Several Arguments For and Against Superintelligence/"Singularity"

Ionuț ISAC, *Senior researcher, PhD*
Department of Philosophy, Institute of History "G. Barițiu"
Romanian Academy Branch Cluj-Napoca
isac.ionut@cluj.astral.ro

Abstract

We are only at the dawn of a technological revolution in informatics, robotics and computer sciences. However, we try to imagine how our world will look years, decades and centuries after. In this respect, one of the boldest ideas ever advanced by researchers is that of singularity, understood as the result of a very sudden and fast technological progress, leading humankind to the possibility of building a supposedly more-intelligent-than-humanity "almighty" machine. Such an extremely complex technical system endowed with an enormous potential is actually seen as a possible solution to humanity's most difficult problems (i.e. as an entity capable of forever solving issues, in view of the best desirable future of homo sapiens*).*

But how could one sustain this position? Among expressed fears and desires, exercises of imagination and speculations of all kind, many arguments have been formulated for and against the rise of superintelligence/singularity, that deserve a serious discussion. The purpose of this paper is to comment on several of them, according to some positions already implicitly or explicitly affirmed. In our view, the subject of singularity is able to rise from a simple scholar talk up to the highest levels of ontological and philosophical analysis. Thus, the paper advances and supports the thesis that, from the point of view of the nowadays philosophy of technology, one is compelled to rethink Kant's antinomies, rephrased according to the subject in discussion: the "singularity" is possible (and, consequently, will emerge) – the "singularity" is not possible (and, consequently, it will not emerge).

Keywords: *singularity, machines, superintelligence, evolution, trans-humanism.*

1. Let us begin with the following idea: theoretically, if a human-built machine could be brought to bear greater problem-solving and inventive skills than humans, then it may be able to design a yet more capable machine. If built, this "more-capable-machine" then could design a machine of even greater capability

(and so on). This iteration could accelerate, leading to a "recursive self-improvement", i.e. to an "intelligence explosion" (I. J. Good).

Firstly, we have to say that there is no certainty that such a machine, once reaching a very high degree of intelligence, complexity and speed of its actions, would still be capable or willing to design a different machine, "better" than itself. To sustain such an idea would be nothing else than applying pure induction, inspired by the assumption that the "classy" generations of intelligent machines must aspire for "perfection" as their supreme goal. Can one be sure of that? And what would be the meaning of such a projection? Once we know, in principle, how machines do work nowadays, as well as how humans use to cope with "better-and-better", it is much harder to yield the road to such a simplistic overview. Who could guarantee us that, for instance, maybe because of some inherent limits of our own, we are still not aware of some reasons of self-protecting on behalf of which the aforementioned "utmost-evolved-machine" would rather be tempted to stop itself somewhere in the process of "recursive self-improvement"? Consequently, a very intelligent machine may decide to multiply "in itself and by itself", mainly at the same level of complexity already acquired, anticipating its evolution in "small steps", according to the area of "problem-solving" within a paradigm.

Secondly, once having reached an outstanding level of intelligence, creativity and action, those machines might also decide to further create and develop some not "superior" but, on the contrary, rather "inferior" machines (however much more intelligent than humans), for the purpose of reserving for themselves an unassailable pre-eminence in the world for an unknown period of time (most probably, as long as possible). It may occur that those machines would not be willing to expose or endanger their outstanding place inside the whole of the existence; or, if once having decided to build a machine "more capable" than them, this could mean exactly as to design their future disappearance. Nothing can prevent us from imagining that those "classy" machines would prefer to communicate with their inferior 'mates' as well as with "accompanying" humans in terms of "lower" knowledge, keeping the "supreme" truths and axioms just for their own benefit, with no direct implication toward their alleged interest on possible extinction of *human sapiens*.

In this respect, one must rethink the metaphysical system of the Romanian thinker Lucian Blaga (1895-1961), whom develops a very peculiar and long ranging metaphysical explanation, starting with a high-level hypothesis on the nature of existence: i.e. the concept of the "Great Anonymous" with its "transcendent censorship". The "Great Anonymous" denotes an entity "centre" or

the "core" of transcendence. (Blaga stated that this is just a possible name, and that one could easily find others; what is essential is to refrain from interpreting it anthropologically, by assigning attributes to it). The Great Anonymous represents the "central existential mystery", defending forever "the derived mysteries" from human knowledge (i.e. it means the self-imposed, absolute, and eternal mystery).

Thus, the Great Anonymous provides a barrier between man and mysteries – the so-called "transcendent censorship", the metaphysical axis of knowledge; it is conceived as a "safety net" or a "firewall" (to use the language of informatics) between the human being as subject and the mysteries of the world as objects of knowledge. Due to this special kind of censorship, all human efforts to reveal mysteries and to obtain a "fully adequate knowledge" (i.e. the striving of all metaphysical systems in history) are in vain. The mysteries are never "revealed", but only "dissimulated" by the transcendent censorship, so people are never aware of this complicated, somehow super-natural process. In other words, in principle, there is the possibility of *this* or *that* knowledge, but it is never possible for one to have *the knowledge* as knowledge of the object *in itself*.

