Current Advances in convolutional and recurrent neural networks - Summer Semester 2018

Ethics in AI - risks of AGI and automation

Marius Hobbhahn

***Abstract*—Artificial General Intelligence (AGI) can lead to high impact scenarios for humankind. We argue that because AGI is superintelligent, unalignable with human goals and uncontrollable, these scenarios will in the vast majority of cases have grave negative consequences. Automation will lead to high inequality within societies and redistribution does not work in a highly automated world. While many researchers are very optimistic about AGI and automation, this paper aims to correct this view in arguing that reality converges towards a more grim scenario.**

***Index Terms*—General artificial intelligence, automation**

I. Introduction - AGI

AGI describes an entity that is capable of adapting to solve problems in nearly every circumstance. The only AGI we know so far is the human, but it is also possible that the same capacity can emerge in non-carbon-based systems, i.e. computers. In the following we assume that an AGI is a computational system that is at least as good as solving problems in nearly all circumstances as humans are. Imagine a world in which such a system exists and is owned by a hobby stamp collector. The collector wants to use the AGI to collect more stamps and therefore gives it access to the internet. At first the AGI browses the web and efficiently buys the cheapest ones from different sites. Later, they realize that printing stamps has less costs and in order to maximize, output it hacks into other privat printers to do that. After every printer in the world prints stamps and all available paper is gone, it searches for new resources to generate paper. After all wood is transformed it searches for other sources of carbon and water. Most animals, including the human, consist mainly of these two ingredients. The scenario that started with a simple, innocent idea, ends in a bloodbath with incredible suffering. This paper argues that nearly all attempts will end with unwanted consequences even when every part in the system has best intentions.

A. the AGI has a goal and acts like an agent

One possiblity is that the AGI can be a tool for other agents purposes, like in the case of the stamp collector. Then they maximize a given utility which is equivalent to having exactly that goal. Another possible implementation is for AGIs to have more complex goals or to develop their own. This could be imagined as a sophisticated agent very similar to their human creators as often visualized in science fiction literature. In all possible implementations, however, the AGI has goals and it has the possibility to choose actions in their environment to fulfill these goals. For a more detailed explanation see [1].

B. AGI is superintelligent

By superintelligence we do not mean the difference between an IQ of 70 and 130 or the difference between an ant and a human but the difference of a scale that is hard to imagine to our minds. There are three major reasons, why they are likely superintelligent. a) Current systems already have significantly more computing power than human brains. This trend is likely to go on. b) An AGI is not confined to one main system like the human brain but can use multiple systems at once and parallelize processes in grand style, i.e. use multiple server farms at once c) Increases in intelligence are likely to rise exponentially. As soon as a system is able to improve their own code, they can eradicate flaws and repeat this process from the improved state. Increasing intelligence is an instrumental goal for maximizing their goal, so it is very likely for the AGI to acquire it. Importantly, there is no relation between the complexity of a goal and the intelligence of the entity following it. As Nick Bostrom argues in [2] these properties are orthogonal. This implies that an AGI can follow really simplistic goals from a human perspective, i.e. maximizing the number of stamps in the universe.

C. goals of AGI are unalignable with human goals

Before going into argumentation we would like to make two clarifications: a) We cannot solve the question of which is the most accurate ethical system in a three page paper. We argue that the consequences to which AGI will likely lead are perceived as negative in most ethical systems, i.e. human extinction or great suffering. b) This argument is neither a rigorous proof nor a strong incentive mechanism, but rather a try to make plausible that it is very hard to make human goals and the agents goals align.

The primary idea is that most goals humans either have a lot of unexplicited underlying premises that we are not aware of while formulating our goals or that human ethical systems have found to be inconsistent in many cases. Imagine you have an AI with the explicit goal to fill a cauldron with water. A human would now always try to fill water in the cauldron until it is full. The agent is unlikely to stop filling water in, flooding the whole room, since there is no punishment for it. We reprogram the utility function, such that water on the floor has a cost attached to it. Since it is hard to keep the cauldron exactly full it might start to just make sure that the measurement device is always reporting a full cauldron, instead of actually filling it. The agent just discovered reward hacking. It is not clear that there will always be a problem with every utility function, but so far it seems intuitive that very often you can create a

counter example.
It is sometimes argued that maximizing water in a cauldron or collecting stamps is a stupid utility function, why not just do something reasonable, such as making people happy? Here again it is very unclear how happiness is defined. The agent might put every citizen on heroin for the rest of their lives, minimizing individual autonomy. Or, since happiness correlates with smiling, force people to smile all the time. If you want to solve this by changing the utility function to prevent suffering, the fastest way is to kill all sentient beings. If you include a cost on killing individuals, it might introduce global abortions for every child to prevent future suffering. So far we have not found an ethical set of beliefs that does not have inconsistencies or flaws. Until that normative system is found, we will always struggle to align an AGI with human goals. Further explanation can be found in Eliezer Yudkowsky's article about alignment problems [3]

