Will Robots Destroy Us?: Teaching Students About Technological Implications

Thomas P. Sturm
Quantitative Methods and Computer Science Department
University of St. Thomas
tpsturm@stthomas.edu

Abstract

This paper focuses on the concerns about the future of robotic technology as articulated by Bill Joy, chief scientist and founder of Sun Microsystems. The seminal article for this discussion was "Why the Future Doesn't Need Us" which appeared in Wired Magazine. (http://www.wired.com/wired/archive/8.04/joy.html) [1]. The article's footnotes provide an excellent set of references for this discussion.

This paper discusses the concepts of futurism as articulated by Joel Barker, in which he distinguishes between process futurism and content futurism. It also discusses Moore's law of computer technology growth, in which technological capabilities double every two years. It discusses the robotics history, research, and forecasts of Hans Moravec, in which he asserts a drastic compression of the two-year doubling time in the future.

Within this context, the consequences of unbridled technological development in robotics are discussed. Possible futures, all of which postulate that the robot will become far superior to the human, are discussed. Mankind's golden age, human domination by robots, and the destruction of all carbon-based life forms are among the possible scenarios. A framework for the evaluation of various technological forecasts is also discussed.

Observations are made about the reactions of students to this material, depending upon their background and the way the material was presented. This reaction ranged from elation in a robotics course to depression in a departmental seminar with a more general audience. An important discussion topic to be prepared for is how we can effect a change in the future.

Observations are also made about how the same type of analysis can be applied to other computer technologies, such as data mining. The futures related to data mining in retailing, for example, can include diverse scenarios from as positive as custom merchandise that suits each individual to as sinister as situational pricing for all retail goods similar to airline ticket pricing.

Introduction

In April of 2000 one of my students in my Artificial Intelligence and Robotics course brought an article by Bill Joy, chief scientist and founder of Sun Microsystems, to my attention [1]. I accessed the article, “Why the Future Doesn’t Need Us,” via the internet at http://www.wired.com/wired/archive/8.04/joy.html. What struck me about the article was that I had read or was familiar with most of the references cited in the article. What also struck me was that none of my students were likely to have read many of these works, at least prior to approaching the article. I asked my students to read the article as an assignment in preparation for a classroom discussion. What struck me during the subsequent discussion was that the students focused on the tremendous potential power of robots, but not the consequences of that power. This future was “going to be cool.” They would enthusiastically embrace this kind of development.

It was after this discussion that Bill Joy’s words really hit home: “Failing to understand the consequences of our inventions while we are in the rapture of discovery and innovation seems to be a common fault of scientists and technologists.” I resolved that I had to do a better job than just “have a discussion” about the article in class. I quickly organized a departmental colloquium, inviting not only students and colleagues in the department, but philosophers, theologians, and the university community in general. I needed to find out if I could get across to computer science students my point about the consequences of technology.

The First Seminar

The announcement for the seminar emphasized that the talk would be based on the article by Bill Joy and cited the web URL. Naturally, I wanted everyone in the audience to pre-read the article. At the same time, I couldn’t presume that everyone would have taken the time to read it in its entirely. Furthermore, my prior experience with my computer science majors taking my robotics course suggested that even if they all had read it, they would not experience the full impact of the article. My approach was to interactively lead the audience through the major points of the article.

Involving and Hooking the Audience

The first thing I wanted to do was involve the audience in the presentation. I also wanted to reward those who had prepared most for the colloquium. I decided that each time I mentioned any reference, I would ask how many people had read it. I would then interact with those people about their opinions of the work. This would precede my making the points I needed to using information from that source.

I began by asking how many people had read Bill Joy’s article. I then asked them to keep their hands raised for a while so that everyone could see roughly how many hands were raised. I then asked how many people, in reading the article, felt that they were coming into the middle of a conversation. Virtually everyone kept their hand raised. The point I made from that was that to fully comprehend all of what Bill Joy was trying to get across, you had to bring to the table quite a bit of background information. At least the first part of the talk was designed to fill them in.

Discussion of Futurism

Since the article's primary focus was the forecasting of events some 20 years into the future, I began with a discussion of how futurists make predictions. I presented the concepts of futurism as articulated by Joel Barker [2]. Joel distinguishes between process futurism and content futurism.

Content futurists specialize in an area of information about the future. They speculate on the "whats" of the future. They predict the future reliably if (and usually only if) the rules do not change. Process futurists, by contrast, deal with how to think about the "what." They teach us how to manipulate what is produced by the content futurists. They also predict rule changes or paradigm shifts. My intent was to show which of the futuristic predictions were the result of content futurism, and which were the result of process futurism.

