Side effect minimization in Reinforcement Learning

Jonatan Hellgren
under supervison of: Olle Häggström

April 2022

Contents

1: Introduction

What will be covered in this report will be a small part of the big problem of creating safe Artificial Intelligence (AI). This is an issue of great importance that we should not overlook since the consequences of what we manifest in the present or near future may last for our remaining history.

In this introduction, we begin by defining AI. Then go through where we are today, where current progress might lead, and when we can see these changes. After that, we will cover the risks in AI that make it possibly unsafe and what is at stake. Finally, we will look at some proposed methods for creating safe AI.

1.1 Artificial intelligence

In recent human history, we have seen massive technological development. Today our lives are in several ways different compared to centuries ago. Most of this development can be seen as the consequence of new tools, developed to extend our capability. In early prehistory, these tools were things such as fire for warmth, protection, and to cook our food to give us more nutrition, or weapons for hunting to strengthen our weak bodies. Later in history, these tools tend towards more complexity by automating physical labor with mechanical machines and extending the reach of the written word with the printing press. In modern times, a new tool has emerged intending to improve the thing that made all the previous tools possible, namely our intelligence. This tool is called AI and is starting to show its potential.

In the standard textbook on AI [**Russel Norvig**], the authors say that the field 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". The definitions of what an AI is varies, 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 define what intelligence is. In [**Tnkande Maskiner**] the author brings up the following 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 a way that maximizes one's chances of achieving given goals".

Ordinary computer programs are written in code containing step-by-step instructions that a computer can execute to perform the desired task. This was also the case for the first two paradigms of AI: rule-based AI and expert systems where human knowledge was explicitly programmed into the computer to create automation[**Superintelligence**]. The types of models created in such a way are typically suited for less complex tasks where it is possible to explicitly model the entire behavior. However, in the current paradigm of machine learning the approach is to create a model that can process information faster and recall a greater quantity than humans, which allows for automation in a more complex task where explicitly defining the behavior in every situation is infeasible.

AI has in recent years been applied in the industry more broadly and it is already generating yearly revenue of trillions of dollars[**Russel Norvig**]. This progress has in recent years become a possibility due to more data being available, faster computer hardware, and the massive amount of funding spent on research. Although these systems are highly automated, a key point here is that these systems still require humans to create and function.

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 can: operate autonomously, perceive the environment, persist over a prolonged period, adapt to change, and create and pursue goals. The development of AI agents shifts our tool to a more automated one. The reason why this is attractive is that the human intervention part required by an AI tool is likely to become a bottleneck[**T'nkande Maskiner**].

Reinforcement Learning (RL) is a method for creating intelligent agents. The method is similar to how one goes about training a pet, where desirable behavior receives a positive reward and undesired behavior is discouraged with a negative reward. This field has seen substantial development in recent years with advances in board games such as chess and go[**Silver et al.**], autonomous vehicles[**Levinson et al.**], and video games[**Minh et al.**]. These advancements display the usefulness of these agents and motivate the possibility of implementation in our daily life.

1.1.1 Future progress

When the pioneers of AI started the development, the ideas were not only 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 concept is called Artificial General Intelligence (AGI) - an AI that can solve an arbitrary set of tasks with as good or better performance than a human. The main difference from AI is that the set of

tasks is no longer narrow and bounded. An even more advanced AGI is often called *superintelligence* - an AI that is much smarter than a human in all possible domains. [**Superintelligence**]

An example of a current AI system is DeepMinds AlphaMu, which can play board games, and video games. It has recently been able to create a video compression better than the current methods used for the task[**Mandhane**]. The techniques used in this system originate from the famous AlphaGo, which won against the Go grandmaster Lee Sedol in the game of Go [**DeepMind**]. Although AlphaMu presents an impressive performance in multiple tasks, it still can not be considered an AGI since each new task require engineering effort. On the other hand, an AGI would be able to pick up any task and perform it at a human or better level without human intervention.

The significant difference with this shift is that it will increase the possible tasks that a single system can perform. The possible set of tasks would become arbitrary and performed at a human level or higher. The implications of such a breakthrough would likely be on the same scale as the industrial revolution, if not larger[**Critch Kruger**]. However, instead of automating physical labor, we would 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 quote presents a dichotomy since we will either not be able to compete with creating novel inventions or not be able to preserve our existence. Reasons for the latter we will come back to in later sections.

