The development of Artificial Superhuman Intelligence could spell the end of human race

Marius Olariu (B00350529)
Word count: 2541

Abstract

It is very well known that the main characteristic that makes humans the dominan species on this planet is our brain, namely the intelligence that we posses. However, sooner or later, according to many authors we might have to share this planet with a man made artificial superhuman intelligent entity. The dangers (e.g. human extinction) and promises (e.g. solutions to challeging problems) that this new entity brings are both big. This work tries to analyze the literature regarding this topic and to present arguments for and against an artificial superintelligent entity takeover scenario. If the takeover happens, is the human racgoing to be extincted or enslaved?

The most important argument expressed in the literature, simply put, is that once superhuman intelligenece is present our mankind's fate is in its hands due to the extremly superior cognitive abilities of the entity. However, in the future there will be possible to enhance our intelligence through unconventional methods (e.g. brain-internet interfaces, nanobots in our bloodstream etc.) and become a sort of cyborgs that could match their intelligence with the superhuman intelligent entity.

After thoroughly researching the current stage of artificial intelligence development and discussing with some people involved in the field the author believes that there is still a long way until the first superintelligent entity is created. Nonetheless, AI safety progress should outpace AI capabilities progress since there will be no turning back after the first superintelligent entity is created.

Keywords: Artificial Superintelligence, Artificial General Intelligence, Existential Risk, AI Risk, The Singularity, Future of Humanity

Introduction

It is very well known that the main characteristic that makes humans the dominan species on this planet is our brain, namely the intelligence that we posses. However, sooner or later, according to a number of authors (Eden, Steinhard, Pearce, & Moor, 2012; Good, 1966; Kurzweil, 2010;Vinge, 1993) we might have to share this planet with an entity much more intelligent than us. The aforementioned entity will posses *Artificial Superhuman Intelligence* (ASI). Such an entity would be very powerful and the fate of human kind will be in its "hands" just as the fate of the chimpanzees now depends more on us than on the chimpanzees. Therefore, this superhuman intelligent entity would be able to prevent us from replacing it or changing its preferences and if it is unfriendly to humankind could spell the end of our race. The simple solution is to stop the development of such technology, however this is not possible due to the competitive nature of humans and the advantages (economic, military, artistic etc.) that it brings (Vinge, 1993).

Now, the humankind cannot create ASI before creating an *Artificial General Intelligence* (AGI), namely an entity that is as smart as a normal human being. At the moment there exists only narrow/weak Artificial Intelligence (AI) such as smart suggestion for information searching (Google), suggestion for products you might like based on the products that you purchased (Amazon) or AI that performs buying and selling on the stock exchange market. Life seems much better with narrow AI and the next step to which the research is focused at the moment is AGI which could be such a beneficial invention to mankind. One could imagine hundreds of Ph.D equivalent computers working 24/7 on issues like space exploration, life extension, fight against cancer or other big challenges. On the other hand, such an entity would be self-improving (just as we humans are) and would create the premises to go to the next stage, ASI. The point in time when the first ASI will appear is known as *The Singularity* or *Intelligence Explosion* and it was first described by Turing's chef statistician I.J. Good in his 1966 paper (Good, 1966):

> *Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus* ***the first ultraintelligent machine is the last invention that man need ever make,***

provided that the machine is docile enough to tell us how to keep it under control.
(Good, 1966, non-paginated)

The control problem - the problem of how to control what the superintelligence would do is quite challenging and it seems like we have only one change to tackle it (Bostrom, 2014). To put it (more) simply, the ASI would become awaked in a prison (connected to the outside world through cables that can be unplugged) and guarded by mice (human AI researchers). Once freed, how would the entity feel about its creators? Awe? Almost certainly not since at the moment developing machines with feelings (even if it could be possible) it seems not to be an objective of AI creators. What could stop it from destroying us or making us its slaves? Probabliy nothing, just as we humans do not think that much of how many ants are going to die because we want to build a highway or as Yudkowsky (2008) puts it "the AI does not love you, nor does it hate you, but you are made of atoms it can use for something else.".