Moore's Law

Any discussion of the future of computers requires a discussion of Moore's Law. Moore's Law is named after Gordon Moore, co-founder of chip maker Intel. Moore's Law states that the doubling time of computing power is every two years. The problem I have always had with Moore's Law is that I find it hard to impress upon people, particularly students, the tremendous impact of that law in the long term.

To make my point, I first used four graphs out of Hans Moravec's book *Robot* [3]. The first graph shows how many megabytes of storage and how many millions of instructions per second are needed in a variety of objects, including manual calculation, viral DNA, books, CDs, audio and video channels, various computers, and various living organisms including monkeys and humans. The second graph shows the computing power of various computers measured in MIPS per 1998 kilobuck over the time period from 1890 to date and forecasted through 2030. Of course the MIPS per 1998 kilobuck scale is logarithmic, thus the data points form a straight or slightly curved upward line. The third graph shows the MIPS of computing power available to AI and robotics programs from 1950 to date, again showing MIPS on a logarithmic scale. The fourth graph shows the chess machine performance as measured by chess rating and the computational abilities of various chess playing systems. While these graphs are impressive, I still feared that the point of Moore's Law might be lost on some, and the rest of Bill Joy's argument would be lost as a consequence.

To supplement the graphs, I presented the information in Table 1.

Table 1: Moore's Law

Time into the future	Capability with doubling every 2 years
1 year	1.414 times current capability
2 years	2 times current capability
10 years	32 times current capability
20 years	1,000 times current capability
30 years	32,000 times capability
40 years	1,000,000 times capability
50 years	32,000,000 times current capability

I then pointed out that this was a 120 year trend, so that over the 120 years we have developed 1,000,000,000,000,000,000 times the computing capability Herman Hollerith started out with in 1880. I also pointed out that these forecasts are the result of *content* futurism. So, if the rules don't change, this is a very likely scenario of the future.

I used the following example as a data point for the application of these doubling principles. In 1965 (36 years ago), main computer memory cost $1 per bit in 1965 dollars. Today one can purchase a 128 MB stick of memory for $50. In 1965, could you have been able to build a memory with 128 MB (over 1,073,700,000 bits), it would have cost over $1 billion dollars at a time when union journeyman carpenters were making $4.34 per hour (salary and fringe benefits). Thus a carpenter would have to work over 247 million hours to pay for the memory. By contrast a carpenter today makes $32.86 per hour (salary and fringe benefits), and could pay for the memory in less than 1 hour and 32 minutes. The ratio is 162,600,000 in a period of 36 years. If anything, Moore's Law is quite conservative.

I then asked what *process* futurism would say about this rapid computer development. Hans Moravec and others point out that, based on trends from the last 20 years or so, the rules *do* seem to be changing – the doubling time appears to be shrinking to 18 months. The major hypothesis put forward for this reduced doubling time is that computers are being used to assist in the design of future computers. Thus content futurism would predict that faster computers will shorten doubling times even more. Thus Moore's Law becomes a "slowest growth" scenario.

Discussing the Human Brain

The next question I address is what will it take for a computer to become as powerful as the human brain. From a physiology point of view, it is estimated that the human brain contains between 10 billion and 100 billion neurons. Determining the neuron capacity equivalent of a current computer a bit more problematic. Estimates run from a low of 50,000 to a high of 20,000,000 neurons. This means we have between 1/2millionth of a brain to 1/500 of a brain in computing power. While this is a large variation, these two estimates are "only" off by a factor of 4000, or 24 years as doubling time goes. These estimates place the development of a computer as powerful as a human brain between 2019 and 2043 based on content futurism, significantly sooner based on process futurism. To me, I point out, it is not a question of *whether* it will occur, but *when* it will occur. Bill Joy picks a nice average number of 2030.

However, just because we have a neural network as large as a brain does not mean we have anything resembling the functioning of a brain. I next asked the students how many had read John Searle's book *Minds, Brains, and Science* [4]. Quite to my surprise, very many students had read the book. (I determined later that the book was used in a general education required philosophy course.) This made it easier to discuss Searle's argument against artificial intelligence, which is summarized as follows:

1) Brains cause minds (i.e. the process of mind is caused by the brain)
2) Syntax is insufficient for semantics
3) Programs are defined by their structure
4) Minds have semantic content
5) Therefore no program can give a system a mind

I ask the programmers in the audience whether they agree with the third point. What is in a program? Do you just write code until you obtain no syntax errors and call that a program? I ask whether that isn't like saying the proper use of the English language is defined by its syntax, therefore it is impossible to communicate semantic content in English or any other language with a syntax. However, it is clear that the program is the key ingredient in turning a collection of hardware into something that might function like a brain or a mind.

