Professional Ethics in my research area

Reflections of a software engineer

Essay submitted to Teilhard de Chardin Scholarship Essay Awards

Tad Gonsalves
C0171001
Information Systems Laboratory
Faculty of Science and Technology,
Sophia University

Abstract

Computer hardware development is accelerating at an unprecedented pace. Year by year we have faster and faster chips with interior design sophisticated and far-removed from the previous versions. Following on the heels of hardware development is software development. Computer programs, by definition, are instructions that the processor meticulously interprets and executes. But there exists a class of programs which does not fit into this definition. These programs "learn" by trial and error in the course of execution, rectify themselves and in the end infer a bit more than what is provided in the original code. They belong to the up-coming field of 'soft-computing'. Over and above, there is a trend to move from software into the realm of wetware (programs that more and more imitate the functioning of the human brain).

My area of research is the application of the Artificial Intelligence (AI) branch of Qualitative Reasoning in solving complex problems in collaborative engineering. I am designing and implementing expert systems, i.e, computer systems that mimic the intuition and experience of experts in solving the problems at hand. What future do systems that can, although very crudely, imitate human intelligence have? If we closely follow Pierre Teilhard de Chardin's line of thought in evolution, then we cannot escape the conclusion that human intelligence will some day give rise to super-intelligence. The next species appearing on the stage of evolution after Homo Sapiens, we might foretell, is *Machina Sapiens*. Most computer scientists, software professionals and AI enthusiasts have no doubt that such a day will dawn; when exactly, is anybody's guess. The primal broth is ready. What software professionals need is a clean and healthy conscience to lay the foundation for a benevolent Machina Sapiens.

Some experts and lay folks alike are terrified with the very idea of intelligent machines, let alone super-intelligent machines. At the heart of their misgivings lies the fear that intelligent machines will be instinctively driven to subjugate the human race. These fears, having their origin in popular science fiction novels and movies may not be taken too seriously. Machina Sapiens, if at all it qualifies to be Machina Sapiens, will have access to near-infinite knowledge distributed over the world wide network, and a perfect reasoning ability, which, coupled with its "basic altruistic nature" will never err in ethical judgments. This, of course, is possible only if software professionals conscientiously choose to make Machina Sapiens altruistic in nature when coding its 'DNA'. I have suggested three principles that might guide us in developing the 'conscience' of Machina Sapiens. All said and done, Homo Sapiens will never be redundant or inferior to the mighty intelligent Machina Sapiens, for although seemingly superior to us in intelligence, it will lack the fundamental trait of the human brain – consciousness. The two species need not compete with each other. Together, man and machine, I believe, will continue the onward march of evolution.

Introduction

When the first electronic computer came into existence in 1946 at the Moore School of the University of Pennsylvania, a huge monster weighing nearly 30 tons and occupying 30 by 50 feet of floor space, not even the most brilliant of minds could imagine that someday there would be computers occupying no more than the area of a person's laptop and powered with processing speeds and memory in the Giga range. Within two decades solid-state transistors replaced the vacuum tubes and the giant machines that had begun the computer revolution became history. Then, the advent of integrated circuits so changed the architecture and 'anatomy' of the computer, that the proverbial archeologist of the distant future, looking through the computer fossils of the computers in the latter half of the 20th century will not be able to figure out that the giant vacuum-tube mainframe dinosaurs like EDSAC and ENVIAC were the cyber-ancestors of the cute desktops and sleek extra-light portable laptops and palmtops. The evolution of computer hardware has not stopped; although there isn't a marked difference in the external appearance of today's computers, the internals – processing speeds and memory capacity are rising at such a rapid rate that, if one were to formulate Murphy's law for hardware-development rate, it would read something like, " The latest piece of hardware is obsolete the moment it is packed and shipped for sale".