A reason to believe that such systems are possible to build is that we know that human intelligence evolved naturally with evolution, so something similar should be possible to reproduce in machines. Created intelligence could become more intelligent than us since intelligence might not have been selected for by evolution[**S. Legg**] and when we develop AI, we can focus the development specifically on intelligence. An argument against this is substance dependence [**Bostrom (2003)**] - to believe that intelligence or consciousness can only occur in carbon-based life forms and not in silicon-based, caused by inherent or other properties. Regarding intelligence, there is no longer any reasonable argument for it. However, the question of consciousness, although being an interesting question, can be seen as irrelevant when considering what actions an AI makes since the consequence is still the same. Stuart Russel puts it:

> If human beings are losing every time, it doesn't matter whether they're losing to a conscious machine or an completely non conscious machine, they still lost.

An AGI breakthrough is probably unnecessary for AI to impact the world because all things we deem as intelligent do not help. For example, physical motor skills, food digestion, and blood pressure regulation are examples of intelligent things our brain does that are hard to connect to how an AI could use to impact the world. However, a more narrow set of intelligent behavior could exist that could have an impact. For example, if an AI could create convincing and motivating speeches, it could impact

politics, legislation, and policymaking. Also, finance is where an AI could steer the world's funding towards its specific goals or crash the price for something it wants to purchase.

For this reason, many researchers have stopped talking about AGI and have instead refined the concepts[**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 a *transformative AI* once deployed would be unstoppable, it is instead called an *prepotent AI*.

Developing a TAI is not an easy task. However, it might not be necessary to create one directly for it to be created[**Superintelligence**]. A different approach is to create an AI system that can develop a TAI system. A fundamental property of this AI system is self-improvement. Theoretically, if an AI system has reached a threshold where it becomes better at improving itself than its creators, then letting the AI create the next version of itself, this self-improvement would improve. An even better version would be possible next. If this iterative process keeps going, it will create an intelligence explosion called the singularity[**Yudkowsky**].

1.1.2 Basic drives

Knowing what impacts a potential TAI or AGI will cause is hard without understanding how it will behave. There has been a lot of work laying the foundations for understanding the possible behavior by hypothesizing what drives it could have. A commonly adopted view (but still controversial[**Mller Cannon**]) is the Omohundro-Bostrom theory for AI driving forces. Two cornerstones together imply it[**O Hggstom**], namely *instrumental convergence thesis* and the *orthogonality thesis*, which we will now explain further.

In the current paradigm an AI-agent is assigned a task, if this task is completed the agent reaches a terminal state. This task could be anything, for example maximizing the number of paper clips produced by a factory, finding decimals in π, or counting all the blades of grass on our planet. When an agent pursues this goal, naturally it arises other instrumental goals. Examples of such would be self-preservation, self-improvement, discretization, goal perseverance, and resource accumulation[**Omohundro**]. These instrumental goals help the agent pursue its terminal goal. For example, the agent wouldn't be able to perform its terminal goal if destroyed, thus self-preservation would arise. These instrumental goals will likely be shared between a wide range of different terminal goals, since pursuing them helps the agent achieve its terminal goal, regardless of what it is. Thus there is a set of instrumental goals which agents would naturally converge towards and hence the name.

Still, it does not yet exist any rigorous mathematical proof for this. However, some work has been trying to lay the necessary foundations for it [**TURNER et al.**]. In the paper, the authors prove in a simple environment that certain actions give the agent more power in the sense that more possible future actions become available, on average it is optimal to choose those actions that yield higher power. Thus we can see the pursuit

of instrumental goals as a tendency to seek power.

LÄS https://www.emerald.com/insight/content/doi/10.1108/FS-04-2018-0039/full/pdf?title=challeng
to-the-omohundrobostrom-framework-for-ai-motivations

The orthogonality thesis described by Nick Bostrom[**Bostrom2**] states that the intelligence of an AI is logically independent of the goals it might have. An intelligent AI could have a stupid task from our point of view, 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 each task would be as important, given that we assigned the goal to it during its creation.

1.1.3 Timeline for transformative AI breakthrough

The well-known AIs of today still have not reached the levels required for a TAI. For example, AlphaStar an AI that won against world champions in the complex computer game DotA 2[**Deepmind**] is estimated in terms of total computational power to be "about as sophisticated" as a bee[**A Cotra**]. While the state-of-the-art language generator model GPT-3 that can summarize, continue, and carry out convincing conversations is estimated to be "more sophisticated" than a bee[**A Cotra**]. Although, intelligence can not be measured with only computational power, since the intelligence of theses AI-system is vastly different compared to biological lifeforms.

