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

1.1 Artificial intelligence

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 which we will in this report define as;

Definition 1.1.1 (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.

AI have in the recent years been applied in the industry more broadly, this is mostly due to the recent and impressive progress in neural networks, a model inspired by biological neurons which can learn useful correlations in massive dataset. The recent progress have in the recent years become a possibility due to more data being available, faster computer hardware and the massive amount of funding that is spent on research. Although these systems is often quite automated, 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, which brings us to our second definition;

Definition 1.1.2 (AGI)**.** Artificial general 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.

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 implications of such a breakthrough would likely be on the same scale as the industrial revolution, but instead of automating physical labour we would instead have automated mental labour.

Take for example DeepMinds AplhaGo that won against the world 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]. 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.

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

[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. Never the less, the set of task is still limited.

funding being focused on it, we are undoubtedly getting ever closer. Since AGI is defined 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, often referred to as the *singularity*.

1.2 Issues with AI

All tools can be applied in multiple ways, some might be beneficial and some might be ill intentioned. Take for example a hammer, you could either use it to build a house where you can live or you could use it to beat another person to death. The same is the case for AI because it still is but a tool, although the consequences might be more prominent since we do not understand the tool as well, and we thus can't guarantee that a well intentioned use of them won't cause negative consequences.

In the recent years we have seen some examples of how AI can have negative consequences from a certain viewpoint. For example the algorithmic bias we can see in models used by lawyers to determine how long of a sentence a felon would receive after committing a crime, it was shown that the models gave afro americans a significantly longer imprisonment. Another one being that social medias exploits our psychology with the help of AI to get our attention and keep us focused on them. These problems are alarming since if we have problem with the AI of today how is the future going to be when the potential power of them will likely be greater.

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. The main concern is that if the goals of the AI is unaligned with our goals This would of course be a higher risk after a potential 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 with goals that are unaligned with ours. The consequences could possibly be existential.

1.3 Basic AI drives

Although we do not yet know what will be the drive for a potential AGI, there have been a lot of work laying the foundations for it. A commonly adopted view is the Omohundro-Bostrom theory for AI driving forces. In it there are two corner stones, namely *instrumental convergence thesis* and the *orthogonality thesis*, which we will now explain further.

1.3.1 Instrumental convergence

Given a suffiently intelligent agent with a terminal or which can be seen as it's final goal. When the agent pursuis these goals it would naturally arise other instrumental goals, examples of such would be self-preservation, self-improvemnt and resource accumilation. The resoning behind this is that these things basically helps the agent in it's pursuit of it's final goal, the agent wouldn't be able to perfor it's goal if it where destroyed for example. These instrumental goals would likely be shared between a wide range of different agents, and thus there is a set of instrumental goals which agents would converge towards and hence the name.

To this day there doesn't yet exist any rigorous mathematical proof for this, it is still a well grounded speculation. Some work has however been done in trying to lay the necessary foundations for it. TURNER et al.

1.3.2 Orthogonality thesis

1.3.3 The implications of the Omohundro-Bostrom theory

1.4 Potential solutions

1.4.1 Considering future task

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: 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.

2.1 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 2.1.1 (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.

2.2 Reinforcement learning

2.2.1 Q-learning

2.2.2 Deep Q-Learning

2.2.3 Inverse reinforcement learning

3: Methods

3.1 Simulations

4: Results

5: Discussion

6: Conclusion