The 21st. CENTURY ARTILECT

Moral Dilemmas Concerning the Ultra Intelligent Machine

Dr. Hugo de Garis
Brain Builder Group
ATR, Kyoto, Japan

written May 1989

appeared in
Revue Internationale de Philosophie, 1990
(scanned version) p1 p2 p3 p4 p5 p6 p7 p8 p9

ABSTRACT

Within one to two human generations, it is likely that computer technology will be capable of building brain-like computers containing millions if not billions of artificial neurons. This development will allow neuroengineers and neurophysiologists to combine forces to discover the principles of the functioning of the human brain. These principles will then be translated into more sophisticated computer architectures, until a point is reached in the 21st. century when the primary global political issue will become, "Who or what is to be dominant species on this planet - human beings, or artilects (artificial intellects)?"

A new branch of applied moral philosophy is needed to study the profound implications of the prospect of life in a world in which it is generally recognised to be only a question of time before our computers become smarter than we are. Since human beings could never be sure of the attitudes of advanced artilects towards us, due to their unfathomable complexity and possible "Darwinian" self modification, the prospect of possible hostilities between human beings and artilects cannot be excluded.

KEYWORDS

Artilects (Artificial Intellects), Ultra Intelligent Machine, Neuro-Engineering, Dominant Species, Artificial Neuron.

1. INTRODUCTION

A revolution is taking place in the field of Artificial Intelligence. This revolution, called "Connectionism", attempts to understand the functioning of the human brain in terms of interactions between artificial abstract neuron-like components, and hopes to provide computer science with design principles sufficiently powerful to be able to build genuine artificial electronic (optical, molecular) brains (KOHONEN 1987,McCLELLAND et al 1986, MEAD 1987). Progress in micro electronics and related fields, such as optical computing, has been so impressive over the last few years, that the possibility of building a true artilect within a human generation or two becomes a real possibility and not merely a science fiction pipe dream.

However, if the idea of the 21st century artilect is to be taken seriously (and a growing number of Artificial Intelligence specialists are doing just that (MICHIE 1974, WALTZ 1987, de GARIS 1989), then a large number of profound political and philosophical questions arise. This article addresses itself to some of the philosophical and moral issues concerning the fundamental question "Who or what is to be dominant species on this planet - human beings or the artilects?"

2. A MORAL DILEMMA

In order to understand the disquiet which is growing amongst an increasing number of intelligists (specialists in Artificial Intelligence) around the world in the late 1980s (WALTZ 1987, de GARIS 1989), it is useful to make a historical analogy with the development of the awareness of the nuclear physicists in the 1930s, of the possibility of a chain reaction when splitting the uranium atom. At the time, that is, immediately after the announcement of the splitting, very few nuclear physicists thought hard about the consequences to humanity of life in a nuclear age and the possibility of a large scale nuclear war in which billions of human beings would die.

Some intelligists feel that a similar situation is developing now with the connectionist revolution. The intelligists concerned, are worried that if the artificial intelligence community simply rushes ahead with the construction of increasingly sophisticated artilects, without thinking about the possible long term political, social and philosophical consequences, then humanity may end up in the same sort of diabolical situation as in the present era of possible nuclear holocaust.

Within a single human generation, computer scientists will be building brain-like computers based on the technology of the 21st century. These true "electronic (optical, molecular) brains" will allow neurophysiologists to perform experiments on machines instead of being confined to biological specimens. The marriage between neuro-engineers and neuro-physiologists will be extremely fruitful and artificial intelligence can expect to make rapid progress towards its long term goal of building a machine that can "think", a machine usually called an "artificial intelligence", or "artilect".

However, since an artilect is, by definition, highly intelligent, (and in the limit, ultra intelligent, that is, having an intelligence which is orders of magnitude superior to ours), if ever such a machine should turn against humanity, it could be extremely dangerous. An atomic bomb has the enormous advantage, from the point of view of human beings, of being totally stupid. It has no intelligence. It is human beings who control it. But an artilect is a different kettle of fish entirely.

Artilects, unlike the human species, will probably be capable of extremely rapid evolution and will, in a very short time (as judged by human standards), reach a state of sophistication beyond human comprehension. Remember, that human neurons communicate at hundreds of meters per second, whereas electronic components communicate near the speed of light, a million times faster. Remember, that our brains, although containing some trillion neurons, has a fixed architecture, as specified by our genes. The artilects could choose to undertake "Darwinian experiments" on themselves, or parts of themselves, and incorporate the more successful results into their structure. Artilects have no obvious limit as to the number of components they may choose to integrate into themselves. To them, our trillion neurons may seem puny.