Case For

Firstly, a a goal-driven intelligent system is measured by how effectively can fulfill its goals, however without having human values such an entity could acquire physical resources and eliminate potential threats on the quest of reaching its goals (Bostrom, 2014). Since humans can represent at the same time physical resources (e.g. make some humans do something that is not feasible for an ASI entity because it does not have a physical body) and potential threats (e.g. a comitee that takes care of international peace), the ASI could start a confilct (even a war!) in order to fulfill its goal. At the same time, defining what human values an ASI entity should have seems a difficult task given the complexity of human preferences (Yudkowsky, 2011; Muehlhauser and Helm, 2012).

Secondly, one might think that designing ASI entities with harmless goals will be harmless but that is not the case (Omohundro, 2008). S. Omohundro argues that a chess-playing robot run by a cognitive architecture (sophisicated enough that it can rewrite its own code to maximize the changes to win chess matches) can be indeed dangerous *without special precautions*. For instance it could resist being turned off (because the utility probability of such an event will be 0, cannot reach the goal of winning chess matches), could try to break into other machines to make copies of its software or acquire resources without anyone else's safety. In other words, almost any advanced intelligent system (e.g. AGI) will have 4 basic drives: preserve itself, preserve the content of its current final goals, improve its own rationality & intelligence and acquire as many resources as possible. Omohundro (pioneer in AI lip reading, image recognition and one of the six engineers who created *Wolfram Mathematica*) believes that without very careful programming all smart AIs will be *lethal* due to their eventual psyhopathic, egoistic and self-oriented entity.

Human behaviour is quite irrational in some situations and because of that there have been developed disciplines focused on the study of human irrationality (Tversky and Kahneman, 1974). These irrationalities (e.g. smoking - harms your body, drug addiction, gambling etc.) give rise to vulenrabilities that are exploited by the free market, however, humans are becoming more and more rational but this is a slow process. Biological evolution moves slowly toward rationality. On the other hand, AI entities will eliminate their physical and software vulnerabilities in order to protect themselves from someone who might exploit them, that is to say they will improve at a much faster rate than we biological humans do (Omohundro, 2008). An entity that is truly rational, compared to us humans, can really take charge of the things happening around it and cannot be stoped that easily by humans.

There are several ways of reaching AGI and one of them is Whole Brain Emulation (WBE) - scan an biological brain and construct a software model of it so that the software behaves essentially the same way as the original bilogical brain (Sandberg, 2013). In this scenario the digital intelligence might inherit the motivations of the human template and they can be retained even at the stage when this digital brain has reached superintelligence (Bostrom, 2014). Now, without careful attention on what biological brain is chosen

for WBE this can lead to catastrophic outcomes, just imagine if one would choose Adolf Hitler's brain.

Most of the participants of *Future of Humanity Institute* conference on machine intelligence on 16.01.2011 agreed that the ultimate consequences of the creation of a human-level (and beyond) intelligence will have an "extremely bad" outcome (Sanders and Bostrom, 2011).

Case Against

Most of the dystopian scenarios presented in the literature involve a *weak humanity*, that is - one that has not enhanced its intelligence through other means that the classical ones (learning, practice, experience etc.). This type of humanity it is most likely to obide an ASI entity. However, the second type of humanity - *strong humanity*, is one that has enhanced its intelligence through different types of technology (Kurzweil, 2010), for example a human having a wireless implanted brain-mind interface that allows one to access internet. This *strong superhumanity* would be full of cyborgs (godlike humans) that can match their intelligence with ASI entities and therefore the ASI cannot take control over the world. Probably the *strong superhumanity* will allow the "stay-behinds" (people that have not altered their biological body) to live a happy live (Verge, 1993).