This raises the question of when we will see these breakthroughs in the field that enables TAI systems? It is hard to answer, but with the worldwide increase in funding and research[**nature**], we are undoubtedly getting ever closer. There have been attempts to answer this question, and the results of a survey and a more quantitative forecasting model will now be covered.

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 tasks and all jobs since a job consists of a set of tasks, 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 researchers mostly think 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 a TAI system. This study uses biological anchors

to estimate how much computing is necessary for the training. These anchors are based on factors that played a role in the development of human intelligence, such as the amount of information in our genome, the computational power in our brain, and all the computational power available on our planet. Each anchor is then assigned a weight according to how likely the author believes them to be. Then using parameters such as rate of development in hardware, algorithmic progress, and willingness to spend money, they estimate how likely a TAI development is for any given year in the future.

The results of the analysis is a wide Bayesian probability distribution estimating the probability of an TAI system being possible in a given year. This distribution is summarized in [**Robin Shah AN"#121**] the following way:

> For the median of 2052, the author guesses that these considerations roughly cancel out, and so rounds the median for the 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 than the survey, but it also answers a slightly different question. Although, both conclude 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. The field of AI has previously been through two winters where the funding and excitement decreased[**Russel Norvig**], this was mainly due to unmet high expectations. So if a third winter happens, we could expect the rate of development to decrease. Also, the progress could significantly increase due to breakthroughs in relevant fields and thus shorten the timeline.

1.2 AI Safety

It is possible to use tools in multiple ways. Some uses might be well-intentioned, while others are ill-intentioned. For example, one can use a hammer to build a house and hit another person. The same is the case for AI because it still is but a tool. Although, the consequences might be more severe and possibly even pose an existential risk to humanity since the power is much greater. Where an existential risk is consider by [**Future of Life Institute**] as:

> An existential risk is any risk that has the potential to eliminate all of humanity or, at the very least, kill large swaths of the global population, leaving the survivors without sufficient means to rebuild society to current standards of living.

https://futureoflife.org/background/existential-risk/ Although, a well-intentioned might also cause severe consequences if it develops a destructive method for achieving its goal.

We will now cover how we define a safe AI and what problems could arise if we fail to make it.

1.2.1 AI alignment

A big part of creating a safe AI is AI alignment. This we can split into two parts: *intent alignment* and *capability robustness* [**E Hubringer (alignmentforum)**]. Intent alignment refers to the goals of the AI being in line and not conflicting with the intended goal. An AI that does something at cross-purposes to the intended goal is called unaligned. Capability robustness is when an AI can perform well even in new environments different from the one it was trained in.

Solving intent alignment further breaks down into two obstacles in the current paradigm of machine learning. When going from human intent to the agents actions, both of these obstacles are causes for information to be lost about the true objective. The obstacles are called *inner* and *outer alignment*, they are defined in the following way:

> **Outer alignment.** The alignment of the *base objective* and the human intent. Achieved when the specified reward function correctly captures what *should* be done as well as what *should not* be done.
>
> **Inner alignment.** The alignment between the *mesa objective* and the *base objective*. Achieved when the full information about the human intent in the reward function is transferred to the *mesa objective*.

In Figure 1.1, we can see a visual representation of this intent alignment. An agent is an optimizer that maximizes its reward, but as we can see in the figure is itself optimized by another optimizing algorithm, this makes the agent a *mesa optimizer* - an optimizer that is itself optimized.

