Risk of AI in Future Societies

Discussion group 43

Alice Karnsund alicekar@kth.se
Elin Samuelsson elsamu@kth.se
Oliver Åstrand oast@kth.se
Ulme Wennberg ulme@kth.se

Summary of discussion

We started out by identifying three major problems, which we consider as the most important ones when looking into AI and the society in the future. The first problem we discussed was the possibility of automation in many sectors, that will most likely affect a great amount of people in the near future. The second one was that of developing AI weapons for military use. Lastly, we discussed the more distant risk of superintelligent AI.

Alice started the discussion by introducing the statistics that 47% of american jobs are at risk of being automated. She took up the possible solution of universal basic income, where each citizen unconditionally gets an amount of money. That led to the the question of whether we will have to redefine what provides meaning for humans. This, we thought, will have to shift from coming from the work that we do, to the more social aspects and explorative endeavour of life.

We further discussed that robots have been replacing humans ever since the industrial revolution, so what is different this time? We came up with the answer that this time, the effects might be orders of magnitude larger, as we for the first time are replacing cognitive abilities, rather than just physical abilities. Oliver asked the question of what there is that humans can do, except for physical and cognitive abilities? Living in a society where everything basically is automated and excluding humans from almost any task would not be sustainable, since that means that people would get isolated from each other to a much greater extent, whether one chooses it or not. Therefor we said that for sectors where social and physical interaction lies as a basis, it would most likely not be socially accepted to use automatization.

We also discussed the possibility of using AI for military purposes. The group agreed that developing such technology would be a bad idea. Ulme argued that potential loss is the important factor. When initiating a war today, human lives from the own country will be sacrificed, which forces leaders to consider such a decision extra carefully. If it was possible to attack another country with AI robots instead of human soldiers, the threshold for war would become lower and the world more violent. On the other hand, Elin pointed out that as soon as one country has AI war robots, others will immediately follow, leading to the same kind of deadlock as nuclear weapons cause today. No one could ever use their robots anyway, since they know that would lead to an equally powerful counter attack from the enemy.

Oliver discussed whether we should delay the development of AI to avoid it being used for bad purposes. Everyone was convinced that the development of AI is unstoppable, humans are too curious, but maybe we could establish laws to slow down the process. Then, when real AI is reached, there might be fewer conflicts and the world more ready for this technology.

The last topic that we discussed was the risk that superintelligent AI poses to humanity. We talked about both the problem of how to control an agent who’s intelligence far outstrips ours, and briefly touched on the problem of making sure the AI has the same goals and values as us. We all agreed that this of course might pose an even bigger thereat than the other two, but it’s a lot farther in the future, so it’s a lot harder to grasp and come up with concrete solutions. *(Words: 584)*

AI-Risks And Their Effects On Society

Ulme Wennberg, ulme@kth.se

Artificial Intelligence has been highly debated since it was founded as an academic discipline in 1956. The field has undergone several waves of intense hype, to each time be followed by a so called “AI-winter”. In the last few years, thanks to the internet, larger datasets and through better processors, there have been a few notable achievements. With this in mind, it is interesting to look into the future and see what is to be awaited. As I see it, there are three great risks that are associated with AI.

The first risk is rapid impact on the economy that the current system is not ready for leading to a stationary state with a more uneven distribution in wealth. As the world transitions to AI and algorithms performing a greater and greater degree of all products and services in the world, there will be a few obstacles. First of all, these products and services are likely to be more capital intensive goods and services than those performed by man. Instead of investing in really competent personnel, business are likely to instead invest to a greater degree in supercomputers. This, in turn, leads to higher leverage on capital, while wages overall will tend to decrease. This risks leading to a more uneven steady state in the economy.

The second risk I see is the risk of an arms race between different global actors such as companies or even countries. If one actor gains a big advantage in AI, they would be incentivized not to share its advancements with the rest of the world, but rather to use it to gain superiority against other actors. Because of this, it leads to an unhealthy situation where it would be utility maximizing to start a war/bomb another actor, as soon as the former actor suspects that the second has made a breakthrough in AI.