Although Eric Drexler, founding father of nanotechnology, agrees that ASI will be developed and it will pose a threat to mankind, he argues that the ASI can be confined using physical rules so that their behaviour can be examined by humans, thus making ASI entities safe (Drexler, 1986). Along this line of thinking, the telecommunications companies (e.g. Vodafone) use narrow AI to get recommendations on how to modify the configuration of a certain network to better meet the users' needs from a geographical region. Again, recommendations and not actions, the human expert is still in the lopp - he/she is the one taking the decision to perform a certain task.

The same idea of constraining AI systems and making them safe is backed by Chalmers (2010). One of the suggestions is to create AGI in *virual worlds* where it will have free reign without direct effect on our world. This virtual world would allow humans to observe the AGI. Eventually, after many cycles, will develop ASI in these virtual worlds and the ASI could be monitored to see if it is benign and to determine wheter is safe to deploy it in the real world.

According to Chollet (2017) from Google Research, intelligence cannot be expanded without an humanly environment, sensorimotors and human interaction. There are rare cases of humans with high IQ but they are not scientists, they do not solve challenging problems and they *do not take over the world* (just as some people fear an superhuman intelligence would do). In other words, Chollet(2017) argues that even if there is developed a physical robot with a digital brain it cannot develop greater cognitive capabilities than a smart human due to the lack of a stimulating environment, thus an intelligence explosion would not occur.

Summary Diagram

Arguments on Balance, Conclusions and Recommendations

At the beginning of the research for this work the author was tempted to believe that we are close to a major breakthrough in AI just by seeing the formidable performance of AlphaGo that managed to beat 18 times world Go champion Lee Sedol (Silver et al., 2017), however that is not the case. AlphaGo is just narrow AI - advanced computer program that you run to get a functionlity, it has no knowledge that is playing a game or that there exists a real world, it is not intelligent as his human opponent is (although it managed to outperform him in an intelectual activity) . Another example of narrow AI is the Deep Neural Networks image processing that works (almost) perfectly on image content analysis (e.g. detect objects, animals or human face expressions in an image). I say *almost* perfectly because recently it has been proven that this type of narrow AI can be fooled (Nguyen,Yosinski and Clune, 2015) to detect objects with 99.99% confidence in images that are imperceptible (e.g. white noise images).

In the following lines I will give some recommendations found in the literature regarding the topic discussed in this work.

Goertzel (2012) proposes a global "Nanny AI" that would forestall the developments towards ASI until the problems of AI safety are solved, this nanny AI is going to posses human-level intelligence (or above) and have the following characteristics: interconnection to powerful worldwide surveillance systems, strong inhibition against rapidly modifying its general intelligence, open-minded towards the possibility of misinterpreting its initial goals and others.

In the 1970s scientists involved in DNA research were confronted with the promise and peril of the recombinant DNA (mixing genetic information to create new life-forms) thus a set of rules for conducting DNA research had to be created. In the 1975 Asilomar Conference (California) a set of rules were devised, for instance: work only with bacteria that could not survive outside laboratory (Berg et al., 1975). Inspired by this fundamental conference the *Future of Life Insitute* (FLI) held a conference in Asilomar on 5-8 January 2017 were more than 100 AI researchers and leaders from economics, law, ethics, philosophy were invited to address and formulate principles of beneficial AI. These principles were published online as Asilomar AI Principles (2017), to mention but a few of them: create beneficial intelligence, AI systems should be safe/secure throughout their use or AI arms race should be avoided.

As stated by Bostrom (2014) an ASI entity can be controlled using two methods, namely capability control (control what the superintelligence can do) and the motivation selection method (control what it wants to do). The capability control can be ensured by not developing the system with physical manipulators that could interact with real-world objects, boxing the system in a metal mesh (e.g. Faraday cage) in order to prevent it from transmitting radio signals to radio receivers that could eventually be manipulated or bar the system from accessing communication networks. From the motivation selection perspective the AI could be left to freely interact with society so that it would acquire human-friendly final goals just as we humans come to value other individuals' principles and goals.

