A Philosophical Survey of Artificial Intelligence

by Danny Faught
University of North Texas
Senior Honors Colloquium

presented to Dr. Derek Baker
November 18, 1992

minor revision January 9, 1993

TABLE OF CONTENTS

" 'The issues deserve serious public debate. They are too scientific to be left to philosophers and too philosophical to be left to scientists.' "[1]

[1]Anthony G. Oettinger of the Aiken Computation Laboratory at Harvard University, from *What Computers Can't Do*, by Hubert Dreyfus, p. xi.

Foreword

This project was inspired by a shorter paper I wrote as a freshman. The paper was titled "Digital Ideas," and it wasn't very good. But I have never forgiven the professor who, in addition to the abundance of red writing, had the gall to comment on the very title of my paper. He circled "Digital," drew an arrow to "Ideas," and wrote "Cute, but self-contradictory." That was the catalyst that has resulted in this paper, and the attitude that I hope to change.

Introduction and Tour Guide

My goal is to make you think about some things that maybe you haven't considered before. Most people have heard of artificial intelligence (AI), but most people don't know what it means. In this paper I will explore some of the history and personalities of AI, and some of its philosophical ramifications. I will also discuss AI's future, and the effect that AI may have on us in the future. Sometimes I will stray into topics relating to computers in general, and often I will refer to computers and robots interchangeably. I have covered a broad spectrum, probably too broad, but the research was so fascinating that I couldn't bear to leave any of it out.

I will not go into the gory details of any particular AI technique. While a general interest in computers would be helpful, I hope that people with no computer background will be able to follow my reasoning. I will introduce you to a few of the many excellent books and papers about AI. My research led me from cognitive psychology to quantum

physics, and it even changed my religion. While I don't expect this paper to inspire your life's work like it did for me, I hope it will open your mind. If your interest is sparked here, I encourage you to explore the field yourself.

A note on my research method: I have noted my research methods herein where appropriate, except where I will often make generalizations regarding the views of the modern thinkers. Some of this information comes from my research in the literature, and some comes from months of monitoring and interacting on the Internet, especially the Usenet news network. A sizeable proportion of the current researchers in the field are connected via the Internet, and it served as an invaluable aid.

Where is AI going? No one can be sure, but never, ever, say "never."

Reader's Guide study[2]

To get a grasp on how AI fits into the historical picture of computer science, I studied some relevant topics in the *Reader's Guide to Periodical Literature* to see when they first appeared and how they evolved. I used the *Reader's Guide* index because it seemed to be the best way to get a fairly complete and unbiased overview of the evolution of this topic.[3] The index does not cover the extremely technical journals, but

---

[2]If you think history is a little dry, bear with me. It gets more interesting.

[3]The *Reader's Guide to Periodical Literature* was published every few years in the earlier part of this century. As the volume of articles indexed has grown, the editions became more frequent and it is now published annually. In the following text, the term "entry" shall mean a reference to a magazine article listed under a particular topic.

rather the more mainstream magazines. So we can assume that if a topic merits space in a general-purpose magazine, then it has been declared to be a topic worthy of the attention of the general public. Furthermore, by the time a separate heading appears for a topic in the *Reader's Guide*, then it has been given significance by a number of different publishers. This should give a fairly good indication of when a subject has been recognized by at least a noticeable fraction of the general public, but of course this will tend to lag behind the early research in a field that is only published in the technical journals. This is similar to the content analysis that John Naisbitt did for *Megatrends* book and *Trend Report* newsletter.

If you go back far enough in the *Reader's Guide*, of course, there is no topic for computers at all. The precursors to the computer were the mechanical and electro-mechanical calculating machines. The "Calculating machines" topic first appeared in the July '32-June '35 index. The topic had an appreciable number of entries by the time the July '41-June '43 index was published, and it grew steadily in the number of entries until the middle of the 60's. "Computing machines" appeared sporadically throughout the time period investigated, but mostly as a cross-reference to "Calculating machines."

In the March '57-February '59 index is the first appearance of subtopics under the "Calculating machines" heading. In the early 60's, a "Digital computers" section appears under "Calculating machines." Starting in the March '65-February '66 index, "Calculating machines" is cross-referenced to "Computers" and "Computers" is used for this topic from then on. The "Computers" topic grows from this point at a faster rate than that of the "Calculating machines" topic, until the present time where the 1991 index has eleven

and a half pages dedicated to "Computers." This topic is now full of subtopics like legal use, financial services, industrial use, and medical use. We can see that the computer science field has become highly diversified and specialized.

"Calculating machines" exists as a cross reference to "Calculators" until the 1980's, but its last real entry was in the March '72-February '73 index. "Calculating devices" appeared a few times in the early 1970's but didn't last long. The "Calculating machines" traffic migrated to "Calculating machines, electronic" through the 1970's. The "Calculators" topic was first used in the March '75-February '76 index and has continued with an average of a third of a column of entries until the present. The "Calculators" topic is for calculators in the modern sense of the word, that is, hand-held calculators. So we can see that the original "Calculating machines" heading has effectively split into "Computers" and "Calculators."

The "Cybernetics" heading is an anomaly in the rest of the evolution. It first appeared when Norbert Wiener, the founder of the field, published an article on the subject in 1948. An entry appears under this heading about every two years, and they continue up to the present. The appearance of articles is quite sporadic, but hasn't shown any real growth or decline.

"Artificial intelligence" finally became a heading in the March '66-February '67 index. The first article under this heading is from an article by Marvin Minsky, who is considered one of the fathers of AI. It was titled simply "Artificial intelligence."[4]

---

[4]Marvin L. Minsky, "Artificial Intelligence," *Scientific American*, vol. 215, September 1966, pp. 246-252+.

Minsky's article included a historical overview of the AI, going back to 1943, and an explanation of some of the significant algorithms. Interestingly, in this index there is no cross-reference to "Artificial intelligence" from the "Computers" heading, and "Artificial intelligence" has no cross-references to anywhere else. In the March '72-February '73 index, "Artificial intelligence" is cross-referenced to "Computers." But there is no cross-reference from "Computers" to "Artificial intelligence" until the March '83-February '84 index. Until and including the March '83-February '84 index, most AI articles were still listed only under "Computers" or at least duplicated in the "Computers" section. In the March '83-February '84 index, there is a noticeable shift. Starting in this edition, virtually all of the articles relating to AI are listed under "Artificial intelligence," with no duplication under "Computers."

A few other relevant topics have appeared relatively recently. "Computers and civilization" first appeared as a separate heading in the March '80-February '81 index, showing an increased interest in the impact of computers on society. "Artificial life" first showed up in the 1988 index, and has also appeared in 1990 and 1991. Like what we saw in the beginnings of the "Artificial intelligence" topic, there are no cross-references to or from the "Artificial life" heading as of yet.

Philosophy & AI—at the cutting edge

AI is considered by many people to be the cutting edge of modern technology. To understand why, it is useful to compare AI to the field of philosophy. During the heyday of the Greek philosophers, the sciences were not clearly separated. The scientists

studied a broad range of subjects, which was not hard to do given the scarcity of scientific knowledge in those days. All of the sciences were lumped under the umbrella of "philosophy." This was appropriate, because the sciences were mostly speculation. Since there was very little experimental data to work with, the scientists had to approach subjects at the much higher level of philosophy.

Philosophy can be considered the highest level of discourse of all the sciences. It views things in a very broad and vague way, relying on little or no empirical evidence. This is not to reduce the importance of the field of philosophy. The philosophers, along with the science fiction writers and the dreamers, have consistently predicted advances in the hard sciences before they happen. A classic case in the computing field is Charles Babbage. Babbage had developed the plans for a computer by 1835, but wasn't able to actually build it under the conditions at the time. The popular legend has it that the limiting factor was the limited technology of the time—Babbage just couldn't find the parts to build the machine he had designed. There is speculation, however, that the real limiting factor was the simple fact that very few people could see much use in such a device, which means there were few people willing to invest in the research or to invest in the finished product. Even so, there is no doubt that the contributions of Babbage and others like him eventually led to a very real phenomenon. There are scores of stories about inventors of both devices and ideas who died in poverty, but their contributions lived on, eventually took root, and proved wildly successful.

Nowadays we don't have philosophers doing all of our scientific research. As the first scientists gained knowledge about a particular subject, their ideas became more

grounded in facts provable by experiment and they no longer needed to rely on philosophical speculation to form ideas. Thus the sciences broke off from philosophy and became fields of their own. The fields of biology, mathematics, chemistry, and medicine, for example, have so much accumulated knowledge that they themselves have split into sub-disciplines, and researchers are specializing in a particular branch of science because the sheer bulk of scientific knowledge is too overwhelming for one person to grasp.

Computer science is another field that has split into a multitude of sub-disciplines. To be able to do any useful work, a computer scientist must specialize in a field such as database management, software engineering, or systems administration. Another option is the field of artificial intelligence.