Doomsday Scenario

I was anxious, given the limited time of the colloquium, to make sure that students did not go away from the talk with the same "way cool" reaction of my earlier classroom discussion. I took the arguments from Joy and Moravec and I developed a scenario based on the following line of reasoning:

1) In 2030 (give or take 10 years) computers/robots will have more "mental" power than humans
2) Computers/robots will starting designing the next generation of computers/robots
3) Doubling time will shrink to 1 year as robots become twice as powerful as humans (in 2032)
4) Doubling time will further shrink to 6 months as robots become four times as powerful as humans (in 2033)
5) Doubling time will again shrink to 3 months as robots become 8 times as powerful as humans (in 2033 ½)
6) Points 3 through 5 establish a series that converges to 2034 when robots become infinitely more powerful than humans, which is clearly impossible

7) What will happen instead is that robots will exploit all known laws of physics and explode in capacity before 2034

8) Robots will begin to self-replicate

9) Robots will need energy – presumably solar – and matter – presumably sand

10) There is plenty of sand, but limited surface area on earth, which will largely be blocked from sunlight by robot solar cells

11) Carbon-based life forms will become extinct

I thanked them for coming and asked for questions. You would think I had just failed them all out of school. I obviously didn't get the "way cool" reaction. They saw the consequences. Some started to ask what they could do about this problem. Others asked if I had seen the movie The Matrix [5]. (At the time I hadn't, but I have subsequently). At this point I reminded them of some of Bill Joy's main points, namely that technologists develop technology, and that they may not think of all the implications when they are caught up doing development. However, discoveries cannot be undiscovered (genies can't be put back into bottles) and safeguards are not automatically placed in new technologies.

I concluded by observing that someone who is technologically capable and literate must see the consequences and resist the tremendous commercial pressure to insure that destructive scenarios are averted by paradigm shifts (changes in rules)

The Second Seminar

The telephone started ringing the next day. Would I write a paper? Would I give another talk? I agreed to give a presentation entitled "Will Robots Overpower Us?" as part of the University of St. Thomas's Think St. Paul series at the Science Museum of Minnesota. Clearly the talk had to reach a happy medium between the classroom discussion and the colloquium.

Altered Organization

The talk could not end with the doomsday scenario. The talk still had to talk about futurism and Moore's law at the outset. The talk would be attended by a much wider audience. In addition to faculty and students, the majority of the audience would be the general public. I settled on the following organization:

Thinking About the Future

This section covered essentially the same material that I had covered in the colloquium under the discussion of futurism. I again asked the audience how many people had already read Bill Joy's article. Despite the much larger audience, there were many fewer who had read the article, but all continued raising their hands when I asked how many felt they had come in on the middle of a conversation. They were not likely to be bored.

The Future of Computer Hardware

This section covered the material on Moore's Law that I had covered in the colloquium. It also covered the discussion of the size of the human brain and the neural simulation capabilities of computers from the colloquium discussion of the human brain. Searle's argument about computers not having a mind would be saved until *after* the doomsday scenario.

Bill Joy's Scenario

This section consisted of an expansion of the material I had used to respond to students questions during the colloquium. It covered the major arguments in Bill Joy's article:

1) Technologists develop technology.
2) Technologists may not think of all of the implications.
3) Genies cannot be put back into bottles.
4) Future robotic power will be awesome.
5) Safeguards are not automatically placed in new technologies.
6) New technologies can take on a life of their own.
7) GNR technology (genetics, nanotechnology, robotics) will allow self-replication.
8) Self-replication can be destructive.
9) There will be remendous commercial pressure to use GNR technology, which may cause some to violate Asimov's Laws [6].
10) This is not just one person speaking, Bill Joy is quoting Searle [4], Kaczynski [7], Moravec [3], Kurzweil [8], Asimov [6], and Gelernter [9].

Isaac Asimov's Laws of Robotics

Because this was a general audience, I reviewed Isaac Asimov's laws of robotics as articulated in 1950:

"First Law: A robot may not injure a human being, or, through inaction, allow a human being to come to harm.

"Second Law: A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law.

"Third Law: A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law." [6]

I pointed out that robots developed by military interests are unlikely to hold fast to these three laws, and commercial interests are likely to want to increase the priority of the Third Law to save replacement costs.

Doomsday Scenario

At this point I presented the doomsday scenario in generally the same fashion as at the colloquium, but I quickly made a transition to the next section.

Other Points of View

In this section, I discussed John Searle's [4] mind/body problem, namely, that a mind/body connection implies consciousness, intentionality, subjectivity, and causality. I then discussed Searle's argument machine intelligence. In addition, I talked about the origin of the word "Luddite."

I mentioned that my father-in-law, George Martin, a history professor for many years at the University of St. Thomas often told me, "In all of history, no change, however beneficial to mankind, has ever occurred without harm to some group." I continued this line of reasoning by saying that Ted Kacsynski, the Unabomber, would add, "and that effect is unpredictable."