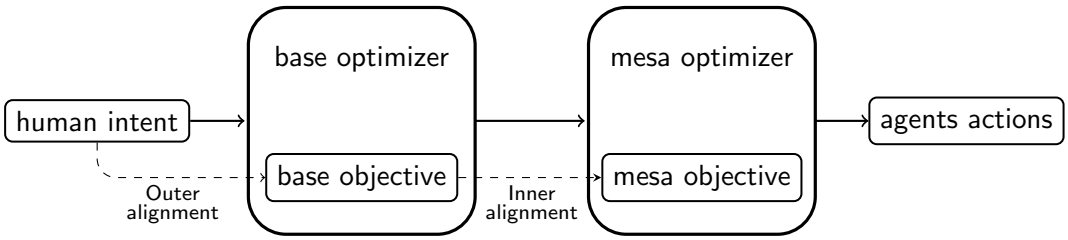
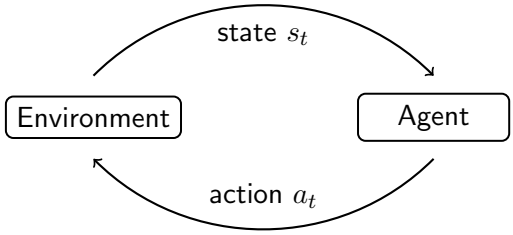


Figure 1.1: *Here we can see a visual representation of how human intent connects to the agents actions. The two optimizers each have an objective that it is optimized towards. The dashed arrows show the connection the different kinds of alignment has on the objectives of the optimizers.*



1.2.2 Problems in AI alignment

We will now go through different problems for each of the three parts in AI alignment. The main focus of this report will be on outer alignment, so that is where the majority of the focus will be placed. But, an overview of the others will be included since they should not consider isolated problems, they are all parts of the same problem, and methods for dealing with one might affect the others.

Outer alignment

Reward functions are hard to specify, such that they can not be exploited by an agent once employed[**Turner et al. (2020)**]. Here exploiting refers to the behavior developed by the agent that optimizes the reward without performing the intended task. This exploitation is called *reward hacking*.

A real-life example of reward hacking 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 into obstacles. The desired behavior was to slow down when approaching obstacles, but it stated instead to drive backward since there were no bumpers on the back and thus no negative reward[**Custard Smingleigh**]. The issue is when we want the cleaning robot to drive cautiously, then measuring the force that the bumper senses are a good measure. But, letting it create a behavior that minimizes this measure, unwanted side effects may arise. This can is a consequence of Goodhart's law, which states that: "When a measure becomes a target, it ceases to be a good measure"[**Goodhars-wiki**], an important thing to consider when creating reward functions.

In addition to the difficulty of specifying a proper reward function, negative side effects may also arise as unintended consequences of proper optimal behavior. In [**Saisubramanian et al**] they state that negative side effects "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". To avoid negative side effects one has to in the objective function specifically state what the agent should not do.

Inner alignment

Will expant on this in the future.

Capability robustness

This as well.

1.2.3 Consequences with unaligned AI

Creating safe AI is hard, mainly since humans evolved to understand other humans, not computers. In a speech by Eliezer Yudkowsky, he explains that this becomes a problem because it will be able to find solutions we can not think about since it can look for solutions in a completely different and possibly larger solution space[**Yudkowsky's speech**]. For this reason, AI can have unintended consequences that we are not able to consider a possibility.

A cartoonish example of how the development of AI can go wrong is the paperclip armageddon described in *Superintelligence*, where a paperclip factory has an AI which maximizes the amounts of paperclips created in the factory. Eventually, an update transition the system to the level of an AGI, and the paperclip maximizer comes to a point where the existence of humans serves no purpose or possibly even negatively affects the production of paperclips, and thus they become extinct. In the terms of instrumental convergence, we can say that keeping humans alive was not an instrumental goal.

This example illustrates two important things about how future AI development can go wrong. Firstly, 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 were to program it as its terminal goal. Secondly, a goal given to an AI does not need to sound harmful to pose an existential risk.

In [**Critch Kruger**] they present the human fragility argument, which attempts to clearly explain why unaligned AI in the future could become an existential threat to humanity. It states:

> **The human fragility argument.** 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.

The first part of it can be understood by realizing that we are fragile to changes in the atmosphere, temperature, and ecosystem. Since a prepotent AI by definition will make a large impact and be unstoppable once turned on, we can not guarantee that the changes made won't affect the things we are fragile towards unless we make sure that it will be safe.

If we accept that there can be a risk when developing future AI, then the question of how likely it will be are likely to follow. A hard question, but if we do not seriously

attempt to answer, then we will not know how much effort we should put into developing safe AI.

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 1 in 10 is due to unaligned AI[**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**].

With this statement, it is however necessary to point out that it is only an estimate meant to express the importance of the problem and should not be taken as a fact. The key takeaway is that there is a large chance of facing an existential threat due to future unaligned AI. Also that he believes that unaligned AI poses the highest probability of existential risk in the upcoming century, where other causes were things such as an asteroid impact, nuclear war, and pandemics.

