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. 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 it realizes, that it has less costs to just print stamps and 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 an 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 will have a goal to maximize

B. AGI is superintelligent

By superintelligence we do not mean the difference between 70 and 130 IQ or the difference between an ant and a human but the difference of a scale that is hard to imagine to our minds. 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.

C. goals of AGI are unalignable with human goals

see rob miles

D. Goal is unchangable once the AGI is instanciated

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. Since it is significantly more intelligent than we are, it is very likely that they are successful in preventing these attempts.

E. AGI will escape controllable environments

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. Results

V. Conclusion

References

[1] C. E. Shannon, "Prediction and entropy of printed english," *Bell System Technical Journal*, vol. 30, pp. 50–64, Jan. 1951.