Blaga does not bring logical arguments to defend his position, according to the tradition of classical metaphysics, since his attempt is a different one. As for the reasons of believing in the finality of this structure of existence, there are no ready-made "solutions"; but one must rather seriously consider the meaning of an entity (e.g. Great Anonymous, which could have other names) playing the role of *the cognitive and ontological centre of existence*. The question is: could the "Great Anonymous" be considered as a hidden technological "God"?

2. When speaking about singularity, another position hard to defend seems to be that of the so-called "infinite" (or extremely large) intelligence. How can one understand the content of this "infinity"? How does it apply to machines (computers, robots etc.)? The idea is that *if* and *when* some intelligent machines shall design other machines even smarter than themselves, this process will cause an exponential growth in machine intelligence, leading to "singularity". But, as G. Hawkins posits, this idea is proliferated based on a naïve understanding of the nature of intelligence. What does it mean when one says "infinite intelligence"? The concept and idea of "infinity" has already set ground for a large number of mythological speculations. Is it, then, something related to the "space" of intelligence, to the time of its life or rather to the speed of its activity? Be it the last, subsequently it should be clear, at least for now, that there is no possibility to accelerate this speed endlessly (e.g. a computer processor or a software system *cannot* operate "infinitely" faster, because there are limitations for all of its

parameters). And, in fact, this is the crucial point: if there is no "infinite" acceleration of a machine's functional parameters, then there is no "singularity" either, at least in the aforementioned meaning!

Upon this claimed "infinity" of the hyper-intelligent machines hinges the problem of their alleged "immortality", i.e. the presupposition that, not being tied to any particular body, the software intelligence is essentially *immortal*. From this trait of their immortality, it has been inferred that the machines would not have neither the need to produce "off-springs" in order to perpetuate their artificial life, nor the experience of an evolutionary lust for love (or emotional feelings) – as Berglas points out. He writes that, in the future, the essential for intelligence is to stay alive, even after centuries (not the case of a human person, of course). The more hardware the artificial intelligence gains, the more intelligent it will become, obtaining again and again a better and bigger hardware. In the "end", this will be *"the" intelligence,* indefinitely extended over space and time. But this way of reasoning looks like an anthropomorphical one, which means to judge on machines' development in terms of human reproduction and competition. Again, it is very hard to argue the "immortality" of machines (no matter how 'superior' they can become compared to humans), because there are countless factors that may stop their evolution at any time (e.g. an unexpected malfunction caused by humans within their software program or by the machines themselves, a cosmic catastrophe like the collision of the Earth with asteroids or comets etc.). What can reasonably make us truly believe that a machine could stay "alive" *forever*? Are we not here rather projecting our ancient desire for eternal survival on these technical systems? As to the issue of perpetuating the artificial "species", there is no reason to stop us from imagining these machines as being interested and motivated to create some kind of "descendants" with "inferior" qualities – but maybe not very much lower that those of their "parents", on the purpose of giving them some more accessible tasks to fulfil (i.e. to keep the maintenance of certain systems, to explore unknown areas of the world, to evaluate critical situations in relationship with humans – potential dangers or conflicts – and send reports to the "central intelligence" etc.). Of course, the sexual desire and the feelings accompanying human reproduction are not to be found within this framework, but who can now tell precisely that what we call "affection" might not have something alike corresponding to the reproductive behaviour of those allegedly extremely evolved machines?

We might get a clue on this issue by comparing the problem of "superintelligence"/singularity with K. Popper's evolutionary view on philosophy:

the "evolution" of philosophy through its history is a trans-generational one, i.e. different generations of philosophers are confronted with the same questions/ problems and work to find answers/solutions. Similarly, different and (continuously improved) generations of machines are better and better prepared to face their tasks, able to correct their possible failures, to become more and more efficient, independent and intelligent.

Popper's very well known schema of conjectures and refutations (see, for instance, *in extenso* works like *Objective Knowledge: An Evolutionary Approach* or *All Life is Problem Solving*) applies not only to the growth of scientific knowledge, since Popper extends it beyond science, to the field of philosophical theories. This schema assumes that theories can be improved, briefly illustrating the progress of scientific and technological knowledge over time. Thus, scientific theories undergo an *evolutionary process* characterized as follows[1]:

$$P_1 \longrightarrow TS_1 \longrightarrow EE \longrightarrow P_2$$

Thus, given a problem (P_1), a trial solution (TS_1) is applied to the problem, for the purpose of attaining a very rigorous (even the most, if possible) attempt at falsification. The process of error elimination (EE) performs for science a function similar to that of the natural selection in the biological evolution. The result is a new problem (P_2) and so on. One can say that 'surviving' theories (as "off springs") are not truer than their "ancestors", but rather more "fit" or applicable to the initial problem PS_1. Consequently, just as a species' "biological fit" does not predict continuous survival, neither does rigorous testing protect a scientific theory from a possible future refutation; this may occur any time, every time when a counterexample is discovered.