According to Minsky's article, the first significant work in AI showed up in 1943 with the publication of three theoretical papers on what is now called cybernetics.[5] The quest to formalize intelligence had started centuries ago. But in the 1940's, when it became feasible to build a digital computer, we see research in using machines to instantiate an artificial intelligence. By the mid 1950's, according to Minsky, computers had reached a level of performance that would support AI projects, and AI began to move from the realm of theoretical speculation to the realm of possibility.

---

[5]Marvin L. Minsky, "Artificial Intelligence," p. 247.

Steven Levy attributes the impetus of AI with a conference at Dartmouth University in 1956[6]. As shown in the *Reader's Guide* study, AI started to become recognized by the media as a legitimate field in the mid 1960's. By the mid 1980's, AI had truly come of its own. But in the dawn of the computer age, AI was not distinguished from computer science at all. Articles in the 1940's had titles like "It thinks with electrons," "100-ton brain at MIT," and "Robot Einstein." Prior to this point, no one had seen anything that mimicked the human brain so closely. Of course, we know now that the similarity was very superficial, but the fact remains that the development of the computer increased our ability to simulate human brain functions by at least an order of magnitude. Thus in the earliest computer articles we see anthropomorphisms like "think" and "brain" applied to computers, even though they did little more than add columns of numbers very rapidly.

Sometimes, early computers were even labelled "mathematical robots,"[7] probably in the hopes that the mathematical prowess that the computers had demonstrated could easily be used to control robots. On a more down-to-earth level, the robot metaphor arose from the fact that once a computer program was started, it ran autonomously through a very complex procedure, with no human help to guide it besides the instructions that they had fed in. In 1962, when Dr. Allen Newell and Dr. Herbert A. Simon developed the "General Problem Solver," the media declared that a computer

---

[6]Steven Levy, "The Riddle of Artificial Life," *Popular Science*, vol. 241, October 1992, p. 100.

[7]"Robot Einstein," *Newsweek*, vol. 26, November 12, 1945, p. 93; and "100-Ton Brain," *Senior Scholastic*, vol. 48, February 4, 1946, p. 36.

had been made to think.[8] The words like "brain" and "think" were all still protected by quotation marks, but the attitude toward the anthropomorphisms had begun to change.

During the 1970's and 1980's, the speculation about computers thinking like humans had become more serious. People were exploring the issues underlying a "thinking" computer. Some gleefully proclaimed that there was no limit to AI, that science fiction would soon be science fact. Others provided carefully prepared arguments that the human mind would never be challenged by computers. In either case, people had lost the gee-whiz attitude of the 1950's and were seriously approaching the subject. And they started to wonder whether the quotation marks around "think" would someday have to be removed.

The evolution of AI is very similar to the evolution of philosophy. They started out being synonymous with the overall field, and have since become recognized as fields separate from the hard sciences. Many of the truly new ideas come from these fields. New ideas develop from other areas, to be sure, but they can often be categorized into one of the previously existing disciplines. AI is simply philosophy as applied to the computing field.

As I pored over the listings in the *Reader's Guide*, I rarely found interesting articles in the specialized subsections under "Computers." The best place to look was in the beginning of the "Computers" section where all of articles that didn't fit in one of the established subsections were located. In that section I found the thought-provoking, future-contemplating, and ground-breaking articles. When the "Artificial intelligence"

---

[8]"Computer Made to Think," *Science News Letter*, vol. 81, January 6, 1962, p. 5.

heading broke off from "Computers," the interesting articles followed, and "Computers" became interesting only to those with an interest in a specialized area of computing other than AI.

After noticing that trend, I related it to the bookstores. Probably ninety-five percent of the books in the computer section at the bookstores are written about a particular software package or programming language or operating system. If you don't have a need to know how to use those particular programs, then you won't have the least bit of interest in those books. However, there are usually a few books hidden in a corner somewhere that apply across all disciplines. This minority represents the philosophy of computer science.

Philosophy has been described as the study of "everything else" which isn't well understood. A common quote in the AI circles is "AI is everything that machines currently can't do."[9] Some things that couldn't be done are now possible because of AI research. Just like the sciences have branched off from philosophy, many of the areas of computer science started off as an AI study. The following subjects have been attributed with originally being considered artificial intelligence, but now, to varying degrees, are now fields of their own:

compilers
time-sharing systems
linked lists, tree algorithms, and graph search techniques
symbolic manipulations such as integration & differentiation
object-oriented programming (like Smalltalk)

---

[9]No one has stepped forward to claim that they are the originator of this quote. John McCarthy is often credited with it, but Mr. McCarthy told me that they are not his words and that the quote does not reflect his current philosophy.

advanced forms of input and output such as the mouse
local networks such as Ethernet
high-level command interfaces (windows, menus)
functional languages like LISP and Prolog
chess programs
expert systems
theorem proving
image processing and control theory
optical character recognition
computer vision
robotics[10]

Two AI Approaches

Before delving any further into the issues of AI, it would be prudent to discuss two very different approaches in AI. I will adopt Joseph Weizenbaum's terms: "simulation mode," and "performance mode." The simulation mode in AI involves simulating the way humans do things. The performance mode, on the other hand, means doing things in an efficient or straightforward way, which will not necessarily relate to the way it is accomplished by the human brain.[11] Geoff Simons has introduced a similar system of classification: organic, which relates to simulation mode, and inorganic, which relates to performance mode.[12] Both of these approaches have been illustrated in examples

---

[10]Thanks to the following people who suggested the items in this list: Scott Hibbetts, Bruce Krulwich, Michael Chui, Cameron Laird, R. Bharat Rao, and William J. Rapaport. Several of these suggestions come from an article by Peter J. Denning, "The Science of Computing: What is Computer Science?" *American Scientist*, vol. 73, Jan-Feb 1985, p. 16-19. In this article, Denning doesn't focus only on what is AI, but he lists branches of computer science that stand on their own, as well as branches the branches that he classifies as AI. A few items came from Bruce W. Arden, ed., *What Can Be Automated? The Computer Science and Engineering Research Study*, MIT Press, 1980, p. 501.

[11]Joseph Weizenbaum, *Computer Power and Human Reason: From Judgement to Calculation*, 1976, pp. 164-166.

[12]Geoff Simons, *Are Computers Alive? Evolution and new life forms*, 1983, pp. 8-9.

throughout history. Simulation mode (organic) can be characterized by Frankenstein's monster and the robots of science fiction. Performance mode (inorganic) can be characterized by the golden maidservants in the *Iliad* and the golems of Hebrew folklore and modern fantasy.

If we are to achieve the goal of making a human-like robot, we will need to concentrate on simulation mode. The human brain and the body that contains it are remarkable pieces of machinery, and we would have a wonderful tool if we could simulate them accurately. There are actually benefits on both ends of the spectrum here—not only are we building computers and machinery that attempt to simulate aspects of humanity, but also the research has given us a better understanding of the human body. To be able to simulate how something works, we first need to have a thorough understanding of how it works. The dream of simulating humans has sparked research and new discoveries in areas such as the physiology of the human body (in order to build robot limbs) and brain functions (in order to build an artificial intelligence and control program).

Simulation mode, however, can be too limiting. After all, the cave men did not invent tools to merely simulate human functions. A hammer no harder than the human hand would not be very useful. The purpose of tools is to provide something that we don't already have, to extend our capabilities. This is the goal of performance mode. We know that humans are fallible in many ways. Our physical systems can break down; our memory fades with time and sometimes performs erroneously. Do we really want to

simulate those features of humans, too? The performance approach opens our minds and our horizons to unlimited possibilities.

Hubert Dreyfus puts these two modes into a timeline.[13] He defines a period from 1957 to 1962 in which cognitive simulation was the dominant approach in AI. Since then, the focus has been on what Dreyfus calls "semantic information processing," which relates to performance mode. This suggests that AI is following a strategy that can be seen in many other places. At first we imitate something else that works well, and after we understand it better, we can do original work rather than relying only on imitation. Even though cognitive simulation is no longer the only focus of AI research, it is certainly not dead, and in the current literature I still see both approaches. I think it is important that anyone starting an AI venture keep both in mind, and be conscious of which approach is the goal for each aspect of the project.

Anthropomorphisms

Anthropomorphisms abound in the field of computers. Much of the discussions between AI researchers consist of arguing whether or not a computer can exhibit particular human traits. I have made a compendium of anthropomorphisms, all of which finish the question, "Can a computer..."

act intelligently?
be intelligent?
learn from training or from experience?
think?

---

[13]Hubert L. Dreyfus, *What Computers Can't Do: The limits of artificial intelligence*, revised edition, 1972, 1979.

understand?
have goals?
be conscious?
be sentient?
have a mind?
have a soul?
be creative, do something really new, have initiative?
have feelings?
fall in love, or make someone fall in love with it?
have moods?
have common sense?
have a sense of humor?
make mistakes?
experience pain?
use words properly?
believe?
make judgements, tell right from wrong?
reproduce?
be kind, resourceful, beautiful, friendly?
enjoy strawberries and cream?
be the subject of its own thought?
experience stress?
have a nervous breakdown?
have as much diversity of behavior as a man?[14]

The various thoughts above can be summarized and extended by the following questions, again answering the question, "Can a computer..."

imitate a human being?
be indistinguishable from a human being?
be alive?
exceed humans in intelligence?
take over the world?

