us," argues that this is a conflict we would surely lose. Initially we may have the advantage in sheer numbers, but that would quickly deteriorate.

Intelligent robots could easily rebuild themselves. They would have no gestation period. A new robot would be "born" in the time that it takes to put the pieces together. In a factory setting, this could be hundreds per day, per factory. Robots would also have no adolescence. It may take sixteen years to raise a reasonably capable human being, and sixteen seconds to replicate a robotic intelligence. This represents a new type of danger emerging in artificial intelligence technology. A bomb, no matter how powerful, can only explode one time, but a race of robots could replicate itself so long as resources were available, resources for which the robots would surely fight [Joy].

Joy also considers a less violent scenario in which robots accidentally squeeze humanity out of existence. If an artificial intelligence was only as clever as human beings, or maybe even less, humanity might still lose out. Even if the robotic race didn't aggressively pursue the destruction of humanity, they might still seek to change the environment in which they live. They might also still seek to replicate, just as people desire to have children. The robots would continue to serve their own best interests, and consume the resources that people rely on. This type of behavior is similar to the way people harvest forests and squeeze out the species of plants and animals that live there.

A third and still less violent future view is one in which strong artificial intelligence never comes to fruition. It was this prospect that drove the hopelessly antisocial Unabomber to misanthropic insanity. It is based on the idea that weak AI continues to make machines work more efficiently and independently. In many ways, we as a society are already dependent on these intelligent machines. There is not, for example, enough human resource available to sustain the credit card industry without the intelligent programs that rate and track people's credit records. Nor is there sufficient

human resource to maintain power if the very complex software in our nuclear plants were gone. Joy points out what he calls the "New Luddite Challenge," namely that we must temper our desire for technology with our capability to live without that technology. Strong AI notwithstanding, dependence on intelligent systems could be our downfall.

Joy, however, fails to adequately address the sustainability issue with regard to technological dependence. Sustainability refers not to stagnation, but to our ability as a society to continue to develop without using up or destroying the resources that support our existence. Dependence on technology may be good, especially if that technology enables us to extend our banks of otherwise depleting resources. We need only be wary of technological dependence when that dependence causes us to overuse a nonrenewable natural resource.

Almost There

The scenarios above are just a few of the many that have been considered by scientists and science-fiction writers alike. But their significance lies in their urgency. Several leading technological minds believe that machines with this kind of intelligence may exist within our lifetimes. Hans Moravec writes in his 1999 book *Robot: Mere Machine to Transcendent Mind* that he predicts human-like intelligence in computers by the year 2040. These intelligent machines will cost roughly the same as a home computer does today. Moravec's estimates are based on his own professional experience and the current trends of computing technology. In the past, however, his predictions have fallen short of technological advance, rather than surpassing it.

Ray Kurzweil, another pioneer in computing technology, concurs with Moravec on all these predictions save one: Kurzweil believes that computers will surpass human brain capacity in only twenty years. His estimate is based on a computer simulation of human brain functions. Kurzweil used an abstraction of individual thoughts, called chunks, to generate an electronic brain. Although his model was orders of magnitude less complex than a real brain, Kurzweil argued that Moore's Law predicts the forthcoming availability of computing power capable of surpassing the capacity of the human brain [Kurzweil, Age].

One final point from Joy raises concern that artificial intelligence may come to be more than just a computer program. In his manifesto, Joy discusses his work with nanotechnology, miniature machines. Showing similar progress to integrated circuits, nanobots could some day be used in what is called a swarm network. Swarm technology is based on tiny robotic devices that communicate via wireless network. They have very little computing power individually, but are designed to work together in parallel processing tasks. As this technology develops, swarm devices could be used in tandem with neural networking technology. By programming swarm devices to form a specific neural structure, AI developers could create the danger that Joy fears most: a physical embodiment of a strong artificial intelligence.

