The Need For Friendly Artificial General Intelligence

Kevin Secretan

SOS 150

The word “Singularity” is overloaded: generally, the meaning implies a point of time in the future when an entity with greater-than-modern-human intelligence exists. However, there are three “schools” of thought that accompany this definition.[7] The first is called Accelerating Change, and is commonly advocated by Ray Kurzweil. Our normal human intuitions are primed to think that roughly the amount of change experienced in our past can be expected in the future, while modern times have uprooted that intuition because with modern technology, the rate of change increases exponentially (and this can be seen using graphs). The invention of the printing press caused a surge in printed materials, and as the printing press improved so did the amount of printed materials. The invention of the modern computer caused a surge in many, many fields, and as the computers get better those fields too get better.

The second school of thought in regards to the Singularity is called the Event Horizon. At the point something smarter than human arises, you can no longer predict its actions, or else you would be equally smart. This is illustrated by a chess game between an average player and a Grand Master. The average player can’t assign high probabilities to individual moves in the game, because if he could tell where the Grand Master would move he would be at least as skilled. However, and this brings in the third school of thought, the inexperienced player can assign a high probability to the final outcome, knowing the initial state that he is inexperienced and playing a Grand Master: the Grand Master will win. Thus the third school of thought contradicts the second, and states that if you know the initial state you can make a good guess at later outcomes. Furthermore, it contradicts the first, predicting faster-than-exponential growth after smarter-than-human intelligence moves in and invents molecular nanotechnology. This third school is what I will be discussing when I speak about the Singularity, and it is commonly called Intelligence Explosion, because when we have an intelligence (most likely a type of Artificial General Intelligence, though an upgraded human could potentially fit the role) capable of modifying itself, one of the things it will likely work on is improving its own intelligence. Even assuming the relatively weak stance that human thought is the highest quality of thought, at the very least it could be sped up. Something that thinks twice as fast as a human with the same quality effectively thinks a year's worth of thought for a human in only six months. Double that speed, and a year's subjective thinking happens in only three months. Double again, and again, and you get a huge amount of thinking power even if it's only at the quality of humans. There is every reason to believe human thought can not only be sped up, but also improved in quality. We suffer from cognitive biases that affect the clarity of our thought: simply removing these by not implementing them in a programmed intelligence would cause a higher quality in rational thought.[3]

With the Singularity drawing closer, so does the outcome for the future of humanity: we must create a Strong, Friendly, Artificial General Intelligence to ensure humanity survives the century without harm from global warfare, nanotechnology, UnFriendly AI, or other existential risks. I picked this topic because I can't think of anything more important for humanity right now, and while I'm fairly confident I'm not smart enough to contribute much in solving the problems involved, I nevertheless feel I can get a feel for them and spread the word.

Imagine a power[6] that does all of the following: it can kill the largest of animals on Earth, it can create and control fire, it can shift water and rock around and create ovens to melt other kinds of rock into different shapes, it can make shelters for surviving cold climates, it can create high pyramids and high towers, it can pave roads, tame other larger beasts to use on the roads, it can treat many different diseases and ailments, it can record history and have a memory longer than any other life-form, it can cause a radioactive nuclear explosion and obliterate land, it can escape the Earth and reach the moon, it can create machines that operate on electrical impulses to perform simple binary logic that can do anything from addition to manipulating molecules. It is a single thing that can do all these tricks, and it is just behind your eyes. Our squishy brain is the singular reason for all of humanity's accomplishments. We are Intelligent, as a species, and the dumbest of us are far superior to the smartest of other species. When people think of intelligence, they think of Einstein instead of simply a human being. Our history isn't composed of Einsteins, yet we've come so far regardless: clearly being human is what intelligence means, rather than being on top of an IQ scale that only applies to humans, and only loosely.