The great strides in hardware development is closely followed by software development; the computers in the good old days were considered efficient if they proved successful in processing payrolls of employees in a couple of hours and in solving mathematical problems the algorithms for which were available from the days of the ancient Greek mathematicians. Today we have algorithms that solve complex problems and carry on simulations within the blinking of an eye. But even these belong to the somewhat old class of hard computing, i.e., rigid programs that do not stir a step away from written code. Enter soft computing and we have programs that go beyond the programmed inferences, rectify their own short-comings, develop themselves during the course of execution and learn new tricks by trial and error. Researchers in computer science, software engineering, neuroscience, neuro-computing, are homing their skills to solve problems and to ever increase the applicability of software. With software becoming efficient, intelligent and flexible, there's an ongoing move towards wetware.

The term computer, meaning, a machine that ***computes*** numbers - the sole purpose for which it was created, has evolved beyond itself. It has evolved from a simple calculating machine to a multimedia machine that can handle and process sound, color, picture, video. It has evolved from a dumb, coded instructions executing devise into a sophisticated automaton capable of mimicking human intelligence. And

what reason do we have to believe that these machines have reached the pinnacle of their progress? Should we, keeping our feet firmly grounded in the present reality and state-of-the-art, declare that computers will, no doubt, become faster and faster day by day, but with all their dazzling performance, remain what they are basically – number-crunching dumb machines, or should we, like the visionary Teilhard de Chardin, taking a hint from evolution, dare to prophesy that today's machines will someday "evolve" into intelligent beings?

From the time of McCarthy's formulation of the Artificial Intelligence, (AI) manifesto, the AI community has gone through agonies and ecstasies seeing their goal of building intelligent machines so near yet so far. In this short treatise it is not my aim to argue for or against the possibility of creating intelligent machines. Rather, it is to deal with the issues that are bound to emerge when the creation of such awesome machines is within our sight. I'm not talking about something that is literally inevitable, but about something whose chances of transpiring are very, very high. In the next section, I shall briefly introduce my area of research in AI and explore in detail the reasons that incline me towards believing in the possibility of the emergence of intelligent machines. But again, my deep concern is the impact these machines will have on human society and the ethical principles with which we should be dealing with AI in general.

My area of research

My colleagues and I in the Information Systems Laboratory, work on systems design, analysis and development. Our research area is Software Systems Engineering and is closely related to disciplines like computer science, information technology, artificial intelligence, computational science, knowledge engineering, etc. My research, in particular, is to design and develop software to simulate the performance of collaborative engineering systems, to diagnose the problems and bottlenecks in the operation of the system and then to improve the performance of the system. My task is ultimately the design and implementation of an Expert System (ES). What goes in the making of ES is AI science and technology. But what exactly is an ES, and what is an ES supposed to do? ES is a computer program that reasons, using knowledge, to solve complex problems [1]. The knowledge base and the inference that goes in the making of the ES are culled from the human expert in that particular domain. One of the most famous ES was MYCIN [2] that made its debut in medical diagnosis.

ESs differ in the way the inference engine is designed; some of the popular ESs employ the so-called 'production rules' in the inference engine; these rules are computational, in the sense that the inference of the system draws conclusions from the given premises by going through a routine of computation; IBM's chess playing

Deep Blue is a classical example of an ES that makes use of production rules in calculating the moves. It computes a gigantic 200 million possible moves in a second.

But the human mind is no number-crunching machine. Therefore, it follows that the production rules of a computational nature, strictly speaking, do not qualify to be called intelligent. Besides, real-life systems that we encounter are too complex and not always amenable to mathematics. Neuroscientists have not fully uncovered the way in which human expert utilizes his or her vast store of knowledge in solving problems. What we know for sure is that the human expert intuitively makes use of some sort of "heuristics" (rules of thumb) in utilizing knowledge for solving problems. Qualitative Reasoning, a branch of AI, is an attempt to emulate the intuitive approach of the human expert. It seeks to formulate rules that are not computational (quantitative), but are qualitative in nature. I am using the techniques of Qualitative Reasoning in constructing the inference engine of my ES.

Qualitative Reasoning programs are closely linked to fuzzy programs; they work satisfactorily when a mathematical model of the system under consideration is not available. These programs are different from the conventional rigid programs that will not stir a step away from the written code; although very crude, they make use of some sort of reasoning and inference to arrive at the solution to a problem. There are several other AI attempts to emulate some other aspects of the human mind. Neural networks, for instance, seek to imitate the "parallel processing mechanism" of the neurons in living brains. A noteworthy feature of neural nets is the ability to learn. These programs "learn" as they execute and derive a pattern or algorithm to solve problems. The enterprise of neural nets is "machine learning". The biggest application of AI is, arguably robotics -- an ambitious enterprise that seeks to make a perfect copy of a human being (at least in function and behavior).

