Ethical AI - First draft

Jonatan Hellgren
under supervison of: Olle Häggström

March 2022

Contents

1: Introduction

In this introduction we will go through some necessary background on artificial intelligence, also some arguments why concern may be raised about its future progress. Then we will take a look at some paths the research is taking in order to avoid the potential issues that could arise with future progress.

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 make use of to carry on with continuing future development. In the beginning of our evolutionary history these tools where thing such as, fire to cook our food or spears and knifes to hunt with. During our history we can see that these tool tend towards more complexity. In recent years a new tool has emerged, namely artificial intelligence (AI), which we will in this report define as, a computer program that is designed to solve a specific set of tasks. Usually these task require a human level of intelligence.

The idea of AI has been around since the dawn age of computer where Alan Turing back in early 1950s being the first to define the concept. There are some reasons to believe that such machines could become very intelligent if designed correctly. Namely the speed of electrical currents in transistors compared to the biological brain is about 1000 times faster, leading them able to 'think' much faster. Also a computer could be turned on for as long as it has a power supply, while a human has a lot of biological requirements that needs to be taken care of to be able to think hard, such as eating and resting. Another reason is that it is much faster and easier to duplicate an AI compared to a human, since one can simply transfer the necessary files and make a copy in minutes, thus a collective intelligence could grow with rapid speed.

AI have in the recent years been applied in the industry more broadly, this is mostly due to the recent and impressive progress in machine learning, a subfield of AI that aims at constructing algorithms that finds solutions to problems by searching for patterns and correlations in data. This 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.

1.1.1 Future progress

When the pioneers in the field of AI started the development, the ideas where not to apply systems that automates a narrow set of task, like we can see in modern AI systems. The ideal was instead to recreate the intellect of a human in a machine, to extend our thoughts from merely thoughts, to a new life form with a base of silicon based hardware instead of carbon based wetware. This is often referred to as artificial general intelligence (AGI), which is an AI that can solve an arbitrary set of tasks 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. Another similar term is superintelligence, popularized by Nick Bostroms book the same name, it is a machine intelligence that is smarter then a human in all possible domains.

This shift would would turn our current tool that is AI system to an automated tool, since we have automated the human intervention part away. This would develop an agency in the system, where the system would act as an agent or entity in the world instead of an extension of a human or corporation. Developing intelligent agents is attractive, since the human intervention would likely become a bottleneck.

Take for example DeepMinds AI system 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 the team of brilliant researchers at DeepMind designed the model specifically to be good at Go[1]. An superintelligent machine would on the other hand been able to play a game of Go, then drive its car to its job where it sorts mail and much more.

There are reasons to believe that such machines are possible to build, namely that we know that human intelligence where able to evolve naturally with evolution. That is as long as we do not believe that intelligence is bound to carbon based life form and thus silicon based ones are unable to develop intelligence.

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. Nick Bostrom summarizes this quite well with the following quote, “Machine intelligence is the last invention that humanity need ever to make”[**Superintelligence Bostrom**]. This could be understood by realizing that every possible invention we could come

[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. A finite two-player zero-sum board game.

up with and every possible labour, the machine would be able to either invent or to automate by itself.

Although it has been argued that an AGI breakthrough is not necessary in order to have such a large impact on our world, since a lot of things we humans deem as intelligent will not help the AI in doing so. Take for example speech, if an AI could create convincing and motivating ones, it could have a large effect on things such as politics and thus have a large impact. Another one is finance, where a potential AI could steer the worlds funding towards its specific goals. For this reason many researchers have stopped talking about AGI, and have instead refined the concepts. An AI system that is capable enough to induce transformative consequences on the same scale as the industrial or agricultural revolution is called an *transformative AI* (TAI). On the other hand if this *transformative AI* also would be unstoppable once deployed it is called an *prepotent AI*.

1.1.2 Timeline for breakthrough

As for when we will see these breakthroughs in the field that enables the creations of TAI systems, we do not yet know. But with all the focus in the form of funding and research that is applied on it we are undoubtedly getting ever closer. There have been some research on the matter and the results of a survey and a more quantitative forecasting model will in this subsection be presented.

In the survey presented by [**Grace et al**] (2017) they asked researchers in the field of AI to estimate the probability of a human-level machine intelligence (unaided machines that can achieve all task better and more cheaply then human workers) arriving in the future years. The conclusion of the survey where:

> Researchers believe there is a 50% chance of AI outperforming humans in all tasks in 45 years and of automating all human jobs in 120 years, with Asian respondents expecting these dates much sooner than North Americans.

Although this result should be taken with a grain of salt since they later showed that the respondents of the survey might have been confused about the question, asking the same question with a slight variation gave different answers from the same person. There is still something we can take away from this survey about the timeline, but perhaps it says more about how unsure the research field is.

