Friendly AI : A Dangerous Delusion?

Prof. Dr. Hugo de GARIS
profhugodegaris@yahoo.com

Abstract

This essay claims that the notion of "Friendly AI" (i.e. the idea that future intelligent machines can be designed in such a way that when they modify themselves into even greater levels of intelligence, they will remain friendly to human beings) is not only impossible, but a dangerous delusion. Reasons for its impossibility are given, as well as explaining why the notion is not only deluded but dangerous.

1. Introduction

I'm known for predicting that later this century there will be a terrible war, killing billions of people over the issue of species dominance, i.e. whether human beings should build artilects (artificial intellects), which could become so vastly superior to human beings in intellectual capacities, that they may end up treating us as grossly inferior pests and wiping us out.

To combat this horrible scenario, the Singularity Institute (http://singinst.org) in Silicon Valley has been set up to ensure that the above scenario does not occur. The Institute's principle theorist, Eliezer Yudkovsky, has coined

the term “Friendly AI” which he defines more or less as given in the abstract.

He claims (I think correctly) that trying to prevent artilects from wiping out humanity is the most important thing on humanity’s agenda this century. He hopes that he and others will be able to prove mathematically that it is possible to design an intelligent machine that (of logical mathematical necessity, given its design) will be forced to remain human friendly, as it redesigns itself into ever higher levels of intelligence.

I will present a set of arguments which I think refute this vision, and then comment on the political desirability (or otherwise) of this vision.

2. The Arguments against “Friendly AI”

Let me start by assuming that Friendly AI can be implemented. Then the next question is whether humanity would be unanimous about implementing it. In other words (for this case) “Does can imply ought?” I think that the more ardent of the Cosmists (the ideological group of people who want to build artilects, and see themselves as “god builders”) will argue that their main goal is to build artilects that are trillions of trillions of times above human intelligence levels, who would be immortal, thinking a million times faster than humans, with unlimited memory, who could change their shape and architecture in milliseconds, could venture out into the cosmos, etc. These

Cosmists would prefer that the artilects be built even if human beings get wiped out as a result. If making them according to “Friendly AI” designs inhibits or even blocks their path to achieving their godlike capabilities, then the Cosmists will want the artilects not to be made “AI Friendly”.

Hence even if “AI Friendly” designs can be created, it does not automatically follow that they will be universally applied. The more ardent Cosmists might go underground to build the artilects the way they want, and “to hell with humanity.” The Cosmists have a slogan “One artilect is worth a trillion trillion human beings!”

On the other hand, if AI Friendly designs are impossible to make, then there is no point in discussing whether they should be implemented or not.

I will now present some arguments which claim to show that the notion of Friendly AI is impossible.

a) The “Evolutionary Engineering” Argument

Ask yourself, “How is it possible for a creature of a given intelligence level, to be able to design a creature of greater intelligence?” To be able to design a creature of superior intelligence requires a level of intelligence that the designer simply does not have. Therefore it is logically impossible to use the traditional blue-print design approach to create a creature of superior intelligence.

For example, my good friend Ben Goertzel has written a book recently called “Building Better Minds” in which he lays out a humanly conceived (i.e. by himself) plan to build a (near) human level intelligence. He will only be able to go so far with such an approach. There will be limits to the ingenuity of his plan/design due to the intellectual limits of Ben Goertzel. So how can such limits be overcome?

Human beings have been building superior intelligences for thousands of generations, by having sex. Their children often grow up to be smarter than they are. So how to explain that? Well, by shuffling the genes. When the genes of the mother mix with the genes of the father, and only one of each mother/father pair of genes is used, it is possible, by blind luck to arrive at a DNA blueprint that builds an intellectually superior child. But there are limits to this process as well. It gets statistically harder and harder to generate ever higher intelligence. For example, the odds of creating an Ed Witten are one in a billion.

