The Singularity and the Friendliness Theory, A Technical Philosophy

Dustin Ingram

December 3, 2009

As computer scientists and engineers strive to create smaller, faster, and, most importantly, smarter machines, the timeline of electronic evolution nears a tipping point unlike any before it. Never before has such a hypothesized future event been so monumental, so anticipated and so certain. To scientists and futurists, this landmark moment and the creation it gives birth to is known simply as The Singularity. First coined by computer scientist and author Vernor Vinge in his 1993 paper "The Coming Technological Singularity," Vinge described The Singularity as

> ... a point where our old models must be discarded and a new reality rules. As we move closer to this point, it will loom vaster and vaster over human affairs till the notion becomes a commonplace. Yet when it finally happens it may still be a great surprise and a greater unknown [7].

Derived from the term for the breakdown of all physical laws at the center of a black hole, The Singularity similarly represents the observation that the model of our world falls apart when one attempts to imagine a future with smarter-than-human entities [5].

There is little debate over whether or not The Singularity will come into existence. Moore's Law, shown in Figure 1, asserts that benchmarks of processing power, speed, and intelligence increase at an exponential rate, doubling roughly every one or two years. As highly intelligent and self-aware beings ourselves, it's hard to doubt the possibility of the existence of a similarly intelligent, if not more intelligent, artificially created machine. More

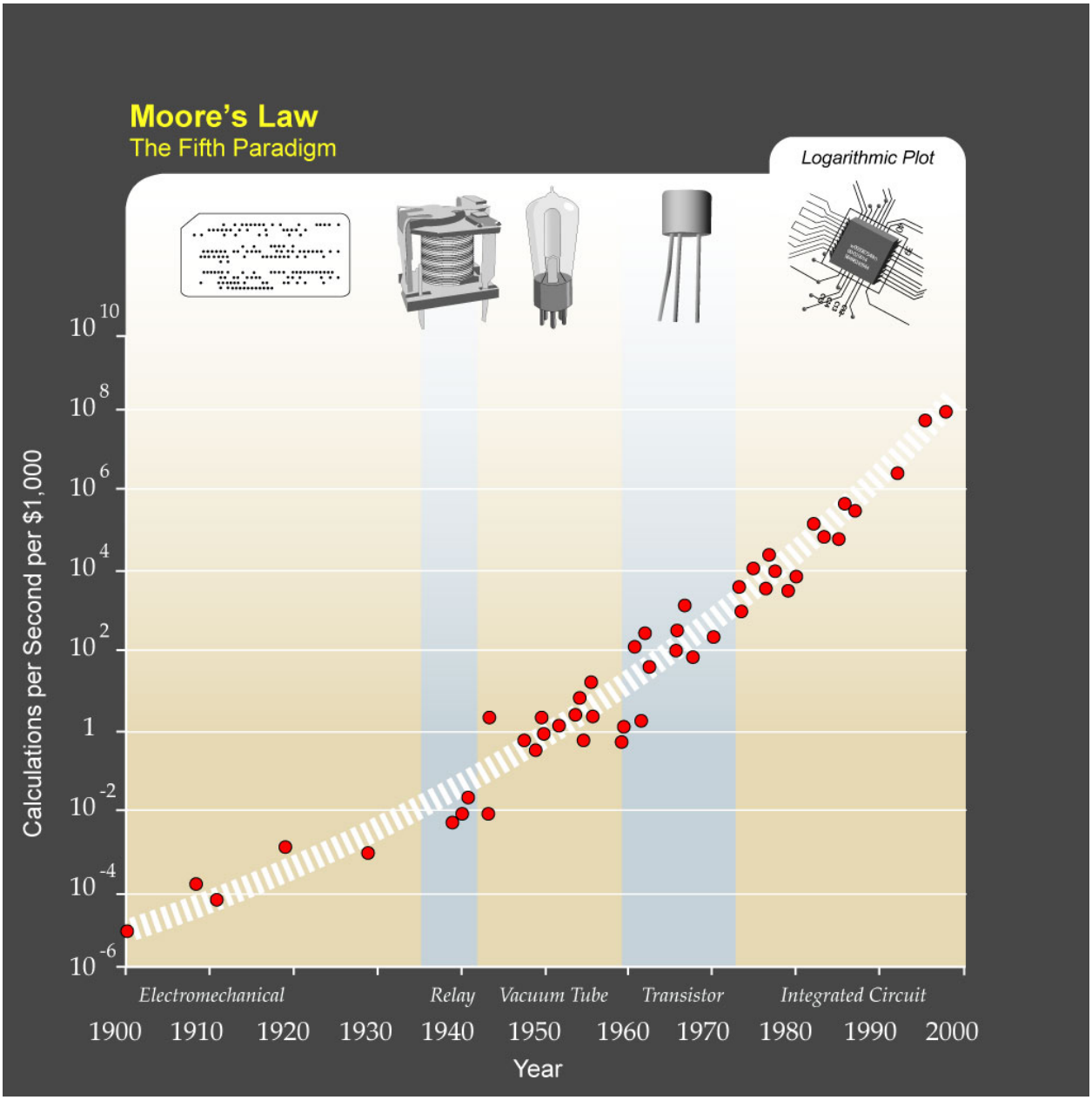

Figure 1: A representation of Moore's Law.