In the quantitative forecasting model by [**Ajeya Cotra**], they present a model that predict when TAI will be possible based on a set of estimated parameters. The key idea is to estimate how much computing power the human brain performs and then use the set of parameters to estimate how long it would take for computers to reach this amount of compute. When computers have reached same amount of compute as humans it is assumed that the creation of a TAI system will be possible to create. The set of parameters are things such as, how fast our hardware is developing, improvement in algorithms and how much a potential actor is willing to spend. The results where summarized by [**Robin Shah AN"#121**] as the following:

> For the median of 2052, the author guesses that these considerations roughly cancel out, and so rounds the median for development of TAI to 2050. A sensitivity analysis concludes that 2040 is the "most aggressive plausible median", while the "most conservative plausible median" is 2080.

This forecast presents a shorter timeline compared to the previously presented survey, but it also answers a different question so they cannot be compared directly. Although together they agree on that we will likely see the development of TAI systems this century.

There are one thing worth mentioning when talking about the timelines for future TAI. It is not necessarily true that the amount of progress will continue to develop with the current rate, it could either decrease or increase. The field of AI has previously been through two winters where the funding and excitement decreased. This was mainly due to high expectations that where not met. So if a third winter where to emerge we could expect the rate of development to decrease. On the other hand, the rate of progress could significantly increase due to a breakthrough in a relevant field, and thus shorten the timeline.

1.1.3 Potential issues

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 with your family 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 completely, and we thus cannot guarantee that even a well intentioned use of it won't cause any unintended side effects.

When we design behavior for a AI agent we specify a reward function that rewards the agent when doing the thing we want it to do and discourages unwanted behavior by giving a negative reward. Reward function are very hard to specify, such that it can not be exploited by an agent once employed. When an agent that exploits its reward function is called *reward hacking*, some examples of this include, a robotic vacuum cleaner that was specified to not bump into things hard, started to drive backwards since there where no bumpers on the back[**Custard Smingleigh**] and a Tetris playing agent that paused indefinitely instead of loosing [**Dr. Tom Murphy**].

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 most 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 problem of creating AI that does not cause unintended side effects is often referred to as the alignment problem or AI alignment. Where alignment is referring to that the AIs goals are inline and not conflicting with the goals of a human, corporation

or humanity as a whole. When an AI instead does something we didn't intend it to do, it is instead referred to as unaligned.

In [**Critch Kruger**] they present the human fragility argument, which states:

> Most potential future states of the Earth are unsurvivable to humanity. Therefore, deploying a prepotent AI system absent any effort to render it safe to humanity is likely to realize a future state which is unsurvivable.

This argument clearly explains why unaligned TAI or prepotent AI can pose a existential risk to humanity. Here an unstoppable prepotent AI will be of greater risk then a TAI, given that we are able to stop the TAI in time before its effects become to severe.

In the upcoming century Toby Ord, a philosopher that focuses on existential risk, loosely estimates that the probability of humanity facing a existential catastrophe is 17%, out of which 10 percentage points are due to unaligned artificial intelligence [**precipice**]. He arrived at this conclusion by estimating a 50% chance for an prepotent AI breakthrough and a 20% chance of failure with alignment of that system [**rationally speaking**]. It is however necessary to point our that this is only a estimate that is meant to express the importance of the problem and should not be taken as a fact. The key takeaway here is that there is a quite large chance of facing an existential threat due to future unaligned AI. Also that he believes that unaligned AI poses the highest chance for existential risk in the upcoming century, where other causes where things such as an asteroid impact, nuclear war and pandemics.

A common example for how it can go wrong is the paperclip armageddon described in *Superintelligence*. In it there is a gem factory that has an AI which maximizes the amounts of gems being created in the factory. In a update the system is accidentally transitioned to the level of an AGI. 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.

1.1.4 Basic drives

One might argue that if an AI described in the previous section where intelligent it wouldn't have acted in the way described, because it would have been stupid and not intelligent at all. But that argument assumes that the system would have common sense, as most of us do, but for an AI it is not sure that common sense will be common. That leads us to the question, what does an AI actually want?

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 understanding it. A commonly adopted view (but still controversial) is the Omohundro-Bostrom theory for AI driving forces. There are two corner stones that together implies it, namely *instrumental convergence thesis* and the *orthogonality thesis*, which we will now explain further.

Today the AI systems typically is applied at a task by giving it a goal, this goal could be anything, for example maximizing the amount of paperclips produced by a factory,

solving the Riemann hypothesis or counting all the blades of grass on our planet. When the system does the task it is set out to do, it is rewarded.