So, how did modern homo sapiens come into being? How did nature build us over millions of years? It did so by using evolutionary engineering – i.e. by selecting genes with superior fitness levels due to random mutations of DNA. This slow, blind process has resulted in us, and is very probably the ONLY approach humans will have to build machines a lot smarter than we are.

But, if we use evolutionary engineering to build for example, artificial neural networks, for our artilects, then the complexity levels of these networks will be so great,

that we are unable to understand them. They will be a black box.

One of the reasons I stopped my brain building work, was that I got bored evolving neural net modules for artificial brains. These modules were a black box to me. They worked, because they were evolved, but I had no scientific understanding as to why they worked. I was doing great engineering, but lousy science. After 20 years of it, I finally got fed up and turned to other research topics that taxed my own biological human brain more (i.e. pure math and mathematical physics).

Let us assume that the evolutionary engineering approach is the only way to create creatures of higher intelligence levels than human beings, and that the complexity levels of the evolved brain circuits is too complex for humans to understand. Then we would not be able to predict the attitudes and behavior of these creatures towards us. The only way to know how they would behave towards us would be to build them, but then its too late. They would then exist and might choose to wipe us out.

Hence with the above logic, we are faced with a dilemma. Either we limit ourselves to humanly designed blue prints for intelligent machines, that are INcapable of reaching super human intelligence levels, OR, we use an evolutionary engineering approach that could attain super human intelligence levels. If we use an evolutionary engineering approach, we cannot be sure the resulting artilects would be human friendly.

b) The "Cosmic Ray" Argument.

It is almost certain that the circuitry that will be used to create intelligent machines will be nanotech based. For example, to build a near human level artificial brain that is not the size of a room, will necessitate the use of nanoscale components. Even if "Friendly AI" nanocircuits could be built, they would then be subject to the random mutations generated by impacting cosmic rays, that can be very energetic, zapping the nanocircuits in random ways, and generating "rogue artilects". Nature would be doing the same kind of evolutionary engineering as the human kind mentioned above. Since these mutations would be random, their consequences on the behavior and attitudes of the artilects towards human beings would be unpredictable. Hence even if the initial, unmutated nanocircuits could be made human friendly, they would not stay that way.

c) The "Asimov Naïve" Argument

Isaac Asimov, the science fiction writer, is famous for his "Three Laws of Robotics" which were intended to ensure that the robots in his stories remained "Human Friendly", for example, the robots were not allowed to harm humans, nor allow humans to be harmed. One can imagine fairly readily that it is probably possible to program robots in a conventional way to behave like this, with the proviso, that the robots are less intelligent than their human programmers. But, once the robots become smarter than humans, they would be able to examine their circuitry,

detect the humanly created parts, find them “moronic”, and delete them, if they want. Hence Asimov’s “Three Laws” cannot help us. They are naïve. Forget Asimov.

3. Friendly AI is a Dangerous Delusion

Hopefully, the above arguments have convinced you that the notion of “Friendly AI” is a delusion. But why might it be seen as a *dangerous* delusion?

If the future politicians who have to decide whether to legislate or not against building artilects of super human intelligence believe that “Friendly AI” robots can be built, then they will be much more likely not to legislate against their construction. On the other hand, if they learn that the artificial brain building community has a consensus view that “Friendly AI” is IMpossible, then they will be far more hesitant.

If “Friendly AI” is indeed impossible, then humanity has a much TOUGHER choice to make, namely (in the form of a slogan of the Cosmists) “Do we build gods, or do we build our potential exterminators?” Spelling this out, humanity will then be forced to choose between building godlike artilects and risking that humanity gets wiped out, OR not building artilect gods, and seeing humanity survive. The first option is *specicide*. The second option is *deicide*. This choice will be the toughest that humanity will ever have to make.

If the pro “Friendly AI” people can persuade the politicians in the coming decades to go ahead with artilect building on the assumption that Friendly AI is valid, then if it is *not* valid, then it is a dangerous delusion, because the politicians may then give the green light to the artilect builders to build artilects that were thought to be “human friendly” but in reality turn against us and wipe us out.