CHALMERS |  GÖTEBORGS UNIVERSITET

AI Risk

Under development

Author
Jonatan Hellgren

Supervisor: Olle Häggström
Examinator: Torbjörn Lundh

Institutionen för Matematiska vetenskaper
Gothenburg Sweden 2022

Abstract

ABSTRACT

Contents

1: Introduction

In recent human history we have seen a massive technological development, our lives today are severally different today compared to a century ago. Most of this development can be seen as the development of tools that we humans can make use of to carry on with increasing future development. In recent years artificial intelligence (AI) have been applied more commonly in the industry. This is currently due to it's outstanding ability to process massive amounts of data, which we are generating more then ever with the recent trend towards digitalization in our society. A key point here is that these systems still require humans to create and function them.

Many experts in the field of mathematics, computer science and even philosophy, believe that the future versions of AI will not require humans to function, they will be able to automate the human intelligence part as well. This is often referred to as artificial general intelligence (AGI). A significant difference with this shift is that it will increase the possible tasks that a single system can perform, in fact the amount of tasks possible would become arbitrary and they would be performed at human level of performance or higher. The effects of such a breakthrough could be on the same scale as the industrial revolution, but instead of automating physical labour we would instead have automated mental labour.

Several AI researchers have raised warnings for future development of AI, Stuart Russel and Max Tegmark, Eliezer Yudkowsky to name but a few. The reason for this concern is that with such massive amounts of power they can have, it would be catastrophic if it where to be used in the wrong way. This would of course be a higher risk after a AGI breakthrough, since it could then also happen due to a slight misspecification of the system during it's creation, which would lead to them ending up unaligned with ours. The consequences could possibly be existential.

In the upcoming century Toby Ord loosely estimates that the probability of an existential catastrophe is 17%, out of which 10 percentage points are due to unaligned artificial intelligence [precipice]. Thus AI alignment is something worth spending resources on for the sake of humanity.

As for how such scenarios could play out a common example is the *paperclip armageddon*. In which an paperclip maximizer is made super intelligent and starts accumulating resources such as hardware and money. Eventually the paperclip maximizer comes to a point where the existence of humans serves no purpose or possibly even having a negative effect on producing paperclips, and thus they become extinct.

The research field of creating safe AI has in the recent years literally exploded. We are a long way from solving the problem, most of what is being done today is mainly speculations and laying necessary foundations for future research. There are a lot of different subfields in this task and this report will specifically focus on the task of minimizing potential side effects on techniques we are applying today. Also with a purpose of spreading the ideas further, Nick Bostrom mentions in his book, *Superintelligence* , that this problem is "*... worthy of some of the next generation's best mathematical talent.*", and to attract those they need to at least know that they are needed. Bostrom

2: Background

2.1 Brief history of artificial intelligence

Definition 2.1.1 (AI)**.** Artificial intelligence (AI), is a computer program that is designed to solve a specific set of tasks. Usually these task require human level of intelligence. Since the set of task usually is limited it is also referred to as *weak AI* or *narrow AI*.

The idea of creating AI, has been around since the dawn age of computer, where one of the founders of of modern computer science Alan Turing being the first one to define the concept. The field of AI research can be dated back to the 1956 Dartmouth summer research project on artificial intelligence, JOHN McCARTHY....

The field has been studied ever since with some periods of breakthroughs and other with less innovation.

Sybolic AI

Expert systems

Neural nets

2.2 Emergence of artificial general intelligence

As previously mentioned in Definition 2.1.1, AI can be referred to as *narrow AI*, this is due to the fact that if we would apply an AI on a task which it have not specifically been trained on the performance would most likely be horrible. Take for example DeepMinds AplhaGo that won against the worl champion Lee Sedol in the game of Go, if we where to apply the same system on the task of sorting mail, it would fail spectacularly. The reason is that a team of brilliant researchers at DeepMind designed the model specifically to be good at Go[1].

This takes us to our next definition.

Definition 2.2.1 (AGI)**.** Artificial generall intelligence (AGI), is an AI that can solve an arbitrary task with as good or better performance then a human is capable of, the main difference from AI being that the set of task is not bounded. AGI is also referred to as *strong AI*.

If an AGI would have been created we it should for example be able to play a game of Go, then drive it's car to it job where it sorts mail and much more. The implications of this

[1] In more recent years DeepMind have released a new AI called AlphaZero which has a more general approach and is thus able to play Go, Chess and Shogi. The set of task it however still limited.

would likely be something similar to the industrial revolution, but instead of automating physical labour, we would instead automate mental labour.

As for predictions of when we are going to see the emergence of AGI there is a lot of uncertainty involved. WHEN EXPERTS GUESS. However with all the research and funding being focused on it, we are undoubtedly getting ever closer.

Since we have defined AGI to be able to solve an arbitrary amount of task, this would also include the creation of newer versions of it self. If it is better then the humans that created it at doing so and it keeps doing so recursively, an intelligence explosion would arise, also referred to as the *singularity*.

2.3 Potential risks and side effects

Instrumental convergence, instumental goals, terminal goals

Orthogonality thesis

2.4 Potential solutions

2.4.1 AI alignment

content

Reward functions are easy to misspecify

example of unaligned; toxic GPT-3

define inner and outer alignment

potential consequences of unaligned AGI

2.4.2 Impact measurements

2.5 Related work

3: Theoretical background

To get a understanding of how intelligent agents(DEFINE AGENCY) will behave in the real world we need to make a few simplifications in order to make the problem feasible. The first one being that instead of modelling the real world we are instead going to make use of Markov decision processes. The second one being that we will have to use Reinforcement learning to achieve intelligent behaviour for agents.

3.1 Markov decision process

A Markov decision process is a stochastic decision process, where the Markov property implies that the process is memoryless, meaning that the previous state do not have an effect on the next choice it only the current state that does. In mathematical terms it can be described as,

$$p(a|s_t, s_{t-1}, s_{t-2}, ..., s_1) = p(a|s_t),$$

where a is an action performed from state s_t in time step t.

Definition 3.1.1 (MDP)**.** An Markov decision process (MDP), is defined as a tuple $(\mathcal{S}, \mathcal{A}, R, p, \gamma)$. $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function, $p(s_{t+1}|s_t, a_t)$ is the transition probability from state s_t to state s_{t+1} given action a_t at time step t, γ is the discount factor typically defined in the range $\gamma \in [0, 1]$.

The process it kept going until either a terminal state is reached or until a certain amount of time steps have been reached. A terminal state is a state where the process terminates, this can be some sort of goal and would thus yield a reward, but it could also yield no reward or negative reward.

The discount factor γ has the important of describing how the agent values future rewards, with low values the agent favours more immediate rewards compared to future rewards, whereas for higher values the agent considers future rewards with more valuable. In environments with high uncertainty lower values of gamma might be more reasonable, since it might not be worth considering future rewards if they are not certain. The opposite holds for more deterministic environments where future rewards are of higher certainty, it might be a good idea to decrease the discount.

4: Methods

4.1 Simulations

5: Results

6: Discussion

7: Conclusion