Software of the Future

Jaron Lanier [10] published a rejoinder to Bill Joy's article. In this article he maintained that while the hardware might eventually be there to build the robots of the doomsday scenario, that stupid software would surely save us from such a fate. He suggest that while evolution is brilliant at optimizing, it is stupid at strategizing and therefore all strategy must come from humans. He also postulates a reverse Moore's Law – as processors become faster, code becomes more bloated, using the resources. He concludes that crummy software will cause every man, woman, and child to be employed at the help desk.

Having heard that "every man, woman, and child" phrase before, I recall that is was said about the effort to tabulate the census in 1880. That led to punched card data processing. It was also said about manually switching telephone calls. That led to automated central office equipment. In other words, there will be a paradigm shift.

I pointed out that software needs an architecture, hence an architect. I also pointed out that programmed computers have difficulty with some problems. For example, they can't solve the halting problem (which means you can't write a program to test whether Windows® will ever give you the blue screen of death). They can't answer every question that can be asked. Finally, they can't solve a large class of problems (NP-hard) in a reasonable time.

Finally, I pointed out that software projects of the magnitude of replicating and improving upon the human suffer from problems of size and scale. Scaling up software is known to be hard to do. Many applications of large scale suffer from the 80% complete syndrome – the project gets to 80% completion and no matter how much effort is exerted, there is always 20% more to do.

Based on the question and answer session and the comments by students after the talk, this approach succeeded in getting students to think about the consequences without the depression syndrome of the earlier talk.

The Future Seminars

I have already scheduled four additional lectures on this topic. In these presentations I hope to enhance the materials based on comments from the audiences.

Probabilities

At the second (public) presentation, one gentleman said very kindly but firmly that he had trouble believing me. It occurred to me that not all predictions made in the presentation are of comparable likelihood. In future lecture I plan to give some rough measure of confidence as to the likelihood of the prediction. For example, the likelihood of computers eventually having 1,000,000 times the capability of current computers is very high. However, the likelihood of having a system that will model 1,000,000 times more neurons than current systems is considerably lower, at least in a comparable time frame. The doomsday scenario is quite unlikely, unless everyone ignores Bill Joy and me.

Alternative Scenarios

Given a longer presentation time (the public lecture was 40 minutes), I would like to explore more robotics scenarios. Ray Kurzweil, Hans Moravec, Bill Joy, and Isaac Asimov through his fiction give us a plethora of possible robotics scenarios. These possible futures will hopefully allow students to think about how they can conduct their professional lives to steer toward a desirable scenario and away from the doomsday scenarios.

Applications to Other Courses

In addition to artificial intelligence and robotics, I frequently teach a senior-level course in database design using Oracle® software. Each semester I spend more time discussing data mining. After discussing the technical aspects of data mining, I engage in a discussion of the current market place. I point out that people riding on any given airplane have typically paid a variety of prices for their airline tickets. I point out that “membership clubs” have different prices for members and non-members. I mention the trouble that Amazon.com got in to when they tried marketing music at different prices to different consumers. I then paint diverse scenarios for the application of results from data mining -- from as positive as custom merchandise that suits each individual to as sinister as situational pricing for all retail goods similar to airline ticket pricing. The theme is the same:

1) Make sure they have plenty of background information
2) Make sure they understand how the mechanism works
3) Point out where we are now with concrete examples
4) Present alternative scenarios of the future, including at least one plausible doomsday scenario
5) Continue with a discussion of obstacles that counter-indicate the various scenarios
6) Point out why they need to make a difference and how they can do it
7) Make them aware that what they do matters and they need to think of the consequences of their efforts.

References

1. Smith, L. (1998). *Guidelines for Preparation of Proceedings Papers*. New York: Doubleday.

1. Joy, Bill. (2000). “Why the Future Doesn’t Need Us.” *Wired Magazine*.

2. Barker, Joel Arthur. (1985). *Discovering the Future: The Business of Paradigms*. St. Paul: ILI Press.

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

4. Searle, John. (1984). *Minds, Brains and Science.* Cambridge, MA: Harvard University Press.

5. Wachowski, Larry and Wachowski, Andy. *The Matrix*. Hollywood: Warner Brothers.

6. Asimov, Isaac. (1950). *I, Robot*. Gnome Press.

7. Kaczynski, Ted. (1997). “Unabomber Manifesto.” *New York Times*.

8. Kurzweil, Ray. (1999). *The Age of Spiritual Machines*. New York: Penguin Books.

9. Gelernter, David. (1997). *Drawing Life: Surviving the Unabomber*. Free Press.

10. Lanier, Jaron. (2000). “One-Half of a Manifesto.” *Wired Magazine*.