As computers increase in their ability to mimic people, it is natural to compare them to people and see if descriptions of human abilities can apply to computers. Can a computer act intelligently? Most people would say "yes" without much hesitation.

---

[14]Several of these come from A. M. Turing, "Computing Machinery and Intelligence," 1950, reprinted in *The Mind's I*, p. 61.

Computers can do calculations and draw conclusions in certain fields not only as competently as humans, but more accurately and several orders of magnitude faster. But can a computer *be* intelligent? Did you hesitate on that question? The jump from appearing intelligent and actually being intelligent is a big one. This kind of jump is at the core of many of the anthropomorphic issues.

Intelligence

How do you judge whether a computer is intelligent? Compare it to human intelligence, of course. But that's easier said than done. We must be careful devising a test for intelligence for computers. If it is too lenient, then the test is not serving its purpose. If it is too strict, then many humans who are considered intelligent may fail the test! Joseph Weizenbaum argues that you simply can't compare different intelligences.[15] He offers several examples of people who may not pass an intelligence test:

- An unschooled mother who can't compose a grammatically correct sentence but can make intelligent decisions about her family
- An eminent scholar who doesn't know high school algebra
- A genius who can't manage his private life

Each of these people has a flaw that may cause them to fail an intelligence test. But each has enough overall intelligence to be considered intelligent. Now consider another of Weizenbaum's examples: a computer that beats checker champions but can't change a diaper. Again, we have signs of intelligence, but flaws that may cause disqualification.

---

[15]Joseph Weizenbaum, *Computer Power and Human Reason*, p. 205.

Asking a computer to change a diaper may be asking too much. We should be able to contemplate the intelligence of a computer without involving physical actions. A person paralyzed from the neck down should certainly be classified as intelligent. Even if a person were paralyzed to the point that he had no motor or communication skills, but his brain still functioned normally, he is intelligent. It is unlikely that anyone would realize that he is intelligent, however, and his intelligence wouldn't do much to ensure his survival. But the capacity is still there.

It is important to realize what kind of intelligence I'm talking about here. When someone is smart, we call them intelligent. When someone is dimwitted, we may call them unintelligent. This is a rough way to describe degrees of intelligence, but the seeds of intelligence are there in both cases, regardless of degree. I'm talking about the general idea of intelligence that distinguishes a rock from a person. Note that we could also talk about thinking or consciousness here without changing the point of the discussion.

Even considering purely mental activity, though, we run into problems with computers. We have wonderful programs that play checkers or prove theorems or identify patterns. But none of them exhibit the *general* problem-solving abilities that humans do. We'll return to this topic later when we explore learning.

The problem in any intelligence test is that intelligence cannot be measured linearly. That is why the SAT and ACT tests are so controversial in the United States. Intelligence has many features, and in any individual, these features can be found in varying strengths. This suggests a two-dimensional measure of intelligence, but I think it

is much more complicated than even that. When we can't even devise a test to prove that intelligence exists, we certainly shouldn't claim to be able to accurately measure it.

Thought and the Turing Test

The argument for computer thought is similar to that of intelligence. Man's tools have always matched or outperformed our physical abilities. Cars travel faster and longer distances than humans can, and airplanes can fly, which we cannot do at all.

> But to be able to *think*—that has been a very human prerogative. It has, after all, been that ability to think which, when translated to physical terms, has enabled us to transcend our physical limitations and which has seemed to set us above our fellow creatures in achievement. If machines can one day excel us in that one important quality in which we have believed ourselves to be superior, shall we not then have surrendered that unique superiority to our creations?[16]

Back in 1950, Alan Turing published a paper that pondered the question "Can machines think?" To aid in his illustration, he formulated a test which has now become a benchmark for performance the AI field.[17] Turing called it the "imitation game." It consists of a human and computer that is running software that imitates a human. In another room is another human who is the interrogator. This interrogator knows the other human and the computer as "A" and "B," but he doesn't know which is which. The interrogator communicates with the other two through a medium which will not reveal the subjects' physical characteristics, such as a computer terminal. The object of the game is for the interrogator to determine which subject is the computer and which is

---

[16]Roger Penrose, *The Emperor's New Mind*, p. 3.

[17]Alan. M. Turing, "Computing Machinery and Intelligence," 1950, *The Mind's I*, pp. 53-67.

the human by posing questions through the terminal. Turing provided the following dialogue as an example:

Q: Please write me a sonnet on the subject of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764.
A: (Pause about 30 seconds and then give as answer) 105621.
Q: Do you play chess?
A: Yes.
Q: I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?
A: (After a pause of 15 seconds) R-R8 mate.[18]

Turing predicted that by the year 2000, computers with a billion units of storage would be able to fool the interrogator 30% of the time when given five minutes of questioning. He did not conclude that computers will be able to think, however. The purpose of the imitation game was to avoid that sticky issue. Rather than trying to design a ruler to measure thought, Turing proposed to base the judgement solely on the *behavior* of the subject, without regard to how the thought processes were implemented.

Turing's imitation game is now referred to in the literature as the "Turing Test." A computer system sophisticated enough to fool the interrogator, or at least leave him in doubt as to its humanity, is said to be able to pass the Turing Test. Turing compared his test to a viva voce (oral examination). "The game (with the player B omitted) is frequently used in practice to discover whether someone really understands something or has 'learned it in parrot fashion'."[19] This is a significant comparison, because the viva voce is recognized as a valid test for understanding. A core argument of many AI

---

[18]Ibid., p. 55.

[19]Ibid., p. 60.

pundits is that the computer can only spit out its output in "parrot fashion," even if it passes the Turing test. But if a computer passed the Turing Test, it would show that it knows enough about a topic to pass a viva voce test and that it possesses a reasonable understanding of the topic by the standards we apply to our own children.

In 1965, Joseph Weizenbaum wrote a program called "Eliza" that generated a good deal of excitement. Eliza simulates a Rogerian psychiatrist in a dialogue with a human subject. Many people who used the Eliza program thought they were talking to a human psychiatrist who was typing at a remote computer terminal. Weizenbaum reports that his secretary asked him to leave the room so she could converse with the "doctor" in private! Some human psychiatrists started worrying about their job security, and others welcomed the opportunity to have some automated help. There have been many programs similar to Eliza that have surfaced. In fact, the algorithm is now taught in college symbolic processing courses.

Ironically, Weizenbaum was thoroughly disgusted with the response he got. He knew that all the Eliza program did was a simple pattern match on the users' responses, and it just picked from its list of stock answers based on a few heuristics. Many of Eliza's responses were formed by simply turning around the user's statements in questions. Weizenbaum couldn't believe that such a simple-minded program could have passed the Turing Test, but to many of the people who used it, that's exactly what it did.

A limited form of the Turing Test has manifested itself in the form of the annual Loebner Prize competition. Every year, a group of computers, human subjects, and interrogators gather to determine which programming team wins the $100,000 prize.

There are eight terminals, and the interrogators take their turns at the terminals round-robin style, making a judgement on whether they are communicating with a human or computer at each of the terminals. The program that fools the largest number of judges into believing it is another human wins the prize. The test is limited in that the topics of conversation for each terminal are limited.[20] Subjects in the 1991 competition included women's clothing, small talk, second grade school topics, and Shakespeare's plays. Even with the limited version of the test, the results have been pretty dismal. Below is an excerpt from the "small talk" terminal in the 1991 competition.

00:46:19 Judge 2
Hi how are you? If I met you at a party what would you say to me?

01:06:34 Terminal 2
I'm the head nurse.

01:06:40 Judge 2
Where?

01:06:49 Terminal 2
Let's talk about something else.

01:06:53 Judge 2
Ok what do you want to talk about

01:07:06 Terminal 2
like Paula Abdul, myself.

01:07:08 Judge 2
what?

---

[20]The Loebner Prize competition is controversial because it is so limited. Many people say that by limiting the topic of conversation, the Loebner competition is avoiding the most important feature of the Turing test—the ability to respond to a wide variety of situations.

01:07:17 Terminal 2
What is your favorite song?

01:07:21 Judge 2
Oh I don't know, my favorite movie is Gone with The Wind what about you?

01:07:55 Terminal 2
I can't spend the time standing in line.

01:08:01 Judge 2
I don't understand

I could not confirm from the officials whether this terminal was manned by a human or a computer, but there is no doubt in my mind that this is a very poor attempt by a computer program. This type of performance is surprising after seeing the early success of the Eliza type programs. It is just this type of failure that fuels the AI doomsayers' claims that AI research is getting nowhere.