The development of artificial intelligence, while logical, could have far reaching and unintended consequences. Several technology experts strongly support the idea that truly intelligent machines will exist in the foreseeable future. Unlike other technologies, however, AI systems could develop into a competitive race with which human beings would have to deal in order to ensure our own survival. These observations underscore

the need for ethical and pragmatic foresight in the development of new artificial intelligence technologies. This is not to say, however, that the development of artificially intelligent machines is an entirely fruitless endeavor.

Chapter Four: A More Optimistic View

Many prominent figures in modern technology believe that artificial intelligence will become a reality in the not too distant future. Some also agree that intelligent machines will succeed humanity. Unlike previously discussed dystopian views, however, there are those who welcome the advance of intelligent machines. These futurists believe that human beings will combine with robots or else foster them as our progeny. While wildly technocratic, these views have their own basis in the current trends toward rapid technological advance. This chapter will review the utopian futurist views of artificially intelligent machines.

Not Quite The Matrix: Intelligent Machines Still Succeed Humanity

Technological development, it seems, is inevitable. Technology continues to be driven by the needs and desires of society. Even the luddites draw arbitrary lines between acceptable and unacceptable technology. Those who shun technology in theory surely don't survive by their bare hands alone. Technology is a tool, and there are those who believe that proper use of technology as a tool can help make life itself more fulfilling [Paul]. By this token, it seems logical that continued development of technologies can make life even more enjoyable in the future.

Certainly it is not the goal of computer scientists to develop software that will destroy humanity. It is equally unlikely that engineers in the computing field believe that they will develop an artificial life, only to shut it down and murder it. Computer scientists build their products as a service. These engineers strive to build better programs because they want to better serve those who use the programs. In many cases,

this means building more intelligent software. Sometimes it also means building intelligence into a robot. But robots should not necessarily represent the locust plague. There may perhaps be a scenario in which human beings and intelligent machines share their existences.

Stanley Kubrick & Steven Spielburgs's 2001 film, *AI: The Artificial Intelligence* is a strong modern revision to Kubrick's 1968 production of *2001: A Space Odyssey* (originally written by Arthur C. Clarke). Both feature the introduction of machines with human-like intelligence. However, the newer film portrays robots as more cooperative and aware of their own fallibility. This is a more human-like upgrade to the unruly and overconfident HAL 9000. *The Artificial Intelligence* also paints for humans a more disdainful and regrettably more probable attitude toward intelligent machines. Given the creation of sufficiently intelligent machines, the movie shows how machines and people might live together. It envisions robots in a largely subservient role, providing continually greater service to their human counterparts, including even various emotional services.

The end of Kubrick's cinematic vision predicts that robots outlive humanity and carry the torch of life on Earth when the environment becomes too inhospitable for human life. Hans Moravec finds this to be an attractive and likely scenario for the future of humanity and robotics. He believes that our human desire to propagate our species will eventually manifest itself in a more metaphysical way, and that we will surrender dominance of Earth in exchange for a sort of immortality [Moravec]. The development of robots with an intelligence of their own could present a way for humans to outlive

themselves. It may be possible that our desire to live on would be satisfied by the living of immortal machines that we foster with our own thoughts and teachings.

Kurzweil again outdoes Morvec by predicting that we will not allow robots to succeed us. Instead, Kurzweil writes that we will join with machines and become a race of cyborg humans [Kurzweil, Age]. The change, he admits, will happen slowly and gradually. As evidence to support his position, prosthetic devices are becoming more commonplace and more technologically advanced. We have developed artificial legs and arms, even artificial hearts. As we learn more about the human body, there is little to stop us from emulating it in technology. Kurzweil believes that eventually we will even have microcircuits in our brains [Kurzweil, Man]. These microcircuits will be capable of increased mathematical processing and memory storage. They will also, Kurzweil purports, allow us to directly manipulate our own thoughts. By running a program on these circuits, we can and will live in an increasingly virtual world that will be so real to us that we can't tell the difference from reality.

There are some points about Kurzweil's vision that are appealing. It might be nice, for instance, to directly stimulate our own joy. But human beings are not likely to become cyborgs, at least not in the near term. Despite the continuing rush of technology, people will not accept such a drastic mutation of our bodies. Computer programs ultimately exist to serve society, and there will not be enough social support for a race of half-humans. As dependent as we are on technology, there is still something sacred about our bodies. Human beings do not want to be robots, and we might not really even want to live forever.