D. Goal preservation as an instrumental goal

Once a goal is established every action will be evaluated in terms of a maximization of this goal. Every attempt to change the current goal is inefficient in maximizing it. Therefore the AGI will not allow its goals to be changed. If you change the goal of a stamp collector AGI, future worlds will contain less stamps which the AGI will try to prevent. Since it is significantly more intelligent than we are, it is very likely that they are successful in preventing these attempts. For deeper analysis on instrumental convergence see [4].

E. AGI will escape controllable environments

Common answers often given to the just outlined problems are: "just implement a stop button" or "just simulate it, to see how it behaves". While these might be good intended concerns they turn out not to solve the problem. Every time the AGI is stopped it will not be able to maximize its utility function. Self preservation therefore becomes an instrumental goal. In a stop button scenario the agent will either try to remove it while improving its code or, if that is impossible, try to manipulate humans into preventing the button from being pressed until a solution for the AI is found, i.e. cloning its code onto multiple independent machines. Similar problems arise in the simulation scenario. Every connection the AGI has to the real world is a connection to a human being that can be manipulated into freeing the agent. If you think there is a save alternative solution that nobody has thought of yet, think whether you would still trust that solution if you knew it needs to control an entity a billion times faster and smarter than you. Further explanation can be found in [5]

II. The first AGI is unlikely to be secure

Given the possible implications of AGI for political or security purposes there is a very strong first mover advantage. Any country that first has a loyal AGI can very quickly establish global dominance if wanted. But even when more benign world leaders are assumed, researchers in AGI have the same incentives. Most grants or fame is in achieving AGI, not in security concerns. Developers who also invest time and money in safety are likely to be outcompeted by less careful ones. Given the massive first mover advantage on all levels it is very unlikely that the first AGI will be maximally save.

III. Conclusion - AGI

AGI is likely superintelligent, uncontrollable and their goals do not align with those of humans. Given the combination of these attributes, most, if not all outcomes, will have very impactful negative consequences for all living beings. This risk is further amplified by the first mover advantage that incentivises individuals on all level to not care about AI safety.

IV. Introduction - Automation

A study conducted by McKinsey [6] estimates that 800 million jobs are replaced by machines in 2030. Automation will likely happen at a fast pace and have very disruptive effects for the employees of our economy. It is often argued that when human muscles were replaced by machines during the industrial revolution more jobs were created than taken. In this paper, we argue that this time our minds will be replaced and it is likely that many people will not find work. Just as the horse was quickly replaced by cars at the beginning of the 20th century, the most valuable resource of humans will be replaced this time.

A. Automation will happen

For the following argumentation we assume that companies will try to maximize profit and the markets are, if not shown otherwise, approximately efficient. And even if public outcry against the automation of human labor might create backlash it is likely to recede after people get used to it.
As soon as it is cost efficient, human labor will be replaced. Robots can work 24 hours, do not make as many mistakes, do not need payment or pensions and do not strike. We can see this trend in closed environments, i.e. car makers have already automated most of their production. Since we now made plausible why companies will automate when the technique exists, the only question is how fast machine learning gets better. And the answer is: more quickly than we expected. The developments in deep neural nets allowed DeepMind to crack Go and Chess and Open AI to create a bot to play Dota 2, all tasks recently thought to take decades to solve. Autonomous cars already drive in many countries as test cars only waiting for legislation. We cannot accurately predict what will be the next scientific discovery, but given the current trend it will likely be sooner than we think.