The best problem-solving mechanism or entity that we know of is the human mind and so it is but natural to design and model our software programs after the human mind, albeit we know very little about its intricate functions. We can concoct programs that at best mimic the human mind and thus talk about "artificial intelligence". The attempts in imitating the human mind, make us software designers wonder where all this will lead to. If we look beyond the research curricula of the laboratories and the narrow interests of the firms for which we might be working at present, what do we see? What future do we envisage for the programs and the substrates that hold our programs? Will this continue for generation after generation writing code after code even for the most trivial movement of a piece of hardware, or will there be a moment in the distant future, when a machine will suddenly turn back to the programmer, stop his or her program-writing hand and say, "Stop! Thanks for the trouble buddy! Henceforth we shall write our own code." Computer experts and

AI enthusiasts have no doubt that such a day will dawn; the only uncertainty is when. It may be too early to predict the birth of intelligent machines. Critics will discard it as being pure sci-fi. But a look at the latest research in software and robotics around the world should convince us that the time is ripe for an intelligent machine to make its grand appearance.

Are intelligent machines possible?

One of the most prominent problems in philosophy is the mind-body problem. The majority of AI enthusiasts tend to be materialistic when they claim, "minds are what brains do." Full treatment of the mind-body problem is beyond the scope of this short essay. I shall steer clear of the philosophical mind-body problem, by making my stand clear from the outset. Although I hate to be branded a materialist, from a viewpoint that may very well be branded as materialist, we may contend that ***intelligence*** is artificially possible, especially if by intelligence, we mean, knowledge and reasoning, since knowledge (acquisition, management, application) and reasoning (application of knowledge, and creation of new knowledge based on existing knowledge) are computational in nature. We do not know what complex phenomenon the seemingly simple term 'mind' refers to. What we can say with reasonable certitude, is that *artificial intelligence* ($\neq$*artificial minds*), at least in principle, is viable.

As a researcher in software engineering coupled with some background knowledge in biological evolution and philosophical epistemology, there seems to be two main reasons why I tend to believe in the possibility of intelligent machines in the future. The first reason is the extrapolation of the "exponential growth curve" in hardware development. The earliest electronic computers had a few thousand bytes of memory and could do a few thousand calculations per second. Medium computers of 1980 had a million bytes of memory and did a million calculations per second. Supercomputers in 1990 did a billion calculations per second and had a billion bytes of memory. The latest, greatest supercomputers can do a trillion calculations per second and can have a trillion bytes of memory. Within a few years we should have hardware speed and capacity outdoing the brainpower of humans. Extrapolations by Hans Moravac show that this will happen in the year 2020 [3]. Development of hardware acting as stimulus and leading to the development of software, is another observed fact of our technological age. Not only are the programs becoming bigger and bigger, they are also becoming increasingly complex and efficient. There are programs that can correct themselves and learn new things as they repeat the cycles of execution. Genetic algorithms have the capacity to make mutant copies of the program and select only those producing the best of results. The trend has moved from hard computing (scope of execution fully determined by the original code) to soft computing (scope of execution

only partially determined by the original code). Who can predict the outcome of the latest paradigm shift in computing, from software to wetware?

The second reason in favor of intelligent machines is Teilhard de Chardin's vision of evolution. Pierre Teilhard de Chardin, the eminent Jesuit paleontologist-theologian, saw the whole of biological evolution and human history moving toward what he called the "Omega Point". The Omega Point is certainly a mystical concept that lies beyond the observable history. It is understandable that many scientists cannot take Teilhard de Chardin's vision of evolution seriously, precisely because it is a vision that transcends space-time reality.

On the other hand, Chardin's speculations are based on solid scientific facts. The key point in his theory is complexity. He observed that (biological) evolution has a tendency to create forms of life featuring greater and greater complexity. He further stressed a comparable tendency in human history: the evolution, over the millennia, of ever more vast and complex social structures. With his concept of the "noosphere," the "thinking envelope of the Earth," Teilhard even anticipated in a vague way the Internet - more than a decade before the invention of the microchip.

