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

Marius Olariu (B00350529)
Word count:

Abstract

March 2016, AlphaGo becomes the first computer program to beat a 9 dan Go professional player in a Go match, an achievement a decade ahead of its time. Next year, AlphaGo Zero is released and beats first version of AlphaGo, 100 – 0. The later version possesses the ability that no human has: develop a skill at professional level by self-instruction. Is it this the first nudge of superhuman intelligence? If so, how long until a machine becomes self-aware and takes actions that the human mind cannot conceive and stop? This work tries to provide an analysis of the current stage of Artificial Intelligence development towards Superintelligence, present the benefits and risks of Artificial Superhuman Intelligence and some possible answers to the above-mentioned questions. The author analysed the scientific papers published regarding AlphaGo, the scientific papers regarding the subject from a philosophical perspective and discussed with experts from domain in order to write this case study. It is clear that the development of Artificial Superhuman Intelligence cannot be avoided due to the competitive nature of humans and the benefits that this breakthrough promises. However, not all hope is lost, there is the concept of Strong Superhumanity that could become one with the new entity and will not forget its roots. The Strong Superhumanity will allow the "stay-behinds" to live their lives and give them the false appearance that they are in control of their life just as we humans do with chimpanzees.

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

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, we might have to share this planet with an entity much more smarter 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 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.*** (I.J. Good)

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 (Bostrum, 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.

Case For

An intelligent agent 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 the humans to do something that is not feasible for ASI 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 .

Designing an ASI with harmless goals is not enough, we as humans have positive goals yet we get to harm each other on our journey to fulfill our positive goals. (Omohundro, 2008)

AI race competition (Bostrum, 2014)

Most of the participants to the *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, 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 (do not take action

automatically!) on how to modify the configuration of a certain network to meet the users' needs from a geographical region.

//congress to establish a set of rules //make them with parts that fail (problem it might detect that)

Summary Diagram

Arguments on Balance, Conclusions and Recommendations

//take some suggestions from SUPERINTELLIGENCE //enter a group of humans always in the decision loop //set of drives that it should hav

1 References

//TODO check they are in alphabetic order and contain all the details necessary (like publisher, year, etc) //TODO make sure there are many

Omohundro, S.M., 2008, February. The basic AI drives. In AGI (Vol. 171, pp. 483-492). Bostrum, N., 2014. Superintelligence: paths, dangers, strategies. Oxford University Press. Oxford Vinge, V., 1993. The coming technological singularity: How to survive in the post-human era. Good, I.J., 1966. Speculations concerning the first ultraintelligent machine. In Advances in computers (Vol. 6, pp. 31-88). Elsevier. Drexler, K.E., 1986. Engines of creation. Anchor. Sandberg, A. and Bostrom, N., 2011. Machine intelligence survey. FHI Technial Report, 1.