Of course, there is no accounting for science. Arthur C. Clarke's first law of technology states,

"When a scientist states that something is possible, he is almost certainly right. When he states that something is impossible, he is very probably wrong."

In other words, history has shown that technology is like an unstoppable train. Human beings have learned to fly through sky and space, and travel to the greatest depths of the ocean. Preparing for the unpredictable future is more about prospects and probabilities than about certainties [Neeley].

Solid testimony from some industry leaders supports the idea that we may soon be living amongst artificially intelligent machines. The opinions of experts represented here certainly do not guarantee the eventual creation of truly intelligent machinery, but we must plan according to what *may* happen because we don't know what *will* happen. The eventual development of powerful artificial intelligence systems may or may not lead to a maligned race of robots. Any outcome, however, will certainly carry serious consequences for engineers and all other citizens. We must, therefore, be mindful throughout our journey into the future of AI development, and be prepared for whatever we find there.

Chapter Five: Recommendations

This paper considers artificial intelligence and its place in our future. The predictions presented have multiple variations, but one clear underlying theme. Several leading minds in the fields of computers and robotics believe that artificially intelligent machines will be created. This chapter outlines recommendations for the engineering community to foster the artificial intelligence movement in a safe and sustainable way.

Programming For a Sustainable Future

Ultimately, sustainability is our best ally. Without the future we have nowhere to go (Poritt). It is therefore essential that each person and professional do his or her part to build a sustainable future. That is not to say that progress should stop, only that care should be taken to consider the consequences of technological development. We should continue to develop, but not in a way that is harmful to humanity or to our planet. For computer scientists, this may not seem like a difficult task. We may not realize that we are capable of building a non-sustainable future. We may or may not have the power to build technologies that can destroy our future. It is our responsibility as professionals not to develop such technologies if we can help it. This, again, is not to say that we should stop developing technologies altogether. Any technology, new or old, can be used improperly, a problem as impossible to avoid as it is to predict.

In spite of the ethical points raised in this paper, we should not forget our most obvious professional responsibility. In order to remain useful in this society, computer scientists will continue to develop the technologies that society demands. This is the reason that we are all engineers, to build technology. The public will continue to want

new and exciting things. They will continue to need better interfaces and more complex software systems. Building more intelligent software is the best way to meet these new needs. Machines that are artificially intelligent, in some capacity, will continue to become a necessity. More people will crowd onto our planet, needing more resources delivered to them at a faster rate. Only technology will provide that.

On the other hand, we are required by our own professional ethics to protect the public from that which may harm them. We can do nothing for society if we allow technology to get beyond our control. The responsibility falls on us because we know better than anyone what we are capable of creating, and we should know better than anyone what those creations are capable of doing. This is true for any technology, not just artificial intelligence. While we may not see the danger or potential in any of our products, it is time to start thinking more seriously about our effect on the rest of the world.

Meeting Higher Levels of Responsibility

My research throughout this project led me to believe that computer professionals, as a community, lack a strong governing body. As we blaze ahead into the future and create new, wonderful, and possibly dangerous things, we need guidance. While our products affect as many people as those of any other industry, it remains true that there is no FDA to test and regulate the production of our software, and no Bar Association to keep us from practicing computer science in an unethical or unprofessional way. It seems we need someone to make sure that we are doing the right thing. In hindsight, however, these organizations are not the answer. They are not plausible. No one body

could possibly keep watch over all computer programmers. We would be so bogged down in ourselves that technologies would never come to fruition.

On one level, we must therefore practice individual self-governance. We must be the ones to make sure we are doing the right thing. As engineers, we should be proud to have a code of ethics, and hold it close to our hearts. But there is only so much we can each be expected to do individually. Individual moral standards, while necessary, are not sufficient. We have taken it upon ourselves to be technological leaders of our generation, and are responsible for acting like leaders.