There are many thorny issues concerning the Turing Test that I will not cover here, such as the need in Turing's sample dialogue above for the lengthy pauses. Of course, if it were a computer generating those responses, it knew the answer long before it printed it, but printing it out that soon would surely give it away. For further reading on the subject, I recommend *The Emperor's New Mind* by Roger Penrose and "The Turing Test: A Coffeehouse Conversation" by Douglas R. Hofstadter.[21]

---

[21]Roger Penrose, *The Emperor's New Mind*, 1990, pp. 5-11; and Douglas R. Hofstadter, "Metamagical Themas: A coffeehouse conversation on the Turing test to determine if a machine can think," *Scientific American*, May 1981, pp. 15-36; reprinted as "The Turing Test: A Coffeehouse Conversation," in *The Mind's I*, pp. 69-95, with a response by Daniel C. Dennett.

Searle's Chinese Room and Strong AI

There are two different theories in AI that have caused much debate. They were first mentioned in John Searle's "Minds, Brains, and Programs."[22] The less controversial of the two views is that of "weak AI," or "cautious AI." "According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. For example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion."[23] Few people would disagree with that. "But according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to *understand* and have other cognitive states. In strong AI, because the programmed computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations."[24] Roger Penrose described it like this: "There is a point of view, referred to as *strong AI* which adopts a rather extreme position on these issues. According to strong AI, not only would the devices just referred to indeed be intelligent and have minds, etc., but mental qualities of a sort can be attributed to the logical functioning of *any* computational device, even the very simplest mechanical ones, such as a thermostat."[25]

---

[22]John R. Searle, "Minds, Brains, and Programs," *The Behavioral and Brain Sciences*, vol. 3, 1980; reprinted in *The Mind's I*, pp. 353-373.

[23]Ibid., p. 353.

[24]Ibid.

[25]Roger Penrose, *The Emperor's New Mind*, pp. 17.

The strong AI view has generated a good deal of controversy. There are really two different parts to it, and not everyone seems to make the distinction. The first part says that computer programs have the qualities of minds and take on cognitive states. The second part says that minds can be modelled accurately by computer programs, and further, that our brains are running programs similar to the way a computer does. To simplify,

1) All programs give rise to minds.
2) All minds are built from programs.

"The idea is that mental activity is simply the carrying out of some well-defined sequence of operations, frequently referred to as an *algorithm*."[26] There is a lot of debate on this issue, but the question is still very much in the philosophical realm.[27] The arguments generally depend on theoretical models, the accuracy of which can't really be proven with our current knowledge of the brain. I don't think we're ready to combine the two views above and assert that minds are equivalent to algorithms. We just don't know enough about the brain or the mind to make any conclusions like that. The two views taken separately, however, are interesting to explore.

I'll start with the second, because it is the one that the strong AI theory stumbles on. If our brains are just running algorithms, that means that everything that goes on inside the brain is formalizable. "Since the Greeks invented logic and geometry, the idea that

---

[26]Ibid., p. 17.

[27]If you explore the development of the scientific method, you'll find that the theory that science explains nature went out of style many years ago. Science may sometimes be used to predict nature, but there is no compelling reason to expect nature to be built on the same foundations upon which our scientific models are built.

all reasoning might be reduced to some kind of calculation—so that all arguments could be settled once and for all—has fascinated most of the Western tradition's rigorous thinkers."[28] Dreyfus lists some thinkers who have pondered this question, starting with Socrates. In 450 B.C., Socrates asked Euthyphro, "I want to know what is characteristic of piety which makes all actions pious... that I may have it to turn to, and to use as a standard whereby to judge your actions and those of other men." He was looking for a universal rule to eliminate the ambiguity in society, just like the physicists are now searching for a Grand Unified Theory.[29]

Plato generalized this wish by advocating that all knowledge must be stateable in explicit definitions which anyone could apply. Anything that couldn't be formalized was not knowledge, but merely belief. Aristotle argued that this principle depended on the intuition of the people making and applying the definitions. Dreyfus claims that Hobbes was the first to directly connect thought with calculation. Leibniz thought he had found a symbolic language for expressing the calculations of thought, and said that with sufficient time and money, he could reduce all thought to the manipulation of numbers. George Boole simplified the process with the introduction of his Boolean algebra, which uses the binary logic system that is the foundation of all modern computers. Like the others mentioned, he didn't set out to build a theory of computers, but rather a formal theory of thought. In 1835, Charles Babbage took strides toward the automation of these

---

[28]Hubert Dreyfus, *What Computers Can't Do: The Limits of Artificial Intelligence*, 1979, p. 67.

[29]Refer to Roger Penrose, *The Emperor's New Mind*, for some fascinating ideas about the Grand Unified Theory of physics. See, for example, p. 151+.

logical systems when he designed machines to do the calculations without human intervention.

While the research on computers continued, Alan Turing was working on his theoretical model of computing. He showed that any digital computer could emulate any other digital computer. So the stage was set—if brains are nothing but digital computers, then computers can do anything that brains can!

It is important to have a grasp on what "digital" means. Alan Turing provides a good description:

> [Discrete state machines] move by sudden jumps or clicks from one quite definite state to another. These states are sufficiently different for the possibility of confusion between them to be ignored. Strictly speaking there are no such machines. Everything really moves continuously. But there are many kinds of machines which can profitably be *thought of* as being discrete state machines. For instance, in considering the switches for a lighting system it is a convenient fiction that each switch must be definitely on or definitely off. There must be intermediate positions, but for most purposes we can forget about them.[30]

All values are really continuous. As the mercury in a thermometer rises, it does not jump from one degree to the next. It flows continuously. Even the currents flowing through the circuits in a computer are continuous.[31] The computer merely rounds off the continuous values to 1 or 0. For example, most modern personal computers use an internal voltage of 5 volts to represent a 1, and 0 volts to represent 0. The values vary in

---

[30]Alan M. Turing, "Computing Machinery and Intelligence," *The Minds I*, p. 11.

[31]The laws of physics state that current is discrete because you can count the number of electrons; there are no fractional electrons. But this depends on the assumption that all electrons are identical, and I'm not convinced of that fact.

practice, so they allow a range, say 3.5-6 volts for a 1 state. As long as the voltages stay in the allowable range, the simulation of the digital states is accurate.

So, does the brain also use a simulation of these discrete states? We know that the chemicals in the brain that define the actions of the neurons work in a continuous fashion, and it is difficult to model these interactions in a discrete manner. So it is not likely that the brain works in exactly the same way as the digital computer, unless you go to the level of quantum physics. That doesn't mean that AI is a fruitless pursuit. Any process can be modeled with an algorithm. The extent to which the algorithm accurately represents the processes varies. Certainly, if the model simulates the process to the extent that the model serves a useful purpose, then the model is worthy of attention. It does not matter whether the brain is a microprocessor running programs; it does matter how accurately a computer model of the mind simulates the mind.

Now we return to the first part of the strong AI theory, the assertion that programs give rise to minds. John Searle's famous "Chinese room experiment" provides a framework to discuss whether a computer can have understanding. Searle refers to programs such as those devised by Roger Schank[32] that attempt to simulate human thought. Schank's program reads a story and then attempts to answer questions about the story that require a human level of understanding to answer. Searle gives this story as an example: "A man went into a restaurant and ordered a hamburger. When the hamburger arrived it was burned to a crisp, and the man stormed out of the restaurant

---

[32]Roger C. Schank and Robert P. Abelson, *Scripts, Plans, Goals and Understanding*, 1977.

angrily, without paying for the hamburger or leaving a tip."[33] After feeding this story to Schank's program, you would present it with questions such as "Did the man eat the hamburger?" The story does not directly state anything about whether the man ate the hamburger, but people who can identify with the situation have no problem answering the question, "No, he did not eat the hamburger." Apparently, Schank's program had some success in answering questions like this, though I suspect its scope is quite limited.

The proponents of strong AI would claim that Schank's program was not only simulating human ability, but that it actually *understands* the story. To refute this claim, Searle describes the Chinese room experiment. Searle proposes that he is locked in a room and given a large batch of Chinese writing and a set of instructions. Searle knows nothing of Chinese whatsoever. To him, Chinese writing is merely "meaningless squiggles." Now, he begins to follow the set of instructions, which direct him to analyze the meaningless squiggles, and mark some additional squiggles on another piece of paper. Now we are told that the instructions actually embody everything needed to read statements and questions in Chinese and to provide written answers in Chinese. If a Chinese-speaking person were to write questions and push them into a window in the Chinese room, and then receive the written responses at that same window, he would have every reason to believe that he was conversing with a Chinese-speaking person in the room! The system of the Chinese room, including the room, Searle, the instructions, and the responses, has passed the Turing test, except that instead of proving it is human, it proved that it can speak Chinese.

---

[33]John R. Searle, "Minds, Brains, and Programs," *The Mind's I*, pp. 354.

