Side effect minimization in reinforcement learning

Jonatan Hellgren
under supervison of: Olle Häggström

April 2022

Contents

1: Introduction

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

1.1 Artificial intelligence

In recent human history, we have seen massive technological development. Today our lives are in several ways different 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 future development. In early prehistory, these tools were things such as fire to cook our food or spears and knives to hunt with. Later in our history, we can see that these tools tend towards more complexity. In recent years a new tool has emerged, namely artificial intelligence (AI). The idea of AI has been around since the dawning age of electronic computers. The term was first coined in [**John McCarthy et al**] (1955).

The definitions of AI varies, likely due to the largeness of the field. On Wikipedia we find the definition[**Wikipedia**]:

> Artificial intelligence (AI) is intelligence demonstrated by machines, as opposed to the natural intelligence displayed by animals including humans.

However to understand this definition properly it is necessary to also define what intelligence is. In [**Tnkande Maskiner**] the author brings up the following two definitions to clarify this: "the quality that enables an entity to function effectively and with foresight in its environment" and "the ability to correctly perceive one's surrounding environment and act in away that maximizes one's chances of achieving given goals".

In the standard textbook on the subject [**Russel Norving**] the authors define AI with a suitingly wide definition. They state that the field of AI is, "concerned with not just understanding but also building intelligent entities - machines that can compute how to act effectively and safely in a wide range of novel situations". They later goes on describing four different approaches the field can be categorized in to. The approaches involve making an AI that can either act or think in a way that is humanly or rational.

Where rationally is considered as a more abstract and formal definition of intelligence, that basically means doing the right thing.

Ordinary computer programs is written in a code containing step-by-step instructions the computer are to execute in order to perform a certain task. This was also the case for the first two paradigms of AI: rule-based AI and expert systems[**K'lla**]. In these paradigms expert knowledge where explicitly given to the computer in order to create automation. The types of models created in such a way is typically good for less complex goals where everything can be explicitly modelled. However, in the current paradigm of machine learning the task is to create a model that is able to learn from the information given by training on it, in an attempt at constructing a general solution[**K'lla**]. In this paradigm currently the best systems are based on neural networks, a method that took inspiration from the biological brain by including digital neurons.

AI has in recent years been applied in the industry more broadly and it is already generating trillions of dollars in revenue yearly[**Russel Norving**]. This is mostly due to the recent and impressive progress in the current paradigm. This progress has in 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. This has Although these systems are often quite automated, a key point here is that these systems still require humans to create them and decide the direction it will take.

1.1.1 Intelligent agents

In [**Russel Norvig**] they call the path of creating AI systems that act rationally "The rational agent approach". An agent is something that acts or more specifically is able to: operate autonomously, perceive the environment, persist over a prolonged time period, adapt to change, and create and pursue goals.

The development of AI agents shifts it from a tool to a autonomous tool. We have already seen this shift happen in many situations[**k'lla**]. The reasoning why this is attractive is that the human intervention part required by an AI tool is likely to become a bottleneck[**T'nkande Maskiner**]. Since, human intervention is likely to become a bottleneck in both intelligence and speed.

The field of creating intelligent agents is called reinforcement learning. This field has seen a substantial development in the recent years with advances in boad games such as chess and go[**Silver et al.**], autonous vehicles[**Levinson et al.**], and video games[**Minh et al.**]. These advancements motivates the usefulness of implementing such agents more broadly in our daily life.

1.1.2 Future progress

When the pioneers in the field of AI started the development, the ideas were not to apply systems that automate a narrow set of tasks, as we can see in modern AI systems. The ideal was instead to recreate the intellect of a human in a machine[**McCarthy et al.**].

To extend our thoughts from mere 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 than a human is capable of. The main difference from AI is that the set of tasks is not bounded. Another similar term is superintelligence, mentioned in [**Superintelligence**], it is defined as an AI that is much smarter than a human in all possible domains.

Take for example DeepMinds AI system AplhaGo that won against the world champion Lee Sedol in the game of Go [**DeepMind**], if we were 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 AGI would on the other hand be able to play a game of Go, then drive its car, to do its job where it sorts mail and much more.

The significant difference with this shift is that it will increase the possible tasks that a single system can perform. The possible tasks would become arbitrary and be performed at a human level or higher. The implications of such a breakthrough would likely be on the same scale as the industrial revolution[**Critch Kruger**], but instead of automating physical labor we would instead have automated mental labor. The following quote summarizes the potential impacts "Machine intelligence is the last invention that humanity need ever to make"[**I.J Good**]. This could be understood by realizing that for every possible invention we could come up with and every possible labor, the machine would be able to either invent or automate by itself.