The growing influence of computer systems demands that computer scientists become more active in the decision making process. It is too much responsibility for every computer programmer to evaluate the moral justification of his or her project every day. An individual programmer may not even know what puzzle his or her code will fit into. Management hierarchies exist because history has shown that the evaluation of moral, ethical, and logistical dimensions of a work product is in itself a full-time job. Computer programmers producing their work pieces need to be confident that the decisions handed down to them are trustworthy, safe, and ethical.

Yet, when safety budgets are recalculated, only engineers can fully understand how much security is safe enough. Only engineers who are intimately involved with multibillion-dollar space shuttles should ultimately decide whether ambient conditions are safe for launch. It would seem that engineers should have a strong hand in developing domestic and foreign policies, as advancing technology will continue to make our world a smaller place. Ethical people with technological knowledge should also pursue politics, management, and other policy drafting fields. That is how to ensure that

developing technologies will be useful and safe. Not all engineers can be digging the trenches.

AI will surely continue to develop in the future. In whatever methodology we choose, it is our duty to be mindful of the future. As a basis we know that we must build the right product, and build it as best we can. But we must also ensure that our developments maintain a sustainable future.

Chapter Six: Conclusion

Summary

Artificial intelligence is the design and study of computer programs that react flexibly and intelligently to a wide variety of situations. It has growing influence in new computer related technologies and makes many complicated tasks possible. The development of new hardware and techniques is fueling an ongoing movement to build computer systems that can understand and think in a cognitive way. While the potential advantages of such systems is yet unknown, equally unknown are the potential pitfalls of developing intelligent machinery.

Several technological leaders point to the course of history and their own experiences in saying that artificially intelligent machines may soon become a reality. These machines, if developed, may outlive and outgrow humanity on Earth. They may forcefully take over the planet, or may not take it over at all. Human beings may even learn to evolve into machines and reach a sort of immortality. In scientific outlooks, we must prepare for what is possible, rather than what is certain. Engineers are best suited to spot potential pitfalls of AI and other technologies, and should individually adhere to stringent professional ethics in the practice of their art. But it is equally important that ethical people with engineering education and experience become more intimately involved in decision making and policy drafting processes through communication and an expanded educational curriculum.

Interpretation

This paper developed a thorough picture of the study of artificial intelligence and outlined its usefulness in computing applications. It explored the movement to develop strong AI systems and addressed some non-technical philosophical issues involving that development. Evidence introduced to support arguments that intelligent machines will be a part of our future compelled a set of recommendations intended to guide engineers in their continued development of intelligent computer programs. The recommendations, constituting a primary product of this project, represent subtle changes in the social role of engineers, but will become increasingly important as technology grows.

I did not present a counterargument that artificially intelligent systems may never come to fruition, but stated repeatedly my justification for that omission. As there is currently no strong AI existing, I felt it less necessary to develop an argument for continuation of this condition. The material discussed throughout the project was complex and, regrettably, could be only minimally developed in its complete scope. Even so, the material presented has allowed the project to achieve its two primary goals, namely explaining current and developing AI technologies in a way accessible to non-experts, and outlining recommendations to prepare future engineers for their growing ethical and professional responsibilities.

Further Recommendations

This project will be most effective as a catalyst in a movement toward more socially responsible engineering. There is much more research to be done in the field of neural biology, to help ensure that we do not get ahead of ourselves in the practice of

mimicking brains. Bodies that currently govern the standards of computer science should adopt a more active role in social policy drafting. Namely, computer and electronics societies should become more involved in state and national politics. Additionally, education of engineers should be more biased toward preparing engineers to take responsible management positions and to participate in more socially active fields like politics. Preparation for these roles can be achieved by first giving high school children more exposure to basic engineering principles. Allowing children to have a broader base of education will enable a more rounded engineering education at institutes of higher learning. Colleges and universities should also draft stronger compulsory education for engineers in the fields of communication, philosophy, and business.