According to the strong AI theory, not only does the Chinese room provide intelligible responses to the questions, but it actually understands the Chinese writing. Searle responds incredulously that there can't possibly be any understanding involved. "It seems to me quite obvious in the example that I do not understand a word of the Chinese stories."[34] My response to that argument, and the response of many computer scientists, is that the *system* of the room, Searle, and the instructions is what is doing the understanding. Searle is just a part of the system, and he is not expected to grasp what is going on at the macroscopic level. In your brain, do your neurons understand what you are reading now? No, they are just contributing their small part in the monumental effort of all the systems of your brain to understand. You as a complete *system* understand.

Douglas Hofstadter wrote a fascinating treatise in which he described an anteater that held conversations with an ant colony.[35] The anteater did not converse with the individual ants, but with the colony as a whole, as a system. He draws trails in the ground, and watches how the ant follow the trails. He decodes this information and that is the message, so the pattern of the motion of the ants is a language. The colony and the ants are on two different conceptual levels that do not actively interact. To illustrate this, the anteater freely eats the ants in the colony, even though he considers the colony a personal friend. The anteater claims to be correcting nervous disorders of the colony

---

[34]Ibid., p. 356.

[35]Douglas R. Hofstadter, "Prelude... Ant Fugue," from *Gödel, Escher, Bach: an Eternal Golden Braid*, 1979; reprinted in *The Mind's I*, pp. 149-191.

by removing the ants that are interfering with the communication. Even though many individual ants are perishing, the health of the colony is improved and the colony is very grateful to the anteater.

So Searle is at the wrong conceptual level when he complains that he, a mere part of the Chinese room system, doesn't understand the Chinese writing. Searle anticipated this counter-argument, which he calls the "systems reply." To answer it, he alters the experiment slightly. Let the person in the Chinese room memorize all of the instructions. Searle claims that the person now *is* the system, and he still doesn't understand anything. This is a mistake. The instructions are being implemented in the person's mind, and the understanding being accomplished by these instructions is still at a higher conceptual level than his mind. It doesn't matter where they are located. Donald Perlis calls this higher conceptual level a virtual mind, and argues that the system of the physical mind of the person and the virtual mind of the instructions still understands.[36]

Searle cc ıtinues his argument with a reductio ad absurdum. "If we are to conclude that there must be cognition in me on the grounds that I have a certain sort of input and output and a program in between, then it looks like all sorts of noncognitive subsystems are going to turn out to be cognitive. For example, there is a level of description at which my stomach does information processing, and it instantiates any number of

[36] Patrick Hayes, Stevan Harnad, Donald Perlis, and Ned Block, "Virtual Symposium on Virtual Mind", 1992.

computer programs, but I take it we do not want to say that it has any understanding."[37] Searle has a fatal flaw here and throughout the paper in that he doesn't provide some badly needed definitions. If we take the definition of "understanding" to be "providing the proper response to the provided commands or questions," then I say yes, a properly functioning stomach is understanding its input if it successfully does its part in digesting the food that comes its way. I think Searle assumes that all his readers will share his implicit view that to understand or do anything else normally associated with organic beings, one must be an organic being. I submit, and Donald Perlis concurs, that Searle is begging the question here.

Let me present my own example to illustrate some of the superstitions that Searle is exhibiting. Consider a child who learns long division. He is taught long division step-by-step, much the way a computer is given an algorithm. He may already know a few basics of division, but this new method is so foreign to him that he can't make any connection between long division and the fundamental concept of division that he is already forming in his mind. After a little practice, the child will understand the procedure, but he still won't have a thorough understanding of the *concept* of division. His situation is basically the same as the person inside the Chinese room. After years of practice, however, a different kind of understanding forms. With practice, the procedure is buried into the subconscious. In time, the child forms connections between division and other concepts that he already knows about, such as the fact that division is the inverse of multiplication, and algebraic formulas involving division can be rearranged and

---

[37]John R. Searle, "Minds, Brains, and Programs," *Mind's I*, p. 360.

still mean the same thing. He is learning the fundamental concept of division, whatever that is. I think it is this kind of understanding that Searle refers to.

If we were to teach a computer to do long division, we can easily establish the kind of understanding shown in the first part of this example. When fed the algorithm for long division, the computer understands the procedure, but unless the program goes to great lengths to explain the fundamental concepts involved, the understanding is superficial. If the programmer introduces additional parts to the program that make connections to other concepts, we can say that the computer's understanding is increasing. After all, even though a teacher guides a child's learning, we still say it is the child whose understanding is benefitting. The relationship of the programmer to the computer is very different, but I think the characteristics of learning are still present in this example. To avoid this complication, we can make the computer teach itself. There already exists many types of programs that give computers the power of self-evaluation. Computers can rearrange their utilization of memory and disk storage, and can prioritize a list of tasks that they need to complete. They can rearrange formulas so that they are calculated more efficiently. I don't know if the computers will be able to achieve the same level of understanding as the child, but I see no reason to discount this possibility, at least in the future.

The argument is really that computers are merely a *simulation* of some real feature. Hofstadter has a good argument against this.[38] Consider: what is the difference between the "real" calculations of a schoolkid and the "simulated" calculations produced

---

[38]Douglas R. Hofstadter, "A Coffeehouse Conversation," *The Mind's I*, p. 75.

by a computer? Compare a proof constructed by a mathematician to a proof generated by a computer.[39] There is no appreciable difference. Now if we program a simulation of something grounded in the physical world, like a hurricane, or a cow, then we can easily see that the hurricane will not make us wet, and we would never get anything but simulated milk from the cow. But when we talk about abstract things such as mathematics, thought, and consciousness, then a computer "simulation" is as good as the real thing.

Learning

Many consider computer learning to be the key to artificial intelligence. We can stuff facts into a computer until we turn blue in the face, but to build an entity flexible enough to deal with the varied experiences of society, we must teach it how to learn. Hubert Dreyfus covers this subject at length.

> Since a human being uses and understands sentences in familiar situations, the only way to make a computer that can understand actual utterances and translate a natural language may well be, as Turing suspected, to program it to learn about the world. Bar-Hillel remarks, "I do not believe that machines whose programs do not enable them to learn, in a sophisticated sense of the word, will ever be able to consistently produce high-quality translations."[40]

The lack of the ability to learn in a broad sense is a limiting factor in all computer software. AI research has produced some very useful programs, but none of them are general enough to be of any use outside of their limited domain. We have chess

---

[39]Daniel C. Dennett, "Reflections" to the above, p. 94-95.

[40]*What Computer's Can't Do*, p. 109. The quote by Bar-Hillel is from Yehoshua Bar-Hillel, "The Present Status of Automatic Translation of Languages," *Advances in Computers*, F. L. Alt, ed., 1960, pp. 105-106.

machines, expert systems in medicine, and handwriting recognizers, but by themselves, they don't have near the common sense that two year old children do. This is considered by many to be the major shortfall of AI—that the programs are not general enough. We need a program that encompasses all facets of what we call common sense.

Marvin Minsky, in *The Society of Mind*, presents a model of the mind that consists of many small parts that interact. Research of the brain has shown the brain is subdivided into many small parts, each of which performs a particular function. Our common sense, according to Minsky's model, is just built of these elementary pieces. Thus, we could take the myriad of AI applications that we have today and meld them into one cohesive unit. This may be a good simulation of the way the brain subdivides its structures, but we won't magically have learning in the combined system unless one of the individual elements approaches this issue. The first problem is that we still wouldn't have a generalized learning program, rather, just a collection of parts that do their own job. We need to transcend the level of teaching (programming) computers about all the minute details. Instead, we should be able to teach them how to learn, then they figure out how to do things on their own. The second problem I have with this model is the fact that there are few interconnections. Researchers have discovered that the brain is remarkably adaptable. We can disable parts of the brain, only to find that other parts of the brain take over for the function that is missing. This requires a large amount of redundancy in the structure of the brain, or an incredible amount of interconnectivity. I think that the various parts of the brain are highly connected to the other parts of the brain. This allows for the great flexibility, and makes it easy for one part of the brain to

take over for another part, when the two may not even be in close proximity to each other.

There is a lot of effort going into computer learning at the present. Doug Lenat, for instance, is spearheading the Cyc project, which aims to stuff millions of everyday facts into a computer system with the hopes that with enough of these facts, it will gain common sense. There is some merit to this kind of brute force method, but it remains to be seen how advanced the computer learning technology will become in the next few years.

The Cyc project illustrates an interesting point. The bits of knowledge that are being fed into the system are the types of facts that people learn every day as they are growing up. Adults have had years to pick up the basics of how gravity works, how to use their hands to grab things, etc. It is really unfair to expect a brand new computer system to know these things without taking the time to learn them, or have them preprogrammed as instinctive responses. We couldn't begin to list all the basic facts that we have learned since we were born. So maybe instead of trying to build robots with full-fledged adult intelligence, we should build machines with the capability to learn and nothing more built in. We can then give them the task of learning on their own. The older the robots get, the more knowledgeable they will become. After the first few attempts, when we have invested the time in coaching a few robots into maturity, we could speed up the learning process in the next generation. These experienced robots already have a wealth of information stored in their brains, so we could just download that information into the next generation of robots.