The third risk I see is associated with the difficulties of controlling AIs with superhuman planning abilities. This is the problem that Oxford Philosopher Nick Bostrom discusses in his book Superintelligence. Namely, how are we going to deal with the difficulties of controlling future AIs with superhuman planning abilities? Though institutions such as MIRI in Berkeley are currently pursuing research in how to assure that superhuman planning systems are safe, there is still a long way to go. If we cannot control such systems, we have many reasons to fear for our survival. What makes us humans rule the world today, instead of for example tigers or apes, are not our shear strength, but rather our intelligence and our ability to plan and carry out such plans more effectively than any other animal. If we build superhuman planning systems, and it turns out that these systems have values that are misaligned with our own, then it is unlikely that humans will remain in charge. This is the line of reasoning that people such as Elon Musk follow when claiming that Artificial Intelligence is one of the greatest existential threats to humanity.
[Word Count: 502]

The Role of AI in Future Societies

Alice Karnsund, alicekar@kth.se

Lately we have seen major advances in Artificial Intelligence (AI), Machine Learning (ML) and Mobile Robotics (MR). Together with the enormous amount of data that is being collected, sophisticated algorithms now allow both routine and non-routine based tasks to be automated.[1] This means that major parts of the low and middle wages jobs are at risk for being computerized. This includes unemployment in sectors like services, sales, production, transportation, office and administrative support etc. This corresponds to approximately 47 percent of the total US employment occupations are at risk at being automated over the next decades or so.[2]

The development in the areas of AI, ML and MR is happening at an absurd pace and may affect a greater amount of people in a much shorter time than earlier industrializations has. Some believe that the society will adapt, just as it has done before and that many new jobs will be created as we go. But if this will not be the case, this means that a large part of the population may become unemployed. Which in turn means that they will not have a stable income, and therefore cannot engage in the trading market as before. This will in the long term affect the world economy. In [3] Martin Ford brings up a few ideas how to tackle this concern and one is that to give every adult an annual basic income, which would help keeping the economic activity. Ford also mentions more than once that this technical progress will lead to a "winner takes it all scenario" where the winners are the leading companies like Google, Tesla, Facebook etc. And with this the economic classes differences will get even more extreme. To make this change happen at some customized pace, I believe that it is up to the world leaders to set up clear regulations in preventing robots etc. in too large extent.

Overall when it comes to the risk of AI I believe that sooner than later we will have developed an AI that can do everything better than humans. This is something one should be very concerned with since AI is making great progress in the military for instance. For example, the Autonomous Weapon System is "a weapon system that once activated, can select and engage without further intervention by a human operator".[4] This together with face and audio recognition etc. will turn the battlefield into something never seen before. Unfortunately, I think that AI in this area might be the most difficult growth to prevent. Even though great parts of the world may try to agree on certain constraints, there will always be people with other objectives.

Turning to the bright side of AI I believe a lot will have happened in the next ten years or so. As Elon Musk says in [5], almost all cars produced at that time will most likely be autonomous, and autonomous cars has a lot of advantages compared to human drivers. The number of collisions will probably fall drastically as well as the driving will be more environmental friendly. I do also believe that the health sector will benefit a lot from a range of classification systems and advanced robots for surgery. Hopefully this will help us find solutions to major health problems like cancer and heart diseases etc. *(Words: 549)*

[1] Frey and A. Osborne. "The Future of Employment: How Susceptible are Jobs to Computerisation", 2013, pp. 27.
[2] Frey and A. Osborne, pp. 38.

[3] Martin Ford. “The Rise of the Robots”, 2015. pp. 270
[4] “The Ethics of Autonomous Weapons Systems”, 2014.
https://www.law.upenn.edu/institutes/cerl/conferences/ethicsofweapons/
[5] Elon Musk, https://www.youtube.com/watch?time_continue=2555&v=2C-A797y8dA

The Risks of AI and Its Role in Society

Elin Samuelsson, elsamu@kth.se

Humans have always had the urge to optimize and improve. This has lead to our society successively becoming more automatized, which makes people excited, but also terrified. Will AI robots steal all jobs? Well, we have been replacing humans with machines ever since the industrial revolution, but the affected workers have in general belonged to the less privileged part of the population. The machines of our time are not restricted to performing physical tasks, they can think too. They are capable of planning budgets, writing articles and even code new programs. This affects other groups of people, who have larger influence over the public opinion than any poor factory workers ever had. This can explain why the skepticism towards AI is so spread. Historical examples point in a more optimistic direction. As soon as society has adapted to the changes, new technology tends to eliminate some jobs, but also create new ones, and improve people's standard of living in general.