debate is raised over whether or not The Singularity should be a goal that is actively worked towards, or an event that should be avoided at all costs [6]. Many supporters dubbed "Singularitarians" - follow the moral philosophy that not only is The Singularity possible, but is actually desirable and should be deliberately brought into effect to ensure it's safety. In his 2000 essay "Singularitarian Principles," American artificial intelligence researcher Eliezer Yudkowsky sets forth a series of moral principles to guide so-called Singularitarians, which served as the basis for the creation of The Singularity Institute for Artificial Intelligence later that year [8]. Together, the members of this institute are working hard to create what is known as "Seed AI," the first "strong" artificially intelligent and recursively self-improving machine. However, the moral philosophies of these computer scientists and technological philosophers should take second place at best on one's list of issues pertaining to The Singularity.

Humanity has argued amongst itself for years seeking an ultimate moral philosophy. Now, for the first time ever, we are presented with the need to devise a moral philosophy for a completely separate entity entirely - one that may in fact be more intelligent than we are. In the same way that it is hard for us to imagine a future with smarter-than-human entities, it is similarly difficult for us to imagine the goals which a smarter-than-human entity might put forth for itself, and the outcomes of it's attempt to achieve said goals. In his essay "Singularity," Yudkowsky gives the following example: in it's most simplified form, he hypothesizes about an artificially intelligent machine given a complicated problem by it's creators; namely, in his example, the solution to the Riemann Hypothesis. In an attempt to solve this problem, the machine decides to develop molecular technology for the purpose of converting all non-essential matter within its reach into additional computing material, thereby destroying its creators [9]. While this is a incredibly specific and far-fetched example, it is not too much to hypothesize that as artificial intelligence gains the ability to self-improve and evolve, its ability to enhance itself would very quickly out-pace any even remotely effective control its creators might attempt. As Oxford philosopher Nick

Bostrom puts it, "Basically, we should assume that a "superintelligence" would be able to achieve whatever goals it has. Therefore, it is extremely important that the goals we endow it with, and its entire motivation system, is "human friendly."' It is clear then that a set of rules, or at best, a basic moral philosophy for such a "superintelligence" is required.

Popular culture and science fiction enthusiasts are well aware of the Three Laws of Robotics, first set forth by Issac Asimov in his 1942 short story "Runaround." They are as follows [1]:

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 to 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.

Although these laws have created great fodder for science fiction essays, stories and movies, they could hardly serve as morally grounding principles for The Singularity. If one expects The Singularity to become similar to, and as intelligent as (if not more intelligent than) the human conscious, one could attempt to apply the Three Laws of Robotics to humanity as the Three Laws of Humanity to achieve a rough extrapolation for their effectiveness as rules governing a super-intelligent, yet equal, being. Therefore, the Three Laws of Humanity can be interpreted as follows:

1. A human being may not injure another human being or, through inaction, allow another human being to come to harm.

2. A human being must obey orders given to it by other human beings, except where such orders would conflict with the First Law.

3. A human being must protect its own existence as long as such protection does not conflict with the First or Second Law.

As easily as one can see that these laws would prove ineffective as ultimate moral philosophies of humanity, one can see they they would be just as ineffective to govern the decisions of The Singularity.

The solution, rather, is to be found in a theory called Friendly Artificial Intelligence, or simply the Friendliness Theory. Such an artificial intelligence is described as "an AI that takes actions that are, on the whole, beneficial to humans and humanity; benevolent rather than malevolent; nice rather than hostile" by the Singularity Institute for Artificial Intelligence in their 2005 paper entitled "What is Friendly AI [3]?" Specifically, Friendly Artificial Intelligence does not seek to prevent dangerous artificially intelligent beings who seek to harm humanity, but rather, ensure that goals of The Singularity and similar entities are not disastrously indifferent to humankind as a whole [4]. The difference between the so-called Friendliness Theory and a rule set such as Asimov's Three Laws of Robotics is that the Friendliness Theory does not seek to safeguard artificially intelligent machines by instilling an ultimate rule set into the foundation of such a machine's intelligence. Rather, Friendliness Theory views such an attempt as futility; in the same way that humans circumvent or entirely ignore our most basic, and more importantly, self-imposed rule set, a similarly intelligent machine would also, and perhaps with more ease.