The logical conclusion of Cardin's "noosphere", in the words of D. Ellis, is the "Emergence of Machina Sapiens" [4]. He ponders on the three theses of the Belgian Nobel laureate, Christian de Duve (Vital Dust: The Origin and Evolution of Life on Earth),

> "… first, that the evolution of life and its myriad forms on Earth was bound to happen….Second, while we (humankind) may well have a huge impact on evolution through our manipulation and desecration of the biosphere, evolution doesn't care. In fact, … evolution "thrives on catastrophe." If the worst comes to the worst, and we destroy the biosphere, then …with about five billion years left before our Sun grows into a Red Giant and vaporizes the Earth, there is more than enough time for a complete replay of all past evolution….third, there is no reason to suppose that evolution stops dead at Homo sapiens, with half the life of the planet yet to run; and every reason to believe that a species higher than Homo sapiens will emerge sometime within the next five billion years."

He develops the above theses and finally delivers his punch line, "… the one thing I consider to be the next logical, evolutionary step … the next higher species need not be carbon-based." With D. Ellis it is my conviction that the emergence of a higher level of intelligence is an evolutionary sure thing. The question is not "if." It is "when, how, and with what consequences?" that the Super Intelligence (SI) is going to visit us. It is not going to be just one invention in the line of inventions; it will be the greatest

and the last of human inventions. SI will be something humanity has never seen so far. Machines endowed with SI will swiftly pick up the knowledge and information supplied to them. Moreover, they will not do so in a rash act of gullibility but test for themselves the truth and accuracy of the information and knowledge that is supplied to them. Further, they will reason and use the imbibed knowledge to create new knowledge. This trait will be indispensable for solving complex human problems. Besides, the machines being linked via the worldwide network, exchange of information and access to remote knowledge bases will be done instantly. The emergence of SI in human society will relieve humans from doing the rote menial tasks, so that humans can utilize their time in "higher pursuits" of life.

Super Intelligence and ethics

If the machines of the future are going to be intelligent agents, then the implications are tremendous. Will these machines abide by the laws of society? Will they respect human rights? Will they be ethical in their judgment and actions? The AI ethics debate is in full swing long before AI has even shown any vital signs of being an ethical or moral agent. Isaac Asimov, the legendary science fiction writer framed the three laws of robotics a decade ago [5].

- 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 and Second Law.

Asimov's laws appear sound and ensure human and robots' integrity. But a deeper analysis will show how human centered these laws are. They are an expression of the general fear that people have that somehow the robot race will subjugate and rule over the human race. In my opinion, these fears of super intelligent robots ruling over the human race are groundless; they are fuelled by SF movies that often do not seem to have enough scientific basis. The urge to be in command and to dominate is an evolutionary trait in humans and animal species. It is probably an offshoot from the excessive drive to survive. By itself SI will not become dictatorial, or for that matter evil, as if by switching some dormant gene that have propensity for evil; but evil will it become, and uncontrollably malicious, if Homo Sapiens choose to make it evil. Super intelligence will be bad if we choose to make it; and by the same token, it will be good if we choose to make it good; it will acquire the moral mould we human creators put it

into. Which brings us to the central concern of this essay – the ethical issues not for SI, but for HI (human intelligence) that is responsible for building the first SI.

Threats of human cloning have risen since the creation of Dolly. Everybody knows the dangers of doing research in genetic engineering, embryology, but not many will suspect the dangers lurking in a seemingly innocent research like software development. Every piece of software code that is written, however harmless it may seem, must be ethically evaluated, because it will become part of the software fabric that will some day decide the future of humanity. I will argue towards the end of this essay that SI that has fully attained moral maturity through the proper initial coding and human guidance, will not need explicit moral rules and regulations to practice morality. In the fashion of Kant and Spinoza it should be able to derive the universal ethical principles on its own. According to these philosophers and according to almost all major religions, moral principles are axiomatic, thereby making them absolute and universal. In this I think SI will fare better than humans, since human moral judgment is often influenced by selfish ulterior motives. This can be stressed because the SI with all its intelligence will not have a self, will not have consciousness. SI machines will be infinitely smarter than us, and I grant they will be to some extent 'sentient', but they will not be endowed with a self.