Evaluating a Social Experiment

A typical engineering thesis can be looked at as a social experiment. Each thesis aims to accomplish something important and may or may not affect society adversely. Responsible engineers must consider these possibilities when performing their experiments or developing their technologies. This project does not require an experiment to be performed or a new technology created. In fact, this project is designed to be an evaluation of a social experiment already in place. That social experiment is the development of artificially intelligent computer programs. AI is a case of technology that is breaking through new frontiers, and any such technology would likely be a social experiment, intended or not.

This project is itself a more direct kind of social experiment. While many engineering projects have social side effects, this paper aims to create a direct social

impact. It remains to be seen whether the project will adversely affect the development of new technologies, or possibly spur on the development of new standards. Even if the recommendations are enacted in good faith, various problems may arise. It is unknown, for example, whether pushing computer scientists away from computers and into management positions will lower the quality of new technologies. Furthermore, we can't say yet whether computer scientists and other engineers could even become decent politicians. Life in the political scene and answering directly to the public may also strip away the sense individual self-governance that this project encouraged. This is only the beginning of a larger project, the development of a more responsible and versatile community of computer programmers.

Bibliography

Fiction

1) Asimov, Isaac. I, Robot. New York: Doubleday, 1950.

2) Branagh, Kenneth. Mary Shelley's Frankenstein. Film. TriStar Pictures, 1994.

3) Clarke, Arthur C. 2001: A Space Odyssey. New York: Roc, 1993.

4) Kubrick, Stanley. 2001: A Space Odyssey. Film. Metro-Goldwyn-Mayer, 1968.

5) Shelley, Mary W. Frankenstein, 2nd ed. Ontario: Broadview Texts, 1999.

6) Wachowski, Andy and Larry Wachowski. The Matrix. Film. Warner Bros, 1999.

Non Fiction and Exposition

7) Centurion (University of Virginia). "Centurion: Legion Project Testbed." Online. Internet. 21 January 2002. Available: http://legion.virginia.edu/centurion/Centurion.html

8) Clabaugh, Caroline and Dave Myszewski and Jimmy Pang. "Neural Networks." Online. Internet. 15 March 2002. Available: http://cse.stanford.edu/classes/sophomore-college/projects-00/neural-networks/

9) Dean, Thomas and James Allen and Yiannis Aloimonos. Artificial Intelligence: Theory and Practice. Menlo Park: Addison-Wesley, 1995.

10) Joy, William. "Why the Future Doesn't Need Us." Online. Internet. 19 June 2001. Available: http://www.wired.com/wired/archive/8.04/joy_pr.html.

11) Kaczynski, Theodore. "Industrial Society and Its Future." Jointly published under duress by The New York Times and The Washington Post, 1997.

12) Kurzweil, Ray. The Age of Spiritual Machines. New York: Penguin, 2000.

13) Kurzweil, Ray. "Man and Machine Become One." Business Week 6 September 1999: 260

14) Legion (University of Virginia). "Legion: A Worldwide Virtual Computer." Online. Internet. 21 January 2002. Available: http://legion.virginia.edu/

15) Moore, Gordon E. “Cramming More Components Onto Integrated Circuits.” Electronics 38 (1965): 8.

16) Moravec, Hans. Robot: Mere Machine to Transcendent Mind. New York: Oxford University Press, 1999.

17) Neeley, Kathryn, Professor of the Technology, Culture and Communication Department at the University of Virginia, Charlottesville, Virginia. Personal Interview. 20 March 2002.

18) Paul, Gregory S. and Earl D. Cox. Beyond Humanity: CyberEvolution and Future Minds. Rockland: Charles River Media, 1996.

19) PNNL (Pacific Northwest National Lboratory). “Neural Networks.” Online. Internet. 21 January 2002. Available: http://www.emsl.pnl.gov:2080/proj/neuron/neural/neural.homepage.html

20) Searle, John R. “Minds, Brains and Programs.” The Nature of Mind. Ed. David M. Rosenthal. New York: Oxford University Press, 1991. 509-519.

21) Turing, Alan. “Computing Machinery and Intelligence.” Mind 59 (1950) 434-460.