On the other hand, AI is not just any new technology. It is in fact more powerful than anything we have created before, more powerful than the human brain itself. Specifically, instructions to an AI must be given carefully. The machine will interpret them literally, without the intuition and context humans have from living a life. For instance, how do we teach an AI to be a "good citizen"? We could provide it with the statute book and a selection of court cases, and then forbid it to break any laws. However, there are probably some logical gaps in the laws, that an intelligent machine would find. A human might intuitively say that a "good citizen" should not exploit the system like that, while the AI only reasons based on the information it has access to.

A topic that creates many discussions is AI in war industry. Consider the following scenario: a killer robot, totally superior to humans, so accurate that it only kills soldiers, never civilians. One can argue that this would be to prefer over emotionally controlled, irrational humans. But what if the robot interprets our instructions in an unexpected, undesired way? A killing action would be made before anyone even realized what was about to happen. Even more importantly, there is something fundamentally wrong with giving a robot permission to kill humans. It is not progressive at all, but that might be an argument applicable on the war industry in general, rather than AI technology in particular.

Lastly, I would like to discuss the possibility of developing an ideal, perfect AI in a distant future. Simply, a computer at least as complex as humans, and totally superior. We could never understand such a system, neither predict what it goals and motivations would be. Maybe its values will not at all be compatible with ours. On the other hand, the leaders of humanity has made some quite unreasonable decisions over the years. Maybe, letting someone/something else decide is not a bad idea. Well, given that AI will care about human lives of course, which, when you think about it, can not be guaranteed.

To conclude, developing AI is risky, but at the same time unavoidable. Humans are curious and if something can be created, we will do it. We could slow down the process with laws and restrictions, but in the end, we will never be able to resist the temptation.
(Words: 563)

Essay on AI-risk

Oliver Åstrand, oast@kth.se

There are two radically different threats that Artificial Intelligence poses to society in the future. The first threat is Narrow Artificial Intelligence (NAI) disrupting the societal order, by quickly changing the dynamics of a critical area of society. The second threat comes from the potential development of Artificial General Intelligence (AGI), and the problem of controlling and containing such an agent. I will argue that the first threat poses a large risk for society, but not necessarily a risk for humanity. The second threat on the other hand, is a potential threat for the continued existence of the human race.

It is non-controversial to state that the domains in which there exist NAI:s of superhuman performance grows. There is a deluge of professions that are on the verge of becoming obsolete due to AI. While not every single person in these fields will lose their job, there is certainly a high risk that most will do so in the coming decade. Unless there are jobs in new sectors for these people we will see a large increase in unemployment that will be hard for society to buttress. Many people think that we will just move on to do things we cannot imagine today, while this is possible it is also possible that the new jobs require training that is unattainable for most people. Comparing it to the industrial revolution it is for example easier to train a farmer for factory work than to train a truck driver to be a AI-programmer. It is possible that the rate of progress in AI is so rapid that we do not have time to retrain the unemployed as to keep our current economic system. One possible solution to this problem is universal basic income, financed by taxing the product of labor (VAT) from the algorithms.

While it is not as certain that we will have to face the consequences of the second threat in the coming decades, it is still worth discussing now due to the magnitude of the adverse effects of ignoring it. If we are to believe AI-researchers then the development of AGI is only about 30 years away [1]. There are two issues with developing such an intelligence. These are called the control and the value alignment problems [1]. The control problem is the issue of controlling an agent that is vastly more intelligent than us. A superintelligent AI will almost certainly be able to achieve its goals. For us to control it will be like gorillas trying to control humans. They are not able to conceive of all the ways a human could try to escape from their imprisonment.

The second part is the value alignment problem, it is very hard to make sure that the agent we design has its values aligned to ours. A dummy example of this would be a researcher giving the AGI the goal to solve a very hard mathematical problem. If the problem is too hard for the AGI at is current capacity it might try to get more computing power. Given that it is an AGI it will be able to manipulate the stock market in order to buy processors. This search for more power might not stop until it the whole earth is just a giant computer, crushing us in the way. One possible route of solving this is giving the machine a goal that implies some uncertainty to what it's goal is [1], for example asking it to "do whatever we would have had the most reason the ask the AI to do".

[Word Count: 594]

[1] Bostrom, N., *Superintelligence,* Oxford University Press, 2014