David Lanner

Prof. Carl

Computer Science 356

26 April 2010

# AI Inevitability

The existence of an artificial intelligence (AI) whose mental prowess matches or exceeds that of humans is at once a fantastical, frightening, and richly philosophical thought. According to futurists like Vernor Vinge and Ray Kurzweil, this AI of superhuman intelligence will not be confined to the pages and screens of our science fiction for much longer; it is inevitable. The moment this happens is an event (popularized largely by Vinge) known as as the "Technological Singularity", often shortened to "the Singularity".

The reason why futurists and other Singularity-enthusiasts are so confident in their predictions is based in large part on a mechanistic view of the universe: anything that is observable, measurable, and quantifiable can (at least theoretically) be modeled or reproduced. Consider that science and technology have enabled humans to model with great precision such parts of humans as subatomic particles, atoms, molecules, organelles, cells, cell tissues, organs, etc. The same logically applies to the components of the human brain. Consensus in many schools of thought in the neurosciences reflects as much- that the mind functions mechanically (e.g. the release of ions at synaptic connections, the firing of neurons by electrical signals, etc.). As such, it's conceivable that brain activity can be represented mathematically.

A frustration frequently encountered by physicists, mathematicians, and computer scientists the world over is that the theoretically possible is not necessarily easy to bring about, or find in the world. This is particularly evident in the fact that AI has yet to become intelligent by humans' standards, even though it's possible. Even though the computational power of today's technology far exceeds that of the neuron, a human brain composed of these neurons is still more "intelligent" than a computer. There are two closely related reasons for this.

The first reason is that, unlike modern computers, neurons have tremendous combinatorial strength. "High-speed electronic switching allows computers to explore alternatives thousands or even millions of times faster than biological neurons, but this power pales in comparison with the combinatorial abilities of the billions of neurons and uncounted synapses that constitute a brain." (Dyson, 109). In this way, brains are much more concurrent (that is, more capable of processing multiple operations in parallel) than today's hardware.

This leads into the second reason: there is a serious limit to the hardware that we can produce. Specifically, the hardware is limited by the number of transistors able to be fit onto an integrated circuit. According to Moore's Law (named after Gordon Moore, who famously published his observations on the trend in 1965), this number doubles every two years. "Given that the electrons have less distance to travel, the circuits also run twice as fast, providing an overall quadrupling of computational power." (Kurzweil).

Present-day computer hardware can only perform so many calculations per second (cps). This turns out to fall far below the computing power of the human brain. Kurzweil's "estimate of brain capacity is 100 billion neurons times an average 1,000 connections per neuron (with the calculations taking place primarily in the connections) times 200 calculations per second."

(Kurzweil). He predicts, based on his law of accelerating returns, that

- We achieve one Human Brain capability (2 * 10^16 cps) for $1,000 around the year 2023.
- We achieve one Human Brain capability (2 * 10^16 cps) for **one cent** around the year 2037.
- We achieve one **Human Race** capability[1] (2 * 10^26 cps) for $1,000 around the year 2049.
- We achieve one Human Race capability (2 * 10^26 cps) for **one cent** around the year 2059 (Kurzweil) [emphasis added].

Thus, the second limitation of hardware will, with time, cease to be a problem. The first will also cease to be a problem, but not for the same reason. As time goes on, the cps of machines which are less combinatorially savvy than the human brain will surpass the cps of their biological competition; with time and brute force, a great number of transistors will eventually overwhelm a significantly more coordinated but smaller number of neurons.

When this day comes – when machines are capable of more calculations per second than human brains – well, maybe nothing will happen. This is a possibility admitted reluctantly by futurists and AI enthusiasts (perhaps most reluctantly of all by Ray Kurzweil). The assumption is that at this time, the conditions will supposedly be right for AI to "wake up". But what if simply having the right ingredients doesn't excuse for not having the right recipe?

This is a possibility Vinge is weary of: "A plausible explanation for 'Singularity failure' is that we never figure out how to 'do the software' (or 'find the soul in the hardware', if you're more mystically inclined)" (Vinge). The events following this "Singularity failure" might depend, according to Vinge's speculation, on how far humans can push hardware while still being able to decently take advantage of those advances. "[hypothetical post-Singularity-failure] Software

[1] By "Human Race capability", Kurzweil means a computer which is capable of performing n calculations per second, where n is equal to the sum of the computing power (in cps) of every human brain in the world. At the time "The Law of Accelerating Returns" was published, this amounted to roughly 6 billion people.

projects that endeavor to exploit increasing hardware power fail in more and more spectacular ways. […] Such failures lead to reduced demand for more advanced hardware, which no one can properly exploit" (Vinge). After this, Vinge envisions three possible scenarios.