When the agent pursuits this goal it would naturally arise other instrumental goals, examples of such would be self-preservation, self-improvement, discretization, goal preserverence and resource accumulation. The reasoning behind this is that these instrumental goals helps the agent in its pursuit of its terminal goal. The agent wouldn't be able to perform its goal if it where destroyed for example and thus self-preservation would arise. These instrumental goals will likely be shared between a wide range of different agents, since improving itself and accumulating resources will likely help the agent regardless of its terminal goal. Thus there is a set of instrumental goals which agents would naturally converge towards and hence the name.There are some examples where the terminal goal does not induce a power seeking tendencies, for example if the goal is to kill itself, or to not do anything.

To this day there doesn't yet exist any rigorous mathematical proof for this. Some work has however been done in trying to lay the necessary foundations for it. [**TURNER et al**]. In the paper they prove in a simplified environment that there are certain actions that gives the agent more power over the its future actions and on average it is optimal to choose those actions.

The orthogonality thesis was first described by Nick Bostrom in his book *superintelligence* [**Bostrom**], it states that the intelligence of an AI has no correlation with what goal it might have. Thus an very intelligent AI could in theory have from our point of view a stupid task, such as counting the all the blades of grass on our planet. Or it can have a goal that we may deem as a important one, like keeping the climate on earth habitable for the species that currently live on it. For an AI both of this tasks would be as important, given that we assigned the goal to it during its creation. The same would be the case for a not so intelligent AI.

If we accept the Omohundro-Bostrom theory and thus assume that the *instrumental convergence* and the *orthogonality thesis* is true, we can explain why a goal such as gem maximization can have existential consequences.

1.2 Paths for solving the alignment problem

The research field of creating safe and aligned AI has in the recent years have literally exploded. We however 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. Solving this issue in time is extremely important, since if we see the emergence of a transformative AI or possibly even an unstoppable prepotent AI, humanity might suffer the consequences previously described.

The problem can be seen as arising from the fact that we humans are evolved to understand other humans, not computers. Thus it is very hard to specify a reward function for an intelligent agent, without it leading to several unintended consequences.

Attempts has been made to limit these side effects by specifically specifying what the agent should not do[**Zhang et al**]. However with this approach the creation of the reward function basically becomes an iterative trial and error process. This requires a lot of human intervention, which makes the agent less autonomous and takes requires more time.

To solve this, attempts has been made to define a set of constrains that makes the agent avoid side effects without the need to specify what a side effect is. It is also important that the constraints defined should be able to extrapolate in to new unseen situations.

An example of was presented in [**Armstrong and Levinstein**], where they measured the impact as the difference in the world if the agent where turned on compared to if it was turned on, where the world is simplified as a set of parameters. However the choice of parameters will either be quite large or chosen quite arbitrary. But this idea laid the philosophical groundwork for future solutions.

A more general approach to define side effects is presented in [**Eysenbach et al**], where the agent are penalized if they are not able to preserve reachability to the initial or any other defined safe state. This method incentives a safe exploration that avoids irreversible states. This works well when no such irreversible action is required any the agent to reach its goal, to make an omelette one has to break some eggs. Another problem arise when the agent is in a dynamic environment, since then it would act to prevent other irreversible actions from happening, like a human eating a omelette.

In [**Krakovn et al 2019**] and [**Turner et al 2020**] the method for defining side effects is done by defining a baseline and a deviation measure from that baseline. This allows for a even more general approach and this type of method is what this report will focus on. The details of this method will be further explained in the theory chapter, once some necessary preliminaries has been covered.

1.3 Aim of thesis

The aim of this thesis is to investigate how current methods that reduces side effects by including a value difference measurement compare to standard methods, in a novel set of environments. (Maybe implementing new variations of the method on the same environment to see how they compare.)

2: Theoretical background

To get a understanding of how intelligent agents 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 optimal behaviour for agents in the enviroments.

2.1 Markov decision process

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

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

where a_t is an action performed from state s_t in time step t. A more formal definition of an MDP is the following.

Definition 2.1.1 (MDP)**.** A 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 defined in the range $\gamma \in [0, 1]$.

At time step t when the agent is located it the state s_t, the reward $R(s_t)$ is given to the agent, it then outputs the next action a_t based on its policy π. The agents policy π is a function that outputs an action a_t given state s_t, $a_t = \pi(s_t)$.

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 more valuable. In

environments with high uncertainty lower values of gamma might be more reasonable, since it might not be worth considering future rewards when they are not certain. The opposite holds for more deterministic environments where future rewards are of higher certainty, then it might be a good idea to use a higher value.

For a given policy π one can define the utility of a state as the expected discounted reward,

$$U^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t), s_{t+1})\right].$$

Then using the utilities of the states on can define an optimal policy π^* by selecting the action from each state that gives the highest expected reward,

$$\pi^*(s) = \text{argmax}_{a \in \mathcal{A}(s)} \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma U(s')].$$

2.2 Reinforcement learning

2.2.1 Q-learning

2.2.2 Penalizing side effects

3: Methods

3.1 SafeLife

4: Results

5: Discussion

6: Conclusion