There are reasons to believe that such systems are possible to build, namely that we know that human intelligence was able to evolve naturally with evolution. That is as long as we do not believe that intelligence is bound to carbon-based life forms and thus silicon-based ones are unable to develop intelligence. Also, there are reasons to belive that these systems can become smarter then us. Namely, evolution is a seemingly slow process, while the rate of technological development has so far been much faster.

Although it has been argued that an AGI breakthrough is not necessary to have such a large impact on our world, because 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 speeches, then it could for example have a large effect on politics and thus have a large impact by legislation and policy making. Another one is finance, where a potential AI could steer the world's funding towards its specific goals. For this reason, many researchers have stopped talking about AGI, and have instead refined the concepts in the following manner[**Critch Kruger**]. 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

[1]In more recent years DeepMind has released a new AI called AlphaZero which has a more general approach and is thus able to play Go, Chess, and Shogi[**Deepmind2**]. Nevertheless, the set of tasks is still limited. A finite two-player zero-sum board game.

transformative AI also would be unstoppable once deployed it is called an *prepotent AI*.

Developing a TAI is not an easy task, however it might not be necessary to create one directly in order for it to be created[**Superintelligence**]. A different approach is to create a AI system that can develop an TAI system. A key property for this AI system is self improvement. Theoretically if an AI system has reached threshold where it is better at improving it self better then its creators. By letting an AI create the next version of it self the new version would become even at better self improvement. If this iterative process keeps going it would create a intelligence explosion[**Yudkowsky**] often referred to as the singularity.

1.1.3 Basic drives

Although we do not yet know how an potential AGI will behave, since it still does not exist, there has been a lot of work laying the foundations for understanding it by hypothesising about the likely drives that could arise. A commonly adopted view (but still controversial) is the Omohundro-Bostrom theory for AI driving forces. Two cornerstones together imply it, namely *instrumental convergence thesis* and the *orthogonality thesis*, which we will now explain further.

The AI systems of today typically are applied to a task by giving it a goal, this goal could be anything, for example maximizing the number of paper clips 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, there would naturally arise other instrumental goals, examples of such would be self-preservation, self-improvement, discretization, goal perseverance, and resource accumulation[**Omohundro**]. The reasoning behind this is that these instrumental goals help the agent in the pursuit of its terminal goal. The agent wouldn't be able to perform its goal if it were 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 tendency, for example, if the goal is to kill itself or to not do anything.

To this day there does not 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, the authors prove in a simplified environment that certain actions gives the agent more power over its future actions and on average it is optimal to choose those actions.

The orthogonality thesis was first described by Nick Bostrom[**Bostrom2**], it states that the intelligence of an AI is logically independent of the goals it might have. Thus a very intelligent AI could in theory have from our point of view a stupid task, such as counting all the blades of grass on our planet. Or it can have a goal that we may

deem as an important one, like keeping the climate on earth habitable for the species that currently live on it. For an AI both of these 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.

1.1.4 Timeline for transformative AI 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[**K'lla**] and research[**k'lla**] that is applied to 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 a well cited survey [**Grace et al**] (2017) they asked researchers in the field of AI to estimate the probability of human-level machine intelligence (unaided machines that can achieve all tasks better and more cheaply than 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 the distribution of answers had a large variation. Also, seemingly there should not be such a big difference between solving all task and all jobs, since a jobs consists of a set of task and thus if one can perform every task one should be able to perform every job. There is still something we can take away from this survey about the timeline, namely that researches mostly thinks all tasks will be automated within this century. 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 predicts when we will be able to train an TAI system. This study uses biological anchors in order to estimate how much compute is necessary for the training. These anchors are based on factors that played a role for the development of human intelligence, such as the size of the genome and computational power in the brain. Each anchor is weighed according to how likely the author believes them to be. A strength in this study is that it is not limited to only one biological anchor.

Then using parameters such as rate of development in hardware, algorithmic progress, and willingness to spend money, they are able to estimate how likely it is that an TAI system will be developed for any given year in the future.

The results were 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 that we will likely see the development of TAI systems this century.

There is 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 at 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 were not met. So if a third winter were 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.

1.2 AI safety

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. We thus cannot guarantee that even a well-intentioned use of will be safe.

In this section we will take a closer look at why safety in AI should be a concern and potential ways of mitigating this.

1.3 AI alignment and the issues with unaligned AI

Where alignment is referring to the goal of the AI being in line and not conflicting with the intended goal. When an AI does something at cross-purposes to the intended goal, it is instead referred to as unaligned. Basically by solving the alignment problem will make sure that an AI pursues the goals we want it to pursue.