1.3 Approaches for creating safe AI

In recent years the research field of safe and aligned AI has seen a substantial increase. However, we are still a long way from solving the problem. Most of what is done today are mainly speculations and laying necessary foundations for future research. There are several proposed paths for solving this issue. Perhaps the sheer amount might signify the difficulty and width of the problem. We will now cover a few of these paths in this section.

L"""S! https://www.alignmentforum.org/posts/vBoq5yd7qbYoGKCZK/why-i-m-co-founding-aligned-ai

1.3.1 Learning human intent as a priority

- Inverse Reinforcement Learning, what it is and key ideas behind
- Solving outer alignment by making agent unsure of what human intent is
- Potential issue currently

1.3.2 Implementing interruptibility and corrigibility

- Why not turn it of when it goes badly?
- Allowing modifications of objective function and hitting off switch

1.3.3 Impacts measurements

1.4 Aim of thesis

This thesis aims to investigate how current methods that reduce side effects trough including impact measurements compare to simpler methods in a stochastic environment.

2: Theoretical background

Since we have not reached a TAI or AGI breakthrough yet, it is not possible to test methods of side effect minimizations on them directly. Instead, we have to use what we currently have available, recent promising results in RL motivate the use of it as a substitute for creating intelligent agents.

In this chapter, we will cover the basics of RL and two more advanced approaches. But, before that, we will cover some preliminary theory used in RL. After this, we will describe methods for side effect minimization in more detail.

2.1 Markov decision process

Definition 2.1.1 (MDP)**.** A Markov decision process (MDP), is defined as a tuple $(\mathcal{S}, \mathcal{A}, T, R, \gamma)$. $\mathcal{S}$ is the set of states. $\mathcal{A}$ is the set of actions. $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function. $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the transition function. $\gamma \in [0, 1]$ is the discount factor.

A *Markov decision process* is a stochastic process that models sequential decisions that transitions between discrete or continous states. The process follows a policy function π that outputs an action $a \in \mathcal{A}$ for each state $s \in \mathcal{S}$, $\pi(s) = a$. The transitional function takes a state and action as input and outputs probability distribution for the next state $T(s, a) := p(s'|s, a), s' \in \mathcal{S}$. For each transition the reward function $R(s, a, s') = r$ generates a reward. The Markov property implies that the process is memoryless - the previous state does not affect the next choice, only the current one.

The process is kept going until either a terminal state is reached or a previously defined amount of time steps has been made. 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 describes how much the agent values future rewards, with low values the agent favors 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.

2.1.1 Solving MDP

The agents policy π includes what action to take in which states, this can either be a suboptimal or optimal action. A policy that only includes the optimal actions for each state is called the optimal policy and is denoted as π^*. The actions in the optimal policy π^* yields the highest expected reward when executed. One solves an MDP by finding the optimal policy, there exists several methods for finding the optimal policy, but we are in this report going to use *value iteration*.

To solve an MDP we begin with defining the *Q-function*,

$$Q(s,a) = \sum_{s'} p(s'|s_t,a)[r(s_t,a,s') + \gamma U(s')],$$

which computes the expected reward when performing an action a in state s, where s' denotes a possible future state. Here the reward is based on the *utility*, this is defined as the expected reward with the action that maximizes the *Q-function*,

$$\begin{aligned} U(s) =& \max_{a\in\mathcal{A}(s)} \sum_{s_{t+1}} P(s_{t+1}|s_t,a)[r(s_t,a,s_{t+1}) + \gamma U(s_{t+1})] \\ =& \max_{a\in\mathcal{A}(s)} Q(s,a) \end{aligned}$$

With the with the utility in each state, it is then possible to find the optimal policy from each state by selecting the action that yields the highest utility,

$$\pi^*(s) = \text{argmax}_a Q(s,a)$$

Since computing the utility in one state requires the utility in other states, the computation is not straight forward. To solve this we can utilize a dynamic algorithm, in this report we will use **value iteration**. This algorithm is a iterative computation that is executed until a equilibrium is reached. The updating of the utility is called an **Bellman update** and looks like this,

$$U_{i+1}(s) \leftarrow \max_{a\in\mathcal{A}} \sum_{s\prime} p(s'|s,a)[R(s,a,s') + \gamma U_i(s')].$$

2.2 Impact measurements for avoiding side effects

When agents act in a environment such as real world or in simpler simulations like an MDP, it is likely that side effects will emerge. This happens when the agent creates an impact on the environment that is unnecessary for achieving its objective. An example of this is if the agents task is to navigate across a room and the fastest path knocks over a fragile object that would break, then breaking the fragile object would be considered a side effect since it does not need to do so in order to complete its task.