We believe that the key-point of this schema is the evolution towards something better, be it an extremely evolved machine as an outcome of a multitude of improvements made by generations of its "ancestors". Let us suppose that those technical "ancestors" were, one after another, results of severe tests and critical technological thinking. According to Popper, a successfully tested theory denotes a certain kind of progress, towards more and more *interesting problems* (P_2). The "interplay" between the trial solutions (conjectures) and error elimination (refutations) is for Popper what makes the scientific knowledge advance towards more and more sophisticated problems or, from the point of view of our subject, to more and more sophisticated and intelligent machines.

[1] See K. Popper, *Objective Knowledge: An Evolutionary Approach*, Oxford University Press, 1979, p. 243.

3. Whereas a certain "trans-humanism" is concerned, one cannot avoid considering the problem of "cooperation" between humankind and those possible "super-intelligent" machines. Here, the point is that superintelligence is "different" and, however, superior to human capabilities of all kind. But how? Bostrom considers some of the unusual aspects of the creation of superintelligence:

– superintelligence may be the last invention humans ever need to make;

– technological progress in all other fields will be accelerated by the appearance of an advanced artificial intelligence;

– superintelligence will lead to more advanced superintelligence;

– artificial minds can be easily copied;

– emergence of superintelligence may be sudden;

– artificial intellects are potentially autonomous agents;

– artificial intellects need not have humanlike motives;

– artificial intellects may not have humanlike psyches.

Would, then, humans be left some room in the future? For instance, Bostrom discusses human extinction scenarios having superintelligence as a possible cause. One of them could occur in the event a "subgoal" would be mistakenly elevated to the status of a "supergoal" (e.g. in the process of resolving a difficult mathematical problem, the superintelligent machine can 'forget' about the limited status of the human specialist – the programmer – and perform actions which could endanger his/her life). Here we must ask another question: how far the machine can go in order to perform its tasks up to the "end"? There is no major obstacle to imagine ourselves that once such intelligence was "born" and put at work, the human capabilities should have been already sufficiently advanced to anticipate (almost) any possible collision between the demands addressed to machines and their responses, at least the most dangerous of their possible outputs. So, if Berglas points out that there is no direct evolutionary motivation for an AI to be *friendly* to humans (because an AI does not have human-like evolutionary traits), we can say that there is no direct evolutionary motivation for an AI to be *unfriendly* to us either. An extremely high intelligence should not have any major problem with understanding the kernel of human life, sympathizing with the major problems of humankind, though not as a "classical" biological creature. The demarcation line between these different positions is drawn over the question whether the machine would be not only intelligently enough developed to assume and perform unimaginable (or even unthinkable) tasks for humans, but also whether the "superintelligent" machine could become able to override the ethical commandments set in the processors by its programmers.

As a preliminary conclusion, we assert that different types of perception about the future of superintelligent machines are able to generate and nurture different visions, views and technological forecasts. To speak about "singularity" is, probably, to a larger extent, a question of how we are inclined to conceive the emergence of a possible world ruled by a supposedly extremely intelligent machine. If the coordinates of this process are seen under the fear of a possible oppressive evil system which eventually eliminates the "unnecessary" human being, then the technological 'singularity' would mean the end of humankind's mission in the world. But if the path to singularity is conceived as paved with successful attempts by humans to understand those superintelligent machines and to reach for themselves a degree of intelligence high enough to reasonably cooperate with them, then the technological "singularity" could mean the progress of humankind towards a higher degree of evolution. Regardless of one's preferred view, a lucid and critical discussion should always be welcomed in order to avoid falling into the trap of perpetuating a futile and sterile mythological story about people and machines.

Bibliography:

1. *** *Technological Singularity*, retrieved at http://en.wikipedia.org/wiki/ Technological_ singularity, accessed June 10, 2010.
2. Berglas, A., *Artificial Intelligence will Kill our Grandchildren*, 2008, retrieved at http://berglas.org/Articles/AIKillGrandchildren.html, 2010-06-12.
3. Bostrom, N., "Ethical Issues in Advanced Artificial Intelligence", *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, **2**: 12-17, 2003, retrieved at http://www.nickbostrom.com/ethics/ai.html, accessed July 14, 2010.
4. Good, J., "Speculations Concerning the First Ultraintelligent Machine", *Advances in Computers*, vol. 6, 1965.
5. Popper, K., *Objective Knowledge: An Evolutionary Approach*, Oxford University Press, 1979.