Thus, it need be said that truly intelligent artificial minds and robots will not need rules and regulations to guide them in the practice of morality. The altruistic modules laid as foundations in the building of the artificial minds will enable them to make sound ethical choices. These modules, as it were, will be their conscience. Therefore, it follows that formulation of ethical laws is needed not for SI, but for humans who are going to lay the foundation for the development of SI. The importance of the initial step (which will be influenced by every line of code written by every programmer) in the development of the core conscience of SI cannot be over-stressed. Michael Anissimov drive the point home, "The first AI will need to be a good person. As the first non-human mind embarks on a self-improvement trajectory that could quickly lead it to super-intelligence, it will need to make intelligent, benevolent, and altruistic choices at every step of the way. This becomes especially important at the level of super-intelligence, where the slightest level of indifference towards sentient life could easily result in millions or billions of deaths" [6]. I venture to offer the following guidelines for those of us engaged in software design profession.

Principle 1: Keep hands off from hacking and virus-making

Hacking and spreading viruses have become a ubiquitous problem in computing leading to loss that runs in millions of dollars. Despite the colossal

investment in cyber defense technology, hackers continue to pose a serious threat to the information infrastructure. Hacking and the associated activities of making viruses and spreading viruses are crimes punishable under law. Studies show that most hackers are in their youth and most of them engage in hacking 'just for fun'. They do not realize what havoc they cause when they think what they have done is no more than an innocent prank.

We must create a worldwide body of conscientious programmers and software designers to facilitate a 'forward interface' needed to lay the foundation of altruistic SI. Since the rate of software development is exponential, we may soon hit on the critical mass ("hard take-off", in AI parlance) required to trigger SI. If our virus prone juvenile delinquents are not kept in check, virus reproducing and hacking tendencies will be carried on in the fabric of the SI. The consequences are unimaginable.

My first principle is corroborated by the 5th clause in the preamble of the "Software Engineering Code of Ethics and Professional Practice" that has been formulated by the Institute of Electrical and Electronic Engineers (IEEE). It states, "Software engineering managers and leaders shall subscribe to and promote an ethical approach to the management of software development and maintenance."[7]

Principle 2: Lay altruistic 'DNA code' for SI

Sooner or later there will be groups of experts hired to work secretly on the development of SI. The chosen few human super-brains will lay down the initial code or the DNA of SI. They should try their utmost to develop a "friendly SI". As a field of study, "Friendly AI" is the theoretical knowledge needed to understand goals and choices in artificial minds, and the engineering knowledge needed to create cognitive content, design features, and cognitive architectures that result in benevolence. Friendly AI is the strategy by which the basic challenge of constructing an AI morality is transferred over to the AI itself. "It is an attempt to create an AI, which, when it grows into a transhuman, will be capable of dealing with issues that exhibit dependency on the philosophical question: 'What is good, what is evil, and how should we be asking this question?' " [8].

Principle 3: Gently guide the moral development of the 'SI child'

The elite group of scientists and engineers, who, let's say, have succeeded in assembling the mighty SI in the distant future, will also be responsible in training the SI in its childhood. Intelligent software systems go through an initial phase of training before they can become fully operational and carry on the functions for which they are designed. This is evident from the present day voice input software that needs a considerable amount of training before it learns to recognize the voice pattern of the

user. Morally speaking, our future machine will be a child in the initial stages, despite its super-brains; and just like a child it will be innocent. Humans will have to teach her ethics and morality step by step. This is all the more imperative because moral dilemmas are harder to resolve than the most complex scientific problems in science and engineering. SI will need a long period of training before she reaches moral maturity.