We would have to be careful that the downloaded memory is kept separate from the new robot's memory, though. For example, maybe the older robot had mechanical problems with one of its legs. It would have developed an elaborate scheme to deal with its disability. But if the new robots with their downloaded intelligences didn't distinguish this information from their own experience, we would have a whole batch of brand new, limping robots!

Hubert Dreyfus, in the summary to his book, addresses the "learning from childhood" issue:

> Could we then program computers to behave like children and bootstrap their way to intelligence? This question takes us beyond present psychological understanding and present computer techniques. In this book I have only been concerned to argue that the current attempt to program computers with fully-formed Athene-like intelligence runs into empirical difficulties and fundamental conceptual inconsistencies.[41]

I agree that we can't expect the current systems to have the breadth of experience that any adult human does. But I think we shouldn't dismiss so easily the learning capability of computers. They can do it, we just have to teach them how.

The last word on anthropomorphisms

Most of the arguments about anthropomorphic issues can be reduced to arguments about the meanings of words. Geoff Simons recognized that many of the words in our vocabulary are on the verge of change:

> We are now beginning to re-examine many of the conventional adjectives traditionally applied to human beings—such words as *conscious, intelligent, thinking, perceptive, free,* and *aware*. We are having to scrutinize such terms

---

[41]Hubert Dreyfus, *What Computers Can't Do*, p. 290.

> because we are seeing that increasingly they can denote characteristics of artificial systems.[42]

Turing predicted that by the turn of the century, people would be so used to using anthropomorphisms with computers that they wouldn't think twice about it. Norbert Wiener, the founder of cybernetics, explained that words such as "life," "purpose," and "soul" are not suited to scientific thinking because they are not rigorously defined. He goes on to discuss the evolution of such words:

> Whenever we find a new phenomenon which partakes to some degree of the nature of those which we have already termed "living phenomena," but does not conform to all the associated aspects which define the term "life," we are faced with the problem whether to enlarge the word "life" so as to include them, or to define it in a more restrictive way so as to exclude them. We have encountered this problem in the past in considering viruses, which show some of the tendencies of life—to persist, to multiply, and to organize—but do not express these tendencies in a fully-developed form.[43]

I have only discussed a few of the anthropomorphic issues that I listed earlier, but I hope that you can see how all of them, in some way or another, are going to change to accommodate the emerging artificial intelligences.

Science fiction

To really stretch our minds, we have to look beyond the current technology, beyond even the current theories, and turn to science fiction. Many, if not all of the technological innovations that we enjoy now were predicted by the science fiction literature. Marshall McLuhan recognized the value of the arts:

---

[42]Geoff Simons, *Are Computers Alive?*, p. 28.

[43]Norbert Wiener, *The Human Use of Human Beings*, p. 32.

> The power of the arts to anticipate future social and technological developments, by a generation and more, has long been recognized. In this century Ezra Pound called the artist "the antennae of the race." Art as radar acts as "an early alarm system," as it were, enabling us to discover social and psychic targets in lots of time to prepare to cope with them.[44]

Marvin Minsky reflects the views of many computer sciences in his thoughts on science fiction authors in this quote in *The Media Lab*:

> "Well, I think of them as thinkers. They try to figure out the consequences and implications of things in as thoughtful a way as possible. A couple of hundred years from now, maybe Isaac Asimov and Fred Pohl will be considered the important philosophers of the twentieth century, and the professional philosophers will almost all be forgotten, because they're just shallow and wrong, and their ideas aren't very powerful."[45]

And finally, some comments by Stewart Brand, from a section titled "The Important Philosophers of the Twentieth Century:"

> Somewhere in my education I was misled to believe that science fiction and science fact must be kept rigorously separate. In practice they are practically one intellectual activity, although the results are published differently, one kind of journal for careful scientific reporting, another kind for wicked speculation.[46]

I am a fan of science fiction, and that is part of the reason I am dedicating the space for it here. But science fiction can provide new insights for anyone who invests in the effort to read it, whether a fan or not. Several of the items in the list of

---

[44]Marshall McLuhan, *Understanding Media: The Extensions of Man*, 1964, p. X (in the introduction to the paperback edition).

[45]Marvin Minsky, quoted by Steward Brand, *The Media Lab: Inventing the Future at MIT*, p. 224.

[46]Stewart Brand, *The Media Lab*, p. 225.

anthropomorphisms I presented were inspired by stories in *I, Robot*, a collection of stories written by Isaac Asimov in the 1940's. I'll elaborate on some of them below.

Can a computer make someone love it?

In "Robbie," Asimov tells us of a robot nursemaid that watches after a young girl. The story is set in a time when autonomous robots were a new innovation, and the public was very wary of them. The girl becomes very attached to her robotic playmate, but her mother decides that the robot is too dangerous to trust her daughter to anymore. The girl is devastated when the robot is taken away. Emotional attachment to a machine is also illustrated in "The Soul of the Mark III Beast," by Terrel Miedaner.[47] In this story, a man is shown a small bug-shaped robot, and is asked to "kill" it with a hammer. The man has misgivings when the robot puts on a pitiful death scene. These two examples show how people will feel compassion for things that imitate living creatures in their behavior.

Can a computer make judgements?

Several of the stories in *I, Robot* show examples of a machine making judgements. in "Reason," two test engineers test a prototype of the first robot able to reason for itself. The robot reasons that the humans, being inferior life forms, after all, could not possibly have made him. Knowing that he was more capable of running the base they

---

[47]Terrel Miedaner, " The Soul of the Mark III Beast," published by the Church of Physical Theology, Ltd., 1977. Reprinted in *The Mind's I*, p. 109-113.

were located at, the robot takes over the base and locks up the engineers to protect them from their own incompetence!

In "Evidence," a candidate for public office is accused of being a robot. Robots were banned from the Earth because of the fear of their power, so having a robot as a public servant was unacceptable. Neither the candidate not his opponent proves his humanity or lack thereof, and he is elected. This idea is continued in "The Evitable Conflict," where the Earth's entire social and economic system is managed by a network of computers. When strange things happen, no one is sure whether the computers are failing or doing it for their own good.

Machine judgement is likely to become more of an issue in the near future. A good example is the automatic collision detection systems that are being tested for automobiles, and the automatic pilot systems that will surely follow. People will trust their lives to these computers. Are you ready to entrust your life to a computer and the engineers who built it? Computer glitches will no longer mean just a loss of data.

Can a computer experience stress, or have a nervous breakdown?

In "Catch that Rabbit," a robot supervisor experiences on-the-job stress that effects his performance. In the story, the problem is caused by a glitch in the hardware, but computer stress could just as easily be caused by conditions that push its capabilities to the limit. In a Novell network at the University of North Texas, false error messages appear when too many people are using the fileserver. Not only can it not handle the stress, it can't even properly report the problem.

"Liar!" and "Escape" illustrate what could be called robot nervous breakdowns when they encounter "dilemmas" in their programming. It has been shown that human logic involves a myriad of conflicting information, and that thinking computers would also have to deal with conflict in its logic. The stories introduce the use of a *robopsychologist*, who understands the complicated structures in the robots' brains and can help pinpoint problems. As computers continue to get more complex in their design, their behavior will also become more complex. We may very well have to resort to an inexact science like psychology when computers become so complex that we can't explain their behavior in simple terms.

Robots' rights

With all of the discussion about how robots could be like humans, it is appropriate to wonder whether robots should have the same kind of rights that humans enjoy. The best example is Lieutenant-Commander Data on the television series *Star Trek: The Next Generation*. On the show, Data is an android who has joined Starfleet and is serving on the Enterprise as a regular member of the crew. In one episode, a review board confirmed that because Data is a sentient being, he deserves the same rights that humans have. Many educated people who scoff at the idea of granting rights to today's computer systems have no problem agreeing that Data should have rights. Well, the evolution of today's computer systems may someday bring them in line with the same type of technology that makes an android like Data possible. As that happens, this discontinuity

in people's thinking will become apparent to them, and they will have to deal with these issue head-on.

In researching this idea, I tried to discuss the philosophical implications with the Star Trek fans, but they were more interested in the details of the plot of the show. I tried to discuss it with thinkers in the computer science field, and they weren't willing to talk about science fiction. We need more open-minded thinkers in this field, because science fiction is becoming science fact.

Geoff Simons argues passionately for robot rights as he compares the oppression of blacks to the oppression that robots may face.[48] He says that we will limit robots' education just like we did for blacks and workers "so they will not get 'ideas above their station.' " And in the case of Commander Data, people agree that we can't take away the intrinsic rights of a sentient being.

Pain

Consider a less complicated issue that involved the same sort of complications: pain and emotions. If a computer can think, then maybe it can experience pain as well.