Not only may artilects be superior to humans in quantitative terms, they may be greatly our superiors in qualitative terms as well. They may discover whole new principles of "intelligence theory" which they may use in restructuring themselves. This continuous updating may grow exponentially - the smarter the machine, the better and faster the redesigning phase, so that a take-off point may be reached, beyond which, we human beings will appear to artilects as mice do to us.

This notion of Darwinian experimentation is important in this discussion, because it runs counter to the opinions of many people who believe (rather naively, in my view) that it will be possible to construct artilects which will obey human commands with docility. Such machines are not artilects according to my conception of the word.

I accept that machines will be built which will show some obvious signs of real intelligence and yet remain totally obedient. However, this is not the issue being discussed in this paper. What worries me is the type of machine which is so smart that it is capable of modifying itself, of searching out new structures and behaviours, that is, the "Darwinian artilect".

Since any machine, no matter how intelligent, is subject to the same physical laws as is any other material object in the universe, there will be upper limits to the level of self-control of its intellectual functions. At some level in its architectural design, there will be "givens", that is, top level structures determining the artilect's functioning, which are not "judged" by any higher level structures. If the artilect is to modifiy these top level structures, how can it judge the quality of the change? What is meant by quality in such a context?

This problem is universal for biological systems. Quality, in a biological context, is defined as increased survivability. Structural innovations such as reproduction, mutation, sex, death, etc., are all "measured" according to the survivability criterion. It is just possible that there may be no other alternative for the artilect, than taking the same route. Survivability however, only has meaning in a context in which the concept of death has meaning. But would not an artilect be essentially immortal, as are cancer cells, and would a fully autonomous artilect, resulting from an artilectual reproductive process, but with modified structures, accept being "terminated" by its "parent" artilects, if the latter consider the experiment to have been a failure?

If the offspring artilects do not agree to being "killed", they might be allowed to live, but this would imply that every artilect experiment would create a new immortal being, which would consume scarce resources. There seem to be at least three possible solutions to this problem. Either a limit is placed on the number of experiments being performed, a philosophy inevitably leading to evolutionary stagnation, or artilects are replaced by newer versions, (processes called reproduction and death, in biological contexts), or the growing population of artilects could undertake a mass migration into the cosmos in search of other resources.

This Darwinian modification is, by its nature, random and chancy. The problem for human beings is that an artilect, by definition, is beyond our control. As human beings, with our feeble intellects (by artilectual standards), we are unable to understand the implications of structural changes to the artilect's "brain", because this requires a greater intellect than we possess. We can only sit back and observe the impact of artilectual change upon us. But this change may not necessarily be to our advantage.

The "moral circuits" of the artilects may change so that they no longer feel any "sympathy" for human beings and decide that, given a materials shortage on the planet, it might be advisable, from an artilectual point of view, to reduce the "ecological load" by removing the "hungriest" of the inferior species, namely human beings.

Since human moral attitudes, like other psychological attitudes, are ultimately physical/chemical phenomena, human beings could not be sure of the attitudes of artilects towards human beings, once the artilects had evolved to a highly advanced state. What human beings consider as moral is merely the result of our biological evolution. As human beings we have no qualms about killing mosquitos or cattle. To us, they are such inferior creatures we do not question our power of life or death over them.

This uncertainty raises the inevitable fear of the unknown in human beings. With artilects undertaking experiments to "improve" themselves (however the artilects define improvement), we humans could never be sure that the changing intelligences and attitudes of the artilects would remain favourable to us, even if we humans did our best to instil some sort of initial "Asimovian", human-oriented moral code into them. Personally, I believe that Asimov's "Three Laws of Robotics" are inappropriate for machines making random changes to themselves to see whether they lead to "improvements". Asimov's robots were not artilects.

3. A WORLD DIVIDED

With many intelligists agreeing that it will be technologically possible to build electronic (optical, molecular) brains within a human generation or two, what are the moral problems presented to humanity, and particularly to applied moral philosophers? The biggest question in many peoples minds will be, "Do we, or do we not, allow such artilects to be built?" Given the time frame we are talking about, namely 20 to 50 years from now, it is unlikely that human societies will have evolved sufficiently to have formed a world state, having the power to enforce a world wide ban on artilectual development, beyond an agreed point. What will probably happen, is that military powers will argue that they cannot afford to stop the development of artilects, in case the "other side" creates smarter "soldier robots" than themselves. Military/political pressures may ensure artilect funding and research until it is too late.