The universe has been around for billions of years, and the most complicated thing it has created by chance that we have seen must have been the first replicator, either here on Earth or from somewhere else that made it to Earth. After that, evolution took control:[5] as the replicator replicated, errors were introduced, but not all bad and over time it happened that the replicators which were still around generally had characteristics which helped it stay that way. In a much shorter time scale than it took to create the first replicator, evolution produced us: humans. We are not particularly useful at singular tasks like other animals, such as being poisonous or very strong or having armor. But we do have our intelligence, our one trick that lets us do far more than any other creature thus far. In a mere ten thousand years or so of intelligence-driven technology, taking us a thousands of years just to reach that point, we have become the ultradominant species on this planet, and our greatest threat to survival does not come from a bigger predator, it comes from ourselves. We do not need to wait for evolution to give us wings to compete with the flying creatures; we can discover the power of flight ourselves and accomplish it much faster.

We are a powerful optimization process[1], able to see configurations of matter and make them, instead of simply letting chance dictate where the particles go. We have done so much in our short time as a species. Looking over at the sum of human achievement, this is the power of intelligence brought about by a process not itself intelligent. Now, my thesis proposes we use our intelligence to create a new, better intelligence, and if we do it right we can hope it will also create better intelligences, and take us along for the ride. If in a mere ten thousand years the products of human intelligence have come this far, imagine what an intelligence thousands of times more powerful than our own could do in the same time scale. We can only imagine, but it would be great. Keeping with the third school of thought, an AI, properly constructed, would not simply stick around at human-level intelligence. It would far surpass us, going FOOM, and if we're not careful it could even destroy us. Making an Artificial Intelligence isn't about making an immortal Einstein-bot: it's about creating a new species that makes the difference between human and the AI similar to the difference between an ant and a human, for that is the power of intelligence between species. What we are close to accomplishing will determine if the last ten thousand years of human history have been at all significant. Will we create something that will resound through the universe, or will we simply fade from existence? Friendly AI is the most important problem we have, and it's important not to forget the Friendly part.

When first exposed to the idea of an AI, people think "Oh, like in The Matrix?" or Terminator or some other popular movie they may have seen. The standard reply is "No, and I try to avoid the logical fallacy of generalizing from fictional evidence."[10] Sure, it's possible that we might create an AI barely above human level, that wants to wipe us out, and engages in a war with us... But it's not very likely. The most likely scenario is we create an AI that neither loves humanity nor hates it, but sees humans as forms of matter, like the rocks, that it can use for other purposes. A more likely scenario, as written by Nick Bostrom, to any of the movies is "humanity suddenly going extinct without warning and without being replaced by some other civilization." The humans-vs-robots story is just that, a story. It is best not to use it as a substitute for thinking. On a related note, when confronted with the problem of having a Friendly AI, people again use movies for thinking: just program it with Asimov's three laws of robotics! Or some other wording that involves hard-coding the morals of the programmers into the AI. Imagine for a moment the ancient Greeks creating an AI, and they programming their morals into it. Among other things would be a general disrespect for women; what this is getting at is it is absurd to think humanity has finished its path to morality and found the one true way. Considering another case, imagine an AI that is programmed to value smiling humans. So it captures everyone and ends up surgically altering our faces into perpetual smiles at the least, or simply replaces everyone's head with a giant Wal-Mart smiley-guy. Friendliness isn't a simple problem to solve, but if we don't get it right we might as well not build an AGI in the first place.

Ray Kurzweil is a little more skeptical of the AI approach to the Singularity,[4], and I suspect he thinks reverse-engineering the brain will come first, but he admits that AI work and brain work all tie together into understanding intelligence, and incremental progress in both fields helps. Eliezer Yudkowsky believes what we lack is a piece of knowledge that's not necessarily large, but it involves the mathematical structure of intelligence rather than the particular way neuron patterns give rise to it. For all we know, someone could have cracked the problem right now and unleash an AI next week which subsequently destroys the world after it develops nanotechnology. Kurzweil is not blind to the risks of both AI or uploaded intelligence, however, especially when nanotechnology is concerned. One of the overlooked dangers of nanotech is in brute-forcing advanced intelligence through the increased computing power available, which brings us back to the problem of unFriendly AI. It is something that ought to be guarded against, rather than just nanotech viruses that Kurzweil claims to be working on preventing.