The first he calls "a return to MADness". In this possible future (or lack of future, as the name implies), perhaps the threat of dwindling environment resources is an even greater menace. Maybe international tension over said resources, or the threat of terrorism, causes a regression to the fearful policies of mutually assured destruction. "A return to MAD is very plausible, and when stoked by environmental stress, it's a very plausible civilization killer" (Vinge). This is, from the standpoint of someone who regards the survival of the human race as a positive goal, a terrible scenario. To counter this, Vinge proposes an optimistic alternative- what he calls "the golden age".

This hypothetical future sees an increase in both human population count and the amount of power it can extract or use (what he labels in a graph as "maximum power source", measures in Watts). In addition, research into methods of extending longevity may be given greater attention. Vinge suggests that this is because "[Young] Old People are good for the future of Humanity! […] We have no idea what young very old people are like, but their existence might give us something like the advantage the earliest humans got from the existence of very old tribe members (age 35 to 65)" (Vinge). Perhaps these young old people, because of their relatively extended lives (imagine a 500 year old person), will be able to provide amazing insights and much-needed wisdom. Perhaps, though, this is a much more probable future than this golden age; there is, in Vinge's view.

Vinge considers one more possible future after the Singularity fails to happen: "the wheel

of time". In this future, "we see cycles of [mega]disasters and recovery" (Vinge). If, after the first disaster, mankind fails to reach extinction, we might expect two possibilities: either there are enough resources for humans to rebuild again, or there are not. The consequences of the latter are that mankind is unable to extract precious metals and fuels like petroleum from the earth, and as a result may not be able to develop sufficiently; never again would humans be able to attain the state of development seen today. In the case of there being sufficient resources, however, then we can assume that either the Singularity happens this time around, or it doesn't.

Turning now to the more likely scenario, in which the Singularity actually happens: it is worth examining what will happen afterwards. Many Futurists like Kurzweil predicts that the machines will be able to create more efficient versions of themselves, which, in turn, will be able to improve on their own design, so on and so forth- potentially ad infinitum. With this incredible ability to improve themselves, the rate of change of progress will become so fast that humans will be "left behind", so to speak, by technological and scientific advancements. "Well, for one thing, they [the superhumanly intelligent beings] would come up with technology to become even more intelligent (because their intelligence is no longer of fixed capacity). They would change their own thought processes to think even faster." (Kurzweil). Saying that humans will be "left behind" evokes a sense of dread in most people; it implies a loss of control, which is unacceptable to many.

The desirable outcome in a post-Singularity world, for these people, is that humans still maintain some measure of control over these godlike beings, while deriving utility and pleasure out of them. This is, essentially, the desire to own and use an incredibly powerful, super-intelligent PC. This is understandable in the context of today's consumer-centric world; it is also

understandable when considering the alternative scenario, in which humans are unable for whatever reason to control these machines.

In this scenario, we may expect to see one of two basic outcomes: either humans continue to exist, despite not having any control over the new dominant type of life on earth, or they do not. In the case of the former, the new rulers of earth do not see the human race as a threat (or, as in *The Matrix*, a source of fuel or energy); humans would go on with their lives. In the case of the latter, the reader can refer to any number of science fiction novels and Hollywood movies concerning the various Apocalyptic possibilities, including such notables as *Battlestar Galactica; The Cyberiad; The Matrix; I, Robot; The Terminator* series; *Do Androids Dream of Electric Sheep?;* and *R.U.R. (Rossum's Universal Robots)*. A curious meeting between the scenario in which man goes extinction and the one in which he survives is that man merges with the machine. How might this come about?

One plausible explanation is that humans first start off simply using intelligent machines for convenience, utility, and pleasure as is currently being done with non-intelligent ones. Slowly, people begin to realize the benefits of relying more and more heavily on these machines for augmentation of their human abilities, such as amplifying their strength, speed, dexterity, intelligence, memory, longevity, and so forth. It could come to the point, eventually, that augmentation leads to humans transferring their identities — their personalities and knowledge and memories, and whatever else makes a human "human" — (or at least copies of their identities) into a machine. This is, many futurists believe, the only possible way for humans to survive the Singularity. The only concern someone might have would be the definition of "human" at that point.

But which of these can be expected? Any number of events are coming up on the horizon: whether or not the Singularity happens, whether or not humans will survive it, whether or not humans would even be humans at that point. Certainly no one can be sure of their predictions. This is what futurists like Kurzweil do, though- the prediction of future events based on past performance (or trends). The trend, in this case, is Moore's Law. It is arguable that so long as the rate of technological improvement follows the expected curve, then most futurists will be more or less accurate in their predictions. If this is true, then so is the one thing most futurists agree on: AI is inevitable.

# Works Cited

Dyson, George B. Darwin Among The Machines. Cambridge, MA: Perseus Books, 1998.

Kurzweil, Ray. "The Law of Accelerating Returns". 7 March 2001. 26 April 2010. <http://www.kurzweilai.net/meme/frame.html?main=/articles/art0134.html>

Vinge, Vernor. "What If the Singularity Does NOT Happen?". 15 February 2007. 26 April 2010. <http://www.kurzweilai.net/meme/frame.html?main=/articles/art0696.html>