B. Automation will replace a lot of jobs and unemployment will rise

Given that automation is likely to happen, it will also change the job market in drastic ways. We have already seen that the automobile industry automated most of their production and big retail companies automated most of their warehouses. The next step will be automation of the transport industry. Tesla and other truck makers already presented self-driving

trucks and Uber will replace most taxis in big cities. But besides transportation more complex jobs are also likely to be replaced. Deep learning and neural networks have already managed to write reports that can not be distinguished from human writing. They are already used to find the most accurate length of a sentence in court and other tasks that were always thought to be only doable by humans. This means that also some high skilled work could be automated in the near future. There are a couple of legitimate responses to these concerns that we would like to address in the following. a) the unions will prevent unemployment. b) people can also educate and find other jobs c) the entertainment industry is unlikely to be automated soon, people can always find jobs here. To which we answer a) very unlikely, as we have seen with the carriage industry or manufacturing jobs during the industrial revolution, unions can delay progress but the incentives for companies are too strong to prevent it. b) Theoretically possible but the first people to lose their jobs will be low-educated taxi and truck drivers. It will be hard for someone who primarily drove a truck to get a university degree in computer science and service the automobiles stealing her job. c) As [7] points out there are already bots that write poems and music that are not distinguishable from human compositions, meaning that automation will happen here as well. But even if that was not the case: the market for art is very saturated in most countries already and many can't even make a living from this work. It is unlikely that demand will be rising in the same way as unemployment will rise.
In the end, there will be many jobs replaced and workers will not find new work quickly. Societies will have to adapt and find new big picture strategies to improve the life of their citizens.

C. Redistribution is unlikely to happen

One of the big picture strategies to solve these problems is society wide redistribution, i.e. high tax for the rich and a universal basic income. Here we argue that exactly these measures are possible to happen but significantly less likely than certain. In a competitive market individuals try to make as much profit as possible. They only pay wages because their employees are participating in the creation of a product and they pay taxes because individuals have leverage over them, i.e. through striking and unions. Through these two mechanisms, wages and taxes, the profits or products are redistributed over society. Both of these are gone or less likely in an automated world. Robots don't get wages, therefore the owner has a monopoly on profits. Taxes will be less likely because unions and employees cannot use their political capital as leverage anymore, i.e. through strikes. Additionally, a monopoly on profits for a few people also implies more political capital for these few. This might be done through lobbying for liberal policies or high donations to parties or financing their election campaigns. All of these mechanisms make it less likely that redistribution is happening to a sufficient degree and many people will be unemployed and poor.

V. Conclusion - Automation

Automation is likely to happen because of companies incentives and the rapid progress in machine learning techniques. The job market is disrupted in a big manner that implies high unemployment for low skilled jobs in the next decade and for some high skill jobs in two or three decades. It is also unlikely that unions can prevent this change or that now unemployed citizens will find other jobs through continuing training or advanced education. Other possible industries, such as entertainment, are likely to be saturated already and not a possible alternative for the many unemployed. Statewide strategies such as redistribution are likely to only provide an existential minimum, if at all, because workers and unions have no leverage over owners anymore, meaning that profit margins will be huge for very few. Their monopoly on profits makes redistributions even less likely, since their political capital can be used to further their interests. This is likely to result in an equilibrium in which redistribution happens to the degree in which the masses will not revolt but are not very happy. In the end societies are dependent on the good will and altruism of the rich, which, historically speaking, is rarely a good idea.

VI. Conclusion

We just presented two scenarios in which we examined what could go wrong with artificial general intelligence and automation. AGI will likely be super intelligent, unalignable with human goals and unlikely to be controllable once instantiated. Automation will likely lead to high unemployment, high inequality in wealth and redistribution will be hard. Both of these scenarios look grim but the purpose of this paper is to examine the worst case and attach probabilities to them. We think that most people are significantly too positive about a future with AI and automation and do not think enough about their safety and societal implications. We conclude that AGI will most certainly not work in ways that humans expect it to work and will lead to scenarios that are not aligned with human goals even when all participants in their creation have good intentions. Automation could be less problematic if redistribution happens, but is likely to result in an equilibrium of wealth inequality and unemployment.

References

[1] R. Miles, "The Orthogonality Thesis, Intelligence, and Stupidity," Jan. 2018.
[2] N. Bostrom, "The Superintelligent Will: motivation and instrumental rationality in advanced artificial agents," *Minds and Machines*, 2012.
[3] E. Yudkowsky, "AI Alignment: Why Its Hard, and Where to Start," Dec. 2016.
[4] R. Miles, "Why Would AI Want to do Bad Things? Instrumental Convergence," Mar. 2018.
[5] ——, "AI "Stop Button" Problem - Computerphile," Mar. 2017.
[6] McKinsey, "What the future of work will mean for jobs, skills, and wages."
[7] CPGgrey, "Humans Need Not Apply," *YouTube*, Aug. 2014.