This Friendliness business still sounds fairly mysterious, are there any proposals that don't immediately fall flat on their faces? I read about Coherent Extrapolated Volition[9], which romantically states humankind's extrapolated volition is "our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted." An example is given of an AI possessing two boxes, A and B, one with a diamond and one with nothing. A human, Fred, asks for box B, which the AI knows to be empty. Would a Friendly AI give Fred box B, or would it extrapolate that if Fred knew more, he would have rather picked box A, since what Fred really wants is the diamond? It is safe bet that the latter is the case, and were the AI a human friend it would likely give Fred the box containing the diamond.

In addition to this system, we want the AI to maintain it throughout its existence no matter what sort of intelligence upgrades it goes through. We want a system where the AI could consider an upgrade that made it desire to wipe out humanity, but not take it, because the current AI does not wish that to happen. If someone offered Gandhi a pill that made him want to kill people, Gandhi would not take that pill because the current version does not want to kill people, and does not want himself to kill people in the future either. The tricky part to this Friendliness business is of course the math. We can't just give an AI clever-sounding arguments and expect it to follow the rules: we have to prove that its goal architecture will remain consistent and stable. Bayesian

Probability and Decision Theory are two areas where this can be expected to happen.

On the intelligence side of things, Marcus Hutter has a thesis about a method of implementing intelligence he calls AIXI.[2] In it he attempts to develop a foundation in math for Artificial Intelligence, and touches upon developed areas where a rational agent can reason and act optimally in any situation. He draws from reinforcement learning, algorithmic information theory, Kolmogorov complexity, computational complexity theory, information theory and statistics, Solomonoff induction, Levin search, sequential decision theory, adaptive control theory, and others. The math is extensive and I won't try to go into it here, but it nonetheless is an actual proposal to consider that's not simply along the lines of "Well, someone will build it!" Hutter does wave the issue of Consciousness aside, stating that while it is philosophically interesting it isn't really relevant from a practical point of view. Whether an intelligent agent is "conscious" or not simply doesn't matter; only the intelligence matters.

Similarly, Eliezer Yudkowsky has a book called Creating Friendly AI[8] where he delves into the mathematical side, among other things showing how Bayesian Reasoning can be used to create a stable upgrade path. I haven't yet managed to really get into this book, but it is nonetheless fascinating to browse over and there is a simple FAQ section accessible to the lay-person that contains some things you will find in the book. For example, a simple high-level view of Friendliness is given: "Friendliness: The set of actions, behaviors, and outcomes that a human would view as benevolent, rather than malevolent; nice, rather than malicious; friendly, rather than unfriendly. An AI that does what you ask, as long as it doesn't hurt anyone else; an AI which doesn't cause involuntary pain, death, alteration, or violation of personal space. Unfriendly AI: An AI that starts killing people." He admits there is also a "core of unknowability at the center of the Singularity", such that there are things "beyond our ability to anticipate, not in the way Socrates couldn't have anticipated nanotechnology, but in the way that a dog couldn't have anticipated nanotechnology." I think that phrase captures just how hard the problem is and what sort of things might be out there we can't even comprehend: the lower intelligences on this planet have no idea that the sun will swell up in a few billion years and scorch the Earth, who knows what problems we are equally ignorant about in the universe? He finishes with ”In striving to create an AI, we are not striving to create a predictable tool. We are striving to create a messenger to send on ahead to find humanity’s destiny, and the design requirement is that we handle any problem, *any* philosophical crisis, as well and as altruistically as a human.” This problem isn’t just about creating a utopia: it’s about continuing the human drive of seeking out knowledge and whether we’ll make it off this rock or not.