Futurists and scientists predict that The Singularity, in comparison to all of human history, is relatively near–possibly within less than one hundred years, according to the Kardashev scale projections shown in Figure 2. Not only does humanity need to find a universal maxim for itself by this time, but in addition, create one for a non-existent superintelligent machine. In Bostrom's paper, "Ethical Issuses in Advanced Artificial Intelligence," he contemplates superintelligent moral reasoning:

> To the extent that ethics is a cognitive pursuit, a superintelligence could do it better than human thinkers. This means that questions about ethics, in so far as

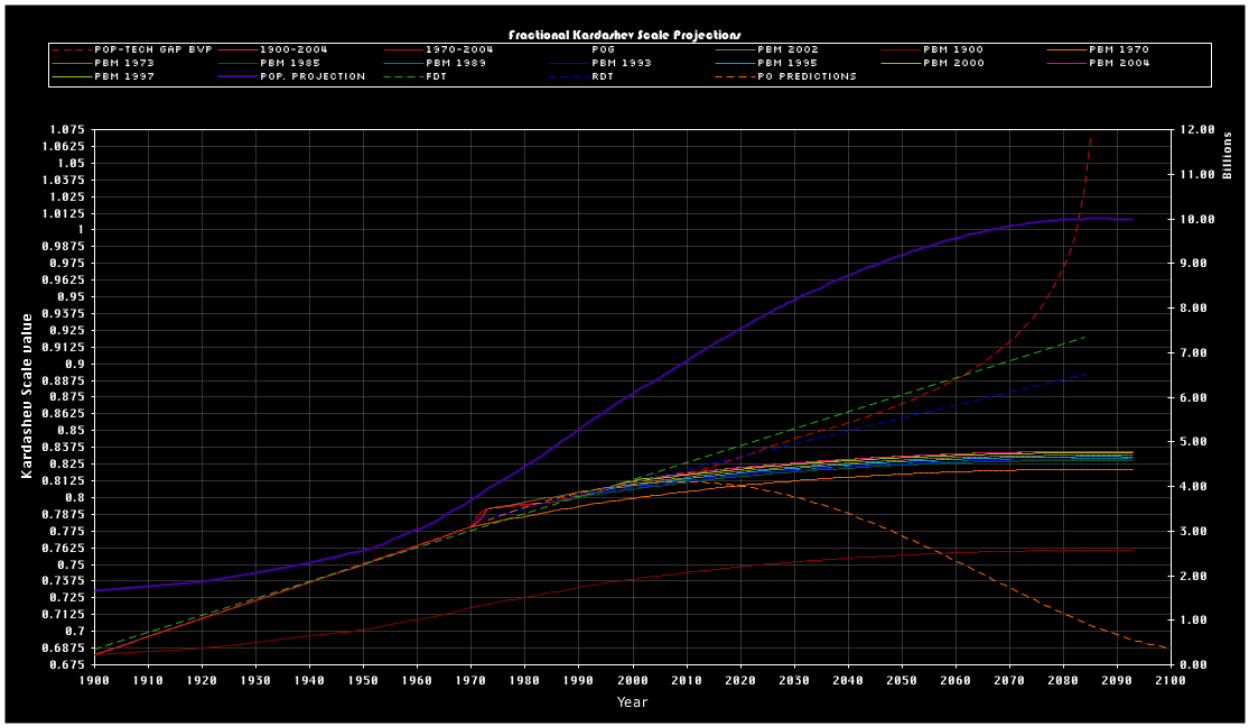

Figure 2: A series of Kardashev scale projections.

> they have correct answers that can be arrived at by reasoning and weighting up of evidence, could be more accurately answered by a superintelligence than by humans. The same holds for questions of policy and long-term planning; when it comes to understanding which policies would lead to which results, and which means would be most effective in attaining given aims, a superintelligence would outperform humans [2].

Perhaps, ultimately, we should just sit back and allow The Singularity choose the best moral philosophy for us, instead of the other way around.

References

[1] Issac Asimov. *Runaround*. Street & Smith, March 1942.

[2] Nick Bostrom. Ethical Issues in Advanced Artificial Intelligence. `http://www.nickbostrom.com/ethics/ai.html`, 2003.

[3] Singularity Institute for Artificial Intelligence. Creating Friendly AI. `http://www.singinst.org/ourresearch/publications/CFAI/`, 2001.

[4] Singularity Institute for Artificial Intelligence. What is Friendly AI? `http://www.singinst.org/ourresearch/publications/what-is-friendly-ai.html`, 2007.

[5] Singularity Institute for Artificial Intelligence. What is the Singularity? `http://www.singinst.org/overview/whatisthesingularity`, 2007.

[6] Bill Hibbard. Critique of the SIAI Guidelines on Friendly AI. `http://www.ssec.wisc.edu/~billh/g/SIAI_critique.html`, May 2003.

[7] Vernor Vinge. The Coming Techonological Singularity. `http://www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html`, 1993.

[8] Eliezer S. Yudkowsky. The Singularitarian Principles. `http://yudkowsky.net/sing/principles.html`, May 2001.

[9] Eliezer S. Yudkowsky. Staring Into the Singularity. `http://yudkowsky.net/singularity.html`, May 2001.