The role and risks of artificial intelligence in our future society

Mårten Nilsson
marten3@kth.se

October 2016

There are currently many opinions of what artificial intelligence (AI) is and is not. Norvig and Russel [1] encapsulated eight definitions of AI into four categories concerning different aspects of animal intelligence. The definition used in this essay is the one of these eight that corresponds best with my intuition of AI, stated by PH Winston as "The study of the computations that make it possible to perceive, reason and act." [2]. For a system to be considered intelligent it has to perceive, reason and act to achieve a specified goal.

An intrinsic risk of goal oriented reasoning agents is unexpected side effects. The plan that the agent creates to achieve its goal is not always the obvious plan for a human. The level of damage caused by unexpected side effects are only limited by the resources of the agent. An AI playing chess can at worst make a move considered counterintuitive by a human, but an AI driving a truck might decide to save fuel by colliding with a pedestrian on the road instead of stopping the truck. A common approach to prevent unexpected behaviour of complex systems is to perform comprehensive tests. Most non intelligent systems yield the same behaviour in similar situations and the possible unexpected behaviour that may arise are small malfunctions that cause the system to fail. For an intelligent complex system testing the decision making process is an elaborate task, since just a small change in input may cause the system to reconfigure it's entire plan.

Another risk with AI is that it may surpass humans at too many tasks too fast. Since the industrial revolution automation of labour have been constantly increasing. This has been a great part of the development of our society to the point it is today. Though automating jobs ultimately frees people of labour it also forces individuals to readjust. The society we live in today relies on individuals to make a living for themselves which only works as long as everyone has the ability to contribute. Historically there have always existed some unqualified labour for people to fall back to when they need to readjust their working situation. In our current globalized society a breakthrough in artificial intelligence could possibly eliminate the majority of all unqualified labour in a matter of decades, possibly less. The need for everyone to work would disappear and instead it would only be necessary to have a small group of experts to program AI:s to perform all labour for us, including intelligent labour. Ultimately this would give people more time to enjoy life, create art and science but if the society is not ready for this change the consequences could be devastating.

I believe that many of our jobs today will be automated in the next decades thanks to recent progress in artificial intelligence and machine learning. Though there are some risks associated with this progress, I believe that as long as we are aware and prepared for the risks and always put our safety first this will change society for the better. People will have more time to understand their surroundings and more powerful tools available to deal with global challenges.

References

[1] S Russel and P Norvig. *"Artificial Intelligence–A Modern Approach"*. Pearson Education, 3 edition, 2014.

[2] Patrick Henry Winston. *Artificial Intelligence*. Addison-Wesley, 3 edition, 1992.

word count: 535.