Through further research I’ve come to appreciate how much human-level intelligence means and the implications of something higher. It seems hard to get people to realize that AI is not about slavery, it’s about progress. It is also heartening to see some real proposals on the subject rather than politicking fluff. Though there is secrecy involved: building a Friendly AI for example is about building an AGI in a particular way, and if the knowledge to build the AGI without the Friendliness part became known, humanity can kiss its chances goodbye. In a personal blog Eliezer related how in his youth he once thought AI should be achieved as soon as possible, and didn’t initially consider the Friendliness problem. He has since had a “Bayesian Enlightenment” (in his own words) and realized the necessity of proving Friendliness before writing any real code. A lot of people shout out “Show me the source!”, even if they wouldn’t understand it, just because it would be comforting to know some progress is being made. But Eliezer has stated that once it is known how to build an AGI, actually programming it won’t be the hard part. He is against the idea of trying to brute-force the program, but he isn’t too concerned yet because it’s such a hard problem.

I believe the spirit of my sources and the math and science they cite, though I’m not convinced (and I don’t think the authors are fully convinced) that the proposed solutions are the One True Way to go. Certain fundamental aspects such as the use of Bayesian reasoning seem like they will be used in the final product, but overall there is still that piece of knowledge missing and it could be 5 years or 5 decades before we have it. My own estimates do not extend past that due to Kurzweil’s predictions, since if we don’t get AI in this century we’re at least likely to get nanotech, and while that has many benefits it is not comforting knowing modern-day humans with their high fallibility will be controlling it. Nanotechnology has the power to completely exterminate all life on this planet, and it doesn’t even have to be maliciously intended.

If more people learn about the goal of Friendly AI creation, the sooner we can improve this world. We can look at more seriously solving problems that most people don’t even realize are problems, such as death. We can get rid of the insane death rate of 1.8 people per second. In a few billion years when the sun expands, we can expect to still exist (if in a different form) and expand out of this solar system. The future of humanity is at stake, and the more people that realize it the better. We need our best minds working on this problem, not on String Theory. And for those of us who can’t contribute intellectually, because it is a very difficult problem, we can still contribute financially: if you want to help, find the best thing you are able to do, do it, and then contribute your earnings. It’s baffling that athletes can be paid millions of dollars to throw balls around, while scientists in general doing humanity a true service often struggle with funding. One of my goals with this paper is to educate on a high-level view, that this is an important problem, that a lot of our initial intuitions about it are wrong, and that there are real serious proposals to it and it is not unsolvable.

Bibliography

[1] Nick Hay. The stamp collecting device. URL: http://www.singinst.org/blog/2007/06/11/the-stamp-collecting-device/.

[2] Marcus Hutter. Universal algorithmic intelligence: A mathematical top→down approach. In B. Goertzel and C. Pennachin, editors, *Artificial General Intelligence*, Cognitive Technologies, pages 227–290. Springer, Berlin, 2007. URL: http://www.hutter1.net/ai/aixigentle.htm.

[3] Daniel Kahneman, Paul Slovic, and Amos Tversky. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, 1982. URL: http://www.amazon.com/Judgment-under-Uncertainty-Heuristics-Biases/dp/0521284147/.

[4] Surfdaddy Orca R.U. Sirus. Ray kurzweil: The h+ interview. *H+ Magazine*, 2009. URL: http://hplusmagazine.com/articles/ai/ray-kurzweil-h-interview.

[5] Robert Wright. *The Moral Animal: Why We Are, the Way We Are: The New Science of Evolutionary Psychology*. Vintage, 1995. URL: http://www.amazon.com/Moral-Animal-Science-Evolutionary-Psychology/dp/0679763996.

[6] Eliezer Yudkowsky. The power of intelligence. URL: http://yudkowsky.net/singularity/power.

[7] Eliezer Yudkowsky. Three major singularity schools. URL: http://yudkowsky.net/singularity/schools.

[8] Eliezer Yudkowsky. Creating friendly ai, 2001.

[9] Eliezer Yudkowsky. Coherent extrapolated volition. URL: http://singinst.org/upload/CEV.html, 2004.

[10] Eliezer Yudkowsky. The logical fallacy of generalization from fictional evidence. URL: http://lesswrong.com/lw/k9/the_logical_fallacy_of_generalization_from/, 2007.