The artilect question alone, is sufficient in itself, to provide a very strong motivation for the establishment of a skeleton world government within the next human generation. With the rapid development of global telecommunications and the corresponding development of a world language, the establishment of a skeleton world government within such a short time may not be as naive as it sounds.

For the purposes of discussion, imagine that such a ban, or at least a moratorium, on artilect development is established. Should such a ban remain in force forever? Could one not argue that mankind has not only the power, but the moral duty to initiate the next major phase in evolution, and that it would be a "crime" on a universal or cosmic scale not to exercise that power?

One can imagine new ideological political factions being established, comparable with the capitalist/communist factions of today. Those in favour of giving the artilects freedom to evolve as they wish, I have labelled the "Cosmists", and those opposed, I have labelled the "Terras" (or Terrestrialists). I envisage a bitter ideological conflict between these two groups, taking on a planetary and military scale.

The Cosmists are so named because of the idea that it is unlikely, once the artilects have evolved beyond a certain point, that they will want to remain on this provincial little planet we call Earth. After all, there are some trillion trillion other stars to choose from. It seems more credible that the artilects will leave our planet and move into the Cosmos, perhaps in search of other ultraintelligences.

The Terras are so named because they wish to remain dominant on this planet. Their horizons are terrestrial. To the Cosmists, this attitude is provincial in the extreme.

To the Terras, the aspirations of the Cosmists are fraught with danger, and are to be resisted at any cost. The survival of humanity is at stake.

There may be a way out of this moral dilemma. With 21st century space technology, it may be entirely feasible to transport whole populations of Cosmist scientists and technicians to some distant planet, where they can build their artilects and suffer the consequences. However, even this option may be too risky for some Terran politicians, because the artilects may choose to return to the Earth, and with their superior intellects, they could easily overcome the military precautions installed by the Terras.

4. SUMMARY

This article claims that intelligists will be able to construct true electronic (optical, molecular) brains, called artilects, within one to two human generations. It is argued that this possibility is not a piece of science fiction, but is an opinion held by a growing number of professional intelligists. This prospect raises the moral dilemma of whether human beings should or should not allow the artilects to be built, and whether artilects should or should not be allowed to modify themselves into superbeings, beyond human comprehension. This dilemma will probably dominate political and philosophical discussion in the 21st century. A new branch of applied moral philosophy needs to be established to consider the artilect problem.

5. REFERENCES

(de GARIS 1989)"What if AI Succeeds? The Rise of the Twenty-First Century Artilect", Artificial Intelligence Magazine (cover story), Summer 1989

(EVANS 1979) "The Mighty Micro", Coronet Books.

(JASTROW 1981) "The Enchanted Loom", Simon & Schuster, New York.

(KELLY 1987) "Intelligent Machines. What Chance?", Advances in Artificial Intelligence, Wiley.

(KOHONEN 1987) "Self-Organization and Associative Memory", 2nd edn. Kohonen T., Springer-Verlag, Berlin, Heidelberg.

(McCLELLAND et al 1986) "Parallel Distributed Processing", Vols 1 and 2, McClelland J.L. & Rumelhart D.E. (Eds), MIT Press, Cambridge, Mass.

(McCORDUCK 1979) Forging the Gods, "Machines Who Think", Freeman.

(MEAD 1987) "Analog VLSI and Neural Systems", Mead C., Addison Wesley, Reading, Mass.

(MICHIE 1974) "On Machine Intelligence", Michie D., Edinburgh University Press, Edinburgh.

(WALTZ 1987) "The Prospects for Building Truly Intelligent Machines", Waltz D., Thinking Machines Corporation, Cambridge, Mass.

Dr. Hugo de Garis,
Head, Brain Builder Group,
Evolutionary Systems Department,
ATR Human Information Processing Research Labs,
2-2 Hikaridai, Seika-cho, Soraku-gun,
Kansai Science City, Kyoto-fu, 619-02, Japan.
tel. + 81 774 95 1079,
fax. + 81 774 95 1008,
email. degaris@hip.atr.co.jp
web. http://www.hip.atr.co.jp/~degaris