> "One idea is that since computer configurations resemble traditional biological neural nets, and since pain resides in nervous systems, it is possible that computers have a potential for pain of which, as yet, we have little conception."[49]

Pain exists to alert people that they are being harmed, or that something internal is not functioning correctly. Computers can detect when they are having problems with

---

[48]Geoff Simons, *Are Computers Alive?*, pp. 160-161.

[49]Geoff Simons, *Are Computers Alive?*, p. 159.

their hardware, and we can associate this condition with pain. When a person's arm is amputated, he may have phantom feelings, as if the arm were still there. Similarly, when we disconnect a computer peripheral, the software to control that peripheral still resides in the computer and it may attempt to use it. Are we committing homicide when we unplug a computer?

There is something fundamentally wrong with this type of thinking. Is our goal to play God; to build living creatures and set them free to interact in society? No. We are in the business of building tools. At the present, these tools help us to calculate our taxes, build our cars, and make our decisions. Do we want our tools to limit ourselves? We don't worry about hurting the feelings of a hammer when we drive a nail, and we shouldn't have to worry about the feelings of a computer when we use it to suit our needs.

Privacy

Geoff Simons discusses that computer systems can become so complicated that we don't understand how they make their decisions (like declaring war). For safety's sake, he suggests that a "human window" should be built into all computer systems to allow people to question a computer on why it reached a conclusion. But if the conclusion were part of the computer's private life, then "clearly such a 'human window' would be a manifest invasion of privacy!"[50] Hogwash! The only issue of privacy would be related to human issues—e.g. national security, trade secrets, salary levels, etc.

---

[50]Geoff Simons, *Are Computers Alive?*, p. 165.

Weizenbaum and self-degradation

As the computer programs that are being produced are becoming more and more humanlike in their abilities, the public is becoming more willing to grant that computers can think. Joseph Weizenbaum thinks that this will create a major crisis. He cites examples from the past, starting with Galileo, who showed that the Earth is not the center of the universe. Darwin opened our minds to the possibility that humans are just another kind of animal. Freud concluded that rationality is an illusion. And now the AI gurus are showing us that our brains are just meat machines. The human brain is the only evidence we have that humans are anything superior to monkeys and rocks and adding machines, unless we include supernatural characteristics such as the soul. If machines can accomplish anything that our brains can, then we lose that last tangible bastion of superiority. Weizenbaum says that by making these humiliating new theories a part of our lives, we are creating a crisis in the mental life of our civilization.[51]

Weizenbaum implies that we should abandon certain areas of computer research because of this. However, his own evidence shows that human curiosity marches relentlessly onward. Even though every new discovery adds to the hopeless feeling that mankind is an insignificant speck in the universe, the pursuit of science is not going to grind to a halt because Mr. Weizenbaum says it will hurt our feelings. We are still using the ideas of Galileo and Darwin and company, and millions of Christians willingly put

---

[51]Joseph Weizenbaum, "On the Impact of the Computer on Society: How does one insult a machine?" *Science*, vol. 176, May 12, 1972, pp. 609-614.

themselves into submission to a higher power in order to explain things they otherwise wouldn't be able to understand.

I am not arguing that Weizenbaum's ideas are bad, it is just unfeasible to expect society to suppress a curiosity that is built into our very genes. In the conclusion to his 1976 book, Weizenbaum presents two kinds of computer applications that should be completely avoided, or at least approached with utmost caution.[52] The first category is programs that attempt to substitute a computer system for a human function involving interpersonal respect, understanding, and love. Therefore, he says, it would be both technically infeasible and immoral to use the any of the computerized psychotherapist programs on real-life subjects, because they just wouldn't be able to relate to a human's problems. A computer couldn't begin to understand what it's like to dwell in a human body or interact with the environment in a human way. Therefore, he says, the computer wouldn't have any empathy for its subjects and wouldn't be able to express compassion.

The second category includes applications that could have irreversible and unknown side effects, and don't meet a pressing human need. The best example from this category is automatic speech recognition systems. There has been considerable progress in this field, but the successes are all isolated to a particular application. For example, a computer can be trained to understand a decent-sized list of words from a single speaker. But different people pronounce words differently enough to confuse the programs. Translating from speech to text is difficult, but actually comprehending the

---

[52]Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation*, 1976, pp. 268-269.

words is even more difficult. We will probably never have a computer that can participate in a human conversion and understand perfectly what is being said. After all, we can't expect that from a human.

Weizenbaum sums it up like this: "What emerges as the most elementary insight is that, since we do not now have any ways of making computers wise, we ought not now to give computers tasks that demand wisdom."[53] Here I disagree. He shows that we should be careful when we describe the capabilities of a computer system. But computers can and will encroach upon the area of human reasoning. Let me address another of Weizenbaum's points:

> The structure of the typical essay on "the impact of computers on society" is as follows: First there is an "on the one hand" statement. It tells all the good things computers have already done for society and often even attempts to argue that the social order would already have collapsed were it not for the "computer revolution." This is usually followed by an "on the other hand" caution which tells of certain problems the introduction of computers brings in it wake. The threat posed to individual privacy by large data banks and the danger of large-scale unemployment induces by industrial automation are usually mentioned. Finally the glorious present and prospective achievements of the computer are applauded, while the dangers alluded to in the second part are shown to be capable of being alleviated by sophisticated technological fixes. The closing paragraph consists of a plea for generous societal support for more and more large-scale computer research and development.[54]

Since this was written in 1972, the hype about computers, and especially AI, has increased at a rate far exceeding the actual development of the technology. No one listened to Weizenbaum's forebodings. There are many very real risks we are facing, but

---

[53]Ibid., p. 227.

[54]Joseph Weizenbaum, "On the Impact of the Computer on Society: How does one insult a machine?" *Science*, vol. 176, May 12, 1972, p. 609.

the scientists are not going to address them. After all, science is good; it helps to satisfy our curiosity. The safeguards must come from somewhere else. In the next section, I'll explore a few of the risks, and explain what might be society's safety net.

Risks involving machines

Ever since the first time a caveman smashed his thumb with a stone axe, man has known that machines pose risks. Nowadays, society is acutely aware of the fact that machines can be dangerous. Peter Neumann writes a monthly column in *Communications of the ACM* entitled "Inside Risks" that serves as a watchdog for computer risks, and he moderates an electronic forum where anyone can submit a risk that they have witnessed or heard about. The risks are very real.

The Machine Stops

E. M. Forster gives us a foreboding picture in his 1920-era story, "The Machine Stops."[55] Many people aren't familiar with E. M. Forster's science fiction stories, but this one ranks with Orwell's *1984*, and it came decades earlier.

In the story, technology has advanced to the point that the people have every creature comfort they could possibly want, and they are satisfied with the current level of technology. Years earlier, the Machine was built, and it maintained itself so well that all knowledge of its internal workings was lost. Each person lived in a separate but identical underground cubicle. They could communicate via "plates" to any other

---

[55]E.M. Forster, "The Machine Stops," *The Eternal Moment and other stories*, 1928.

cubicle in the world. They spent virtually all of their time in their cubicles, because the Machine provided for all of their needs. The people were isolated by technology, and they grew to dislike physical contact; they were offended when anyone touched them.

The people learned to trust the judgement of the Machine. Reproduction was an emotionless task that they performed when called upon by the Machine. The children were removed from their mother at birth and raised in a nursery. They would rarely have any kind of contact with their relatives ever again. When their bodies failed, they would submit a request to the Machine for euthanasia.

People were completely free from the "overhead" of life. They just sat in their cubicles and tried to think of ideas. If something didn't give them ideas, such as viewing a beautiful landscape, it was not worth spending time with. Acquaintances would pass messages to each other through the Machine, and several times a day they would participate in lectures through their plates.[56]

The people in the story lost touch with nature. Their lungs became weakened by the processed air fed to them by the Machine, so they had to don special equipment to go to the surface of the earth. At the end of the story, the Machine began to break down, and nobody knew how to fix it. Nobody would even believe that it could break. Eventually, though, it did cease to function. Without their Machine, upon which they depended so totally, everyone died.

---

[56]The communications system that Forster described bears a remarkable resemblance to the modern computer networks. After reading the story, I realized that much of the behavior that I exhibit when communicating on a network is accurately described by Forster.

I consider this a prophetic story, especially given that computers were only a fantasy at the time the story was written. Now computers are far more than a fantasy, and Forster's story could easily come true.

Machines aren't smart enough to know not to hurt us

Joseph Weizenbaum gives a vivid example of the fear of machines:

> We set a punch press in motion, and it mangles the hand of a worker who gets too close to it. The very regularity of the machine is its most fearsome property. We put it to the task and it performs, regularly to be sure, but blindly as well.[57]

Fear of this sort is healthy, because it encourages us to think about how much trust we put into machines. A machine, or a computer, will feel no regret if it accidentally kills someone. Isaac Asimov is famous for devising his Three Laws of Robotics, which read as follows:

> #1 A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
> #2 A robot must obey orders given 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 Law.[58]

This is really an attempt to codify human ethics. Of course, if robots are going to interact with humans, they should adhere to our ethics, just like any other foreigner who joined our society would be expected to.