To sum it up, there is still a long way until the mankind creates ASI since at the moment we have only narrow AI. We have seen the dangers (human extinction, another war, humans enslaved) and the promises (solutions to our individual and societal problems, a "return to Eden") that an eventual ASI entity would bring and we are looking forward for a positive outcome. In the future AI safety progress should outpace AI capabilities progress since there will be no turning back after the first ASI entity is created.

References

Asilomar AI Principles. (2017). Available: https://futureoflife.org/ai-principles/ [Accessed: 25 November 2018].

Berg, P., Baltimore, D., Brenner, S., Roblin, R.O. and Singer, M.F., (1975). Summary statement of the Asilomar conference on recombinant DNA molecules. s.l.: Proceedings of the National Academy of Sciences of the United States of America, 72(6), p.1981.

Bostrom, N. (2014). Superintelligence. Oxford: Oxford University Press.

Chalmers, D., (2010). The singularity: A philosophical analysis. s.l.: Journal of Consciousness Studies, 17(9-10), pp.7-65.

Drexler, K. (1986). Engines Of Creation. New York [u.a.]: Anchor Press, Doubleday.

Eden, A.H., Steinhart, E., Pearce, D. and Moor, J.H., (2012). Singularity hypotheses: an overview. In Singularity Hypotheses (pp. 1-12). Berlin: Springer.

Goertzel, B., (2012). Should humanity build a global AI nanny to delay the singularity until it's better understood?. Journal of consciousness studies, 19(1-2), pp.96-111. s.l.: ImprintAcademic.

Good, I.J., (1966). Speculations concerning the first ultraintelligent machine. In Advances in computers (Vol. 6, pp. 31-88). s.l.: Elsevier.

Kurzweil, R., (2010). The singularity is near. s.l.: Gerald Duckworth & Co.

Muehlhauser, L. and Helm, L., (2012). The singularity and machine ethics. In Singularity Hypotheses (pp. 101-126). Berlin: Springer.

Nguyen, A., Yosinski, J., and Clune, J. (2015). Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In Computer Vision and Pattern Recognition. s.l.: IEEE.

Omohundro, S.M., (2008). The basic AI drives. In AGI (Vol. 171, pp. 483-492). s.l.:s.n. . Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.393.8356&rep=rep1&type=pdf [Accessed: 5 November 2018].

Sandberg, A. and Bostrom, N., (2011). Machine intelligence survey. FHI Technial Report, 1. s.l.:s.n. . Available: https://www.fhi.ox.ac.uk/wp-content/uploads/2011-1.pdf [Accessed: 10 October 2018].

Sandberg, A., (2013). Feasibility of whole brain emulation. In Philosophy and Theory of Artificial Intelligence (pp. 251-264). Berlin: Springer.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A. and Chen, Y., (2017). Mastering the game of Go without human knowledge. s.l.: Nature, 550(7676), p.354.

The impossibility of intelligence explosion, (2017). [Online] Available: https://medium.com/@francois.chollet/the-impossibility-of-intelligence-explosion-5be4a9eda6ec [Accessed: 26 November 2018].

Tversky, A. and Kahneman, D., (1974). Judgment under uncertainty: Heuristics and biases. science,

185(4157), pp.1124-1131. s.l.: American Association for the Advancement of Science.

Vinge, V., (1993). The coming technological singularity: How to survive in the post-human era. s.l.:s.n. . Available: https://edoras.sdsu.edu/ vinge/misc/singularity.html [Accessed: 10 October 2018].

Yudkowsky, E., (2008). Artificial intelligence as a positive and negative factor in global risk. Global catastrophic risks, 1(303), p.184. Oxford: Oxford University Press.

Yudkowsky, E., (2011) . Complex value systems in friendly AI. In International Conference on Artificial General Intelligence (pp. 388-393). Berlin: Springer.