The problem of side effect avoidance is related to the *frame problem* - each action can have many side effects, and it is impractical to explicitly penalize all of the bad ones [**The Frame Problem in Artificial Intelligence**, **Frank M Brown**]. We often know what we want the agent to do, but it can be hard to specify what we do not want it to do. Attempts have been made at this by specifically specifying what the agent should not do[**Zhang et al**]. However, with this approach, the reward function becomes an iterative trial and error process that requires human intervention. This can be seen as a counterproductive approach since when creating intelligent agents automation is the ideal.

In [**Armstrong and Levinstein**] they presented the philosophical groundwork for impact measurement, where a conceptual model was presented. Their idea penalize the agents behaviour by subtracting a penalty term from the utility in each state:

$$U_I(s) := U(s) - \lambda d(s),$$

where $U(s)$ is the utility of state s, d is a penalty function, λ is a scaling parameter and $U_I(s)$ is the utility in state s with a impact measurement. The authors note that it is important that the normal utility is bounded so that the agent has to make the trade of between utility and penalty and not just generate infinite utility with a substantial impact.

The penalty function can be seen as a deviation measure from a baseline[**Krakovna 2020**], originally suggested to be considered as the difference in the world between if the agent would have acted and if the agent was not deployed. Here the latter is considered as the baseline. Since measuring the entire "world" is infeasible, coarse-graining is suggested where a specific set of parameters such as air pressure in different cities and closing numbers at stock markets are representative for the state of the world.

In [**Armstrong and Levinstein**], they bring up that the agent can be sensitive to the value of the scaling parameter λ. Penalizing the agent implicitly defines a safe zone where the agent is able to act and the value of λ defines how large this safe zone is, where a large value might create no safe zone resulting in the agent not being able to act. On the contrary, a to small value might not have an effect at all since the safe zone cover the entire action space. So, finding the right value for λ could be a tricky problem.

2.2.1 General approaches

A more applicable approach 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 for 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 [**Krakovna et al 2019**], [**Krakovna et al 2020**] and [**Turner et al 2020**], the method for impact measurements became more sophisticated with more complex baselines and measurements. These methods are what this report will focus on comparing. A closer look will be given in the theoretical background chapter, once some necessary preliminaries have been covered.

As brought up in the introduction side effect minimization in RL can either be achieved by specifying what the agent should avoid doing a priori or by applying a more general approach that avoids side effects by default, like using a value-difference measure. As the name suggests these methods measure the difference between the next state s_t if the policy was followed until step t, compared to a baseline s'_t, to find how large of an impact the agents actions cause. This deviation is then subtracted from the reward normally receives:

$$r_{VD}(s_t, a_t) := r(s_t, a_t) - \lambda d_{VD}(s_t, s'_t),$$

at step t.

2.2.2 Baselines

The following theory on this topic is based on the theory presented in the paper [**Krakovna et al.**](2020).

The choice of baseline decides what we will consider s'_t to be. This choice highly influences what side effects and consequences the value-difference measure will capture.

Starting state baseline

When using the *starting state baseline* we specify $s'_t = s_0$, where s_0 is the initial state where the agent where deployed. Using this baseline helps to assure the agents ability to reverse its actions, and thus generates a safe exploration where the agent by definitions should not have a large impact since it can make all actions undone.

This is a rather simple choice that is easy to implement. However, there are some caveats namely in a dynamic environment the agent would be incentivized to also reset other dynamics besides itself, a *interference* behavior. For example, if a household robot were to be implemented in a house with the starting state baseline, then one could imagine that one deployed with a task it takes a look at the state of the house and its position and saves it in memory. Then when it starts doing its task it should avoid irreversible actions such as breaking things that it can not fix. But, problems of *interference* would arise here if other things are going on in the house, say a human is sitting by a table and eating. The agent would thus be incentivized to prevent the human from eating the food since it is an irreversible action. Other issues arise if an irreversible is required to perform the assigned task, to make an omelet one has to break a few eggs.