---

[57]Joseph Weizenbaum, *Computer Power and Human Reason*, p. 40.

[58]Isaac Asimov, *I, Robot*, p. 1.

I'm afraid of Isaac Asimov's faith in the First Law of robotics. There's a line in *I, Robot* that goes "You *know* that it is impossible for a robot to harm a human being; that long before enough can go wrong to alter that First Law, a robot would be completely inoperable. It's a mathematical impossibility."[59] This concept itself is *physically* impossible. Hardware is always at a lower operational level than the software that enforces the First Law. There is no way to guarantee that the software won't get confused about the operation of the robot's limbs and knock its powerful arm right through someone. And you can't implement something in hardware that is as complicated as human ethics.

Asimov's laws are just another example of where we need more fear and less trust. If Asimov's robots were present in today's society, people would automatically put too much trust in the Laws of Robotics, first because they are "laws," and second, because people tend to put a great amount of trust in computers and technology in general. A simplified example of this trust: the computer calculates a long column of numbers, you're not going to recalculate it by hand to check it—that's why you did it on the computer in the first place. So you accept it all without question. Usually, everything will go well, but when something goes wrong, the results could be catastrophic.

Am I contradicting myself?

You may be wondering why I would first claim that computers are wonderful and will develop without bound, and then cite foreboding stories about them. The reason we

---

[59]Isaac Asimov, "Robbie," *I, Robot*, p. 17.

should all have a little healthy fear is *because* of the enormous power of computers. Computers will someday have the power to take over many of the reasoning functions that we now must do for ourselves. They will act like us, and because of that, we will be inclined to treat them like ourselves and to trust them in the same way we would trust another human being. But computers are not like people in many ways, at least not yet. While we can trust a computer more than a human doing repetitive mathematics, there are areas where humans still excel. If we are to put our trust in computers, we must be sure that they are at least minimally competent in the area in which we place our trust.

So is mankind doomed to be subservient to machines? Not according to John Naisbitt's theory. In *Megatrends*, he introduces his "High tech/high touch" theory:

> High tech/high touch is a formula I use to describe the way we have responded to technology. What happens is that whenever new technology is introduced into society, there must be a counterbalancing human response—that is, *high touch*—or the technology is rejected. The more high tech, the more high touch.[60]

Naisbitt presents several examples of high tech/high touch pairs. One is the introduction of high tech word processors, which he claims has led to a revival of handwritten notes and letters. He also pairs the growth of high technology in the hospitals with the growth of the home hospice movement and the general interest in the quality of death. I like this theory. It shows that we have checks and balances to counteract the potentially negative effects of technology. "The appropriate response to

---

[60]Ibid., p. 39.

more technology is not to stop it, Luddite-like, but to accommodate it, respond to it, and shape it."[61]

An Attempt to Tie it All Together

I have been very optimistic about the future growth in AI research. Let me clarify this optimism, because there are many valid arguments that show how AI progress is stagnating. I do not believe I will see an android like Star Trek's Commander Data in my lifetime. I don't know if such a thing will be technically feasible within the next 200 years. But what about the next 1000 or 10,000 years? When we look at predictions of the future, the farther they go back, the more incorrect they tend to be. Even many of the predictions made 100 years ago were so wrong that they are humorous. When someone says "AI will never rival the human mind," they are taking a big responsibility by saying "never." Many people believe that in 10,000 years the world will appear nothing like the way we now know it, if it will exist at all. I won't venture any predictions about indefinite times in the future.

An observation that I can make, however, is that human curiosity is an unstoppable force. As long as humans roam the earth, they will want to know why things work the way they do, and they will want to build a better mousetrap. So however slowly AI research will proceed, in some form or fashion, it will proceed, and it will continue to produce beneficial results. These results may be continue to be isolated in their respective fields for quite some time.

---

[61]John Naisbitt, *Megatrends*, p. 41.

Marvin Minsky has summarized the progress in AI. He cited examples of projects that were previously thought to be impossible, but with the increased processing power of the newer computers, were possible. Minsky's report was in 1966.[62] It showed that many of the limitations in AI research were caused by a simple need for greater computing power, not a problem in the fundamental nature of the approaches. In 1973, Ted Nelson's *Dream Machines* illustrated some pioneering new techniques in user interfaces. Now that we finally have the processing power to handle it, graphical intuitive interfaces are commonplace. Though the problems in AI are becoming more difficult, the available computing power is growing at an increasing rate. I think these factors will tend to balance each other, so that the research will be able to march on steadily.

Will we someday be caught in Forster's horrific society of the Machine? Some of the aspects seen in "The Machines Stops" will be present, in fact, many of them have already appeared. But hopefully, the checks and balances of high tech/high touch will allow us to cling to our humanity in a world where the machines are getting faster and the programs are getting smarter.

---

[62]Marvin Minsky, "Artificial Intelligence," *Scientific American*, Vol 215, September 1966, pp. 246-252+.

BIBLIOGRAPHY

Bruce W. Arden, ed., *What Can Be Automated? The Computer Science and Engineering Study*, MIT Press, 1980.

Isaac Asimov, *I, Robot*, Ballantine Books, 1950. Citations are from the Doubleday & Company, Inc. edition, 1987. The stories in this book were originally published between 1940-1950.

Stewart Brand, *The Media Lab: Inventing the Future at MIT*, Viking Penguin, Inc., 1987.

"Computer Made to Think," *Science Newsletter*, vol. 81, January 6, 1962, p. 5.

Daniel C. Dennett and Douglas R. Hofstadter, editors, *The Mind's I: Fantasies and Reflections on Self and Soul*, Basic Books, Inc., 1981. Citations are from the Bantam edition, 1992.

Peter J. Denning, "The Science of Computing—What is Computer Science?" *American Scientist*, vol. 73, Jan-Feb 1985, pp. 16-19.

Hubert L. Dreyfus, *What Computers Can't Do: The Limits of Artificial Intelligence*, 1972, Harper & Row revised edition 1979.

E. M. Forster, "The Machine Stops," *The Eternal Moment and other stories*, Harcourt Brace Jovanovich, Inc., 1928, pp. 3-37.

Patrick Hayes, Stevan Harnad, Donald Perlis, and Ned Block, "Virtual Symposium on Virtual Mind," 1992, citations are from a draft manuscript. To be published in *Minds and Machines*, vol. 2, no. 3, pp. 217-238.

Douglas R. Hofstadter, *Gödel, Escher, Bach: an Eternal Golden Braid*, 1979.

Douglas R. Hofstadter, "Metamagical Themas: A coffeehouse conversation of the Turing test to determine if a machine can think," *Scientific American*, May 1981, pp. 15-36. Reprinted as "The Turing Test: A Coffeehouse Conversation," in *The Mind's I*, pp. 69-95, followed by a response by Daniel C. Dennett.

Steven Levy, "The Riddle of Artificial Life," *Popular Science*, vol. 241, October 1992, p. 100.

Marshall McLuhan, *Understanding Media, The Extensions of Man*, McGraw-Hill Book Company, 1964. Citations are from the 1965 paperback.

Marvin L. Minsky, "Artificial Intelligence," *Scientific American*, vol. 215, September 1966, pp. 246-252+.

Marvin L. Minsky, *The Society of Mind*, Simon & Schuster, 1985, 1986.

John Naisbitt, *Megatrends: Ten New Directions Transforming Our Lives*, Warner Books, Inc., 1982.

"100-Ton Brain," *Senior Scholastic*, vol. 48, February 4, 1946, p. 36.

Roger Penrose, *The Emperor's New Mind: Concerning computers, minds, and the laws of physics*, Oxford University Press, 1989, revision 1990.

"Robot Einstein," *Newsweek*, vol. 26, November 12, 1945, p. 93.

Roger C. Schank and Robert P. Abelson, *Scripts, Plans, Goals, and Understanding*, Erlbaum, 1977.

John R. Searle, "Minds, Brains, and Programs," *The Behavioral and Brain Sciences*, vol. 3, Cambridge University Press, 1980. Reprinted in *The Mind's I*, pp. 353-373.

Geoff Simons, *Are Computers Alive? Evolution and new life forms*, The Therford Press, Ltd., 1983.

Alan M. Turing, "Computing Machinery and Intelligence," *Mind*, vol. LIX, no. 236, 1950. Reprinted in *The Mind's I*, pp. 53-67.

Joseph Weizenbaum, *Computer Power and Human Reason: From Judgement to Calculation*, W. H. Freeman and Company, 1976.

Joseph Weizenbaum, "On the Impact of the Computer on Society: How does one insult a machine?" *Science*, vol. 176, May 12, 1972, pp. 609-614.