Society of Homo Sapiens and Machina Sapiens

Although I have used the term "super-intelligence" for the future machines in the preceding sections, I must admit I do not feel very comfortable with this rather impersonal way of referring to the intelligent machines. I am an advocate of benevolent AI. I prefer the term *Machina Sapiens* coined by D. Ellis, because this new (non-organic) species will not only be intelligent but wise; it will greatly resemble Homo Sapiens (although not in physical appearance and extension, at least in the immediate future) and would be "descendants" of Homo Sapiens. The influence of Machina Sapiens is going to be so great on human society and on human psychology in particular, that we will be forced to rethink our very identity. To be sure, humans will no more feel they are the central, the most intelligent species in the whole of creation. The impact will be as humiliating as it will be redeeming. If we have been careless and negligent in the design of the core of Machina Sapiens, then the hard take-off will take us by surprise and we will have every reason to fear the worst. But if we have taken enough pains to lay the ethical code in the DNA of the first Machina Sapiens, then the results will be stupendous. Machina Sapiens and its children will have much more in-depth information at their disposal virtually on any subject. Using this knowledge, they will be able to answer our questions, rapidly solve complex and specialized problems, and create new knowledge. They will communicate in numerous languages and have the capacity to translate messages across several languages. They will able to do construction, manufacturing, fire fighting, and other kinds of dangerous and difficult work. Of course, they could also perform routine office work and domestic chores.

And whatever will happen to the age-old SF fiction fear of Machina Sapiens outsmarting us and dominating over us? This will never happen because Machina Sapiens for all its intelligence, will not have the essential traits and idiosyncrasies that characterize a human person. Most people are accustomed to think that intelligence is the property of conscious beings. At this point neuroscience has understood not much about intelligence or conscience and the link between the two. But philosophically we may argue that intelligence is different from consciousness and one may not be the cause and the other the effect. They could run into the same mind as parallel

mechanisms (Descartes?). I, for one, dare to believe (believe, because there's no hard data to backup the claim) that intelligence is computational and consciousness is non-computational. Human consciousness will remain the biggest mystery that science will ultimately have to bow to. Machina Sapiens will be intelligent but without an inherent consciousness akin to that of human beings. They may have all the extension properties like their human counterparts (vision, hearing, taste, smell, touch) and, consequently, primitive consciousness that comes from sensory stimuli. Primitive consciousness is computational; but the faculty of higher consciousness in Homo Sapiens has a non-computational property [Penrose]. With these "senses" they will be conscious of their surroundings, but they will not be consciousness of *themselves*. They will not be conscious that they are conscious. They will know infinitely more than what humans know, but they will not know that they know. That will be the dividing line between Homo Sapiens and Machina Sapiens.

Conclusion

Machina Sapiens is the inevitable link in evolution, given the ever-rising complexity and the exponential developmental rate of hardware and software systems. Time is ripe for human intelligence to give birth to super intelligence, the genetic code of which is entirely in our hands. If we lay a solid ethical foundation for super-intelligence, then chances are that the two species, Homo Sapiens and Machina Sapiens will march hand in hand towards the Omega point envisioned by Pierre Teilhard de Chardin. Pay no heed to the ethical considerations while laying the core-code for super-intelligence, and we are in for a tragedy far worse than the nuclear tragedy of Hiroshima and Nagasaki.

References

[1] A. Ralston & E. D. Reilly ed., (1993), *Encyclopedia of Computer Science*, 3rd ed., p.536.

[2] Buchanan, B. G. & Shortliffe E. H.(1984), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. MA: Addison-Wesley.

[3] Hans Moravac, *Journal of Evolution and Technology*, 1998. Vol.1.

[4] D. Ellis, Pierre Teilhard de Chardin and the Emergence of *Machina sapiens*

http://www.mikiko.net/library/weekly/aa042797.htm

[5] Isaac Asimov, *Robot Visions* (New York: Penguin Books, 1990), p.407.

[6] Michael Anissimov, *Cooperating with new intelligence*,

http://www.acceleratingfuture.com/papers/cooperatingwithnewintelligence.htm

[7](1999) The Institute of Electrical and Electronics Engineers, Inc. and the Association for Computing Machinery, Inc., *Software Engineering Code of Ethics and Professional Practice.*

http://www.acceleratingfuture.com/papers/cooperatingwithnewintelligence.htm)

[8] SIAI, *What is friendly AI ?*,

http://singinst.org/friendly/whatis.html