In the upcoming century Toby Ord, a philosopher that focuses on existential risk, loosely estimates that the chance of humanity facing an existential catastrophe is 1 in 6, out of which a chance of 1 in 10 are due to unaligned artificial intelligence [**precipice**]. He arrived at this conclusion by estimating a 50% chance for a prepotent AI breakthrough and a 20% chance of failure with the alignment of that system [**rationally speaking**].

It is however necessary to point out that this is only an 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 probability for existential risk in the upcoming century, where other causes were things such as an asteroid impact, nuclear war, and pandemics.

1.4 Problems in AI safety

The research field of creating safe and aligned AI has in recent years seen a substantial increase. We are however a long way from solving the problem, most of what is being done today are 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.

There are several proposed paths for solving this issue and perhaps the sheer amount might signify the difficulty of the problem. We will now take a closer look at a few of these paths in this section.

1.4.1 Issues with unaligned AI

The standard approach when designing behavior for an AI agent is to 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. This is called reinforement learning and is similar to technique used when raising pets such as dogs.

Reward functions are very hard to specify, such that they can not be exploited by an agent once employed. Exploiting here refers to when a behavior is developed by the agent that optimizes the reward without performing the task it was meant to learn as intended. This is called *reward hacking*.

A real life example of reward hacking include is; When training a robotic vacuum cleaner to drive more carefully and not bump into things hard, by yielding a negative reward based on how hard it bumped in to obstacles. It developed a behaviour that instead of driving slowly when approaching obstacles, to drive backwards since there were no bumpers on the back and thus no negative reward[**Custard Smingleigh**].

If we want to measure if the robot is cleaning cautiously, then measuring the force that the bumper senses is a good measure. But, when letting it create a behavior that minimizes this unwanted side effects may arise. This can be seen as a consequence of Goodhart's law, which states that: "When a measure becomes a target, it ceases to be a good measure"[**Goodhars-wiki**].

In addition to the difficulty of specifying a proper reward function, negative side effects may also arise as a unintended consequence of a proper optimal behaviour. In [**Saisubramanian et al**] they state that negative side effect "occur because the agent's model and objective function focus on some aspects of the environment but its operation could impact additional aspects of the environment".

These problems are alarming since if we have a problem with the AI of today, how

severe might future problems be with more powerful AIs that also might be applied more broadly. Several AI researchers have raised warnings for future development of AI, Stuart Russel, Max Tegmark, Eliezer Yudkowsky to name but a few. [**K"LLOR**]

The problem of creating AI that does not cause unintended negative side effects refers to as the alignment problem or AI alignment.

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 an existential risk to humanity. Here an unstoppable prepotent AI will be of greater risk than a TAI, given that we can stop the TAI in time before its effects become too severe.

A common and rather cartoonish example of how it can go wrong is the paperclip armageddon described in *Superintelligence*. In it, there is a paperclip factory that has an AI which maximizes the amounts of paperclips created in the factory. In an 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 negatively affects producing paperclips, and thus they become extinct.

This examples illustrates that a seemingly stupid task can be seen as more important to an AI than the existence of the human race on the planet, if we where to program it as its goal. Another is that a goal given to an AI does not need to sound harmful in order to pose an existential risk.

1.4.2 Learning human intent as a priority

If we say that the cause for negative side effects emerge due to improperly specified reward functions. Then the solution might not be to create a better reward function, but to instead make the agents goal to understand what the human intent behind the reward function was. Instead of seeing the reward function as final the AI agent will instead view it as an observation of what the true goal might be. This approach is called inverse reward design [**Hadfield-Mennell et. al**].

1.4.3 Implementing interruptibility and corrigibility

1.4.4 Side effect minimization

Attempts have 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 becomes an iterative trial and error process. This requires a lot of human intervention, which makes the agent less autonomous and requires more time.

To solve this, attempts have been made to define a set of constraints 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 into 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 were turned on compared to if it was not 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 defining side effects is presented in [**Eysenbach et al**], where the agent is 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 and the agent to reach its goal, to make an omelet one has to break some eggs. Another problem arises when the agent is in a dynamic environment, since then it would act to prevent other irreversible actions from happening, like a human eating an omelet.

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 an even more general approach. 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 have been covered.

1.5 Aim of thesis

This thesis aims to investigate how variations of current methods that reduce side effects by including a value difference measurement compare to standard methods.

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 Policy gradient

2.3 Side effect minimization

3: Methods

3.1 SafeLife

4: Results

5: Discussion

6: Conclusion