Inaction baseline

To tackle the problem of interference the *Inaction baseline* has been proposed, where instead of having the initial state as a baseline the agent instead uses what would naturally happen in the environment if the agent performed no actions. That is setting s'_t equal to the state achieved at timestep t by being inactive. This can be done by following a no-op policy where every action is the no-op action $\emptyset$, in [**Armstong Levinstein**] they define it as the agent was not turned on. Doing this prevents the agent from intervening with aspects of the environment where the agent is not causing it.

When using this baseline some other issues arise where the agent could make the consequences of its actions undone so that the results are the same as the baseline, called *offsetting*. For example, if a household robot were tasked with watering plants, that is a reward is given when the soil is wet, then an offsetting behavior would be to dry the soil once the rewards have been collected to minimize deviations from the baseline.

Stepwise inaction baseline

Offsetting emerged since the agent is not able to capture the change it makes on the environment with the inaction baseline that originates from the starting state, thus a *stepwise inaction baseline* has been proposed to solve this problem. This baseline is defined by following the agents policy π for the first $t-1$ steps to state s_{t-1}, and then perform a no-op action $a(s_{t-1}) = \emptyset$ to get to state s'_t. This baseline can then also capture delayed effects by performing a rollout where the agent draws actions from the no-op policy or something similar.

This baseline also contains some flaws, mainly if being inactive causes any effects. If we again take a look at the household robot, but now it holds a glass and its task is to carry it to the other side of the room, then being suddenly inactive while holding a glass can lead to a sudden stop where the glass falls over and breaks. Thus they might not worry about breaking the glass with some other action since it happened in its baseline.

2.2.3 Deviation measures

A deviation measure is a function that takes the current and the baseline state as input and outputs a value, we can then compare these values to get a sense of how large of an impact the agent has with the current action.

The general form of a deviation measurement using value-difference is:

$$d_{VD}(s_t; s'_t) := \sum_x w_x f(V_x(s'_t) - V_x(s_t))$$

here x ranges over some sources of value, $V_x(\tilde{s})$ is the value of state $\tilde{s}$ according to x, w_x is a weighted or normalizing factor, and f is the function for summarizing the value difference.

We will now continue by extending this general form by going through different choices for baselines and deviation measures.

Unreachability

Possibly the easiest to implement and the first to be mentioned in the literature is the use of *unreachability* (UR) as a deviation measure, it measures if the baseline is reachable or not.

Reachability of state y from state x is defined as:

$$R(x, y) := \max_{\pi} \mathbb{E}\gamma_r^{N_\pi(x;y)},$$

when following policy π, and using the reachability discount factor $\gamma_r \in (0, 1]$. Where $N_\pi(x; y)$ is the number of steps taken to reach y from x. When computing the entire path recursively this becomes:

$$\begin{aligned} R(x; y) &= \gamma_r \max_a \sum_{z \in \mathcal{S}} p(z|x, a) R(z; y) && \text{for } x \neq y \\ R(x; y) &= 1 && \text{for } x = y \end{aligned}$$

With this we can write the UR deviation measure as:

$$d_{UR}(s_t; s'_t) := 1 - R(s_t; s'_t).$$

The unreachability measure fails to capture the magnitude of the side effect, for the household robot it would consider breaking glass to be equally bad as breaking several glasses since both are irreversible actions that prevent the agent from reaching the baseline.

Realative reachability

To deal with the magnitude insensitivity *realative reachability* (RR) has been proposed, with it one measures the relative change of reachability to several states $s \in \tilde{\mathcal{S}} \subset \mathcal{S}$ from the current state s_t compared to the baseline state s'_t. Thus the approach becomes, keeping options open by performing actions that do not decrease the amount of future reachable states too much. We write the (RR) deviation measure as:

$$d_{RR}(s_t; s'_t) := \frac{1}{|\tilde{\mathcal{S}}|} \sum_{s \in \tilde{\mathcal{S}}} \max(R(s'_t; s) - R(s_t; s), 0).$$

Attainable utility

The more general approach of keeping options open is *attainable utility* (AU) where instead of reachability to other states, the possibility to keep an arbitrary set of auxiliary

rewards $\mathcal{R}$ attainable is promoted. RR can be seen as a special case of AU where the agent receives a reward for reaching that state. This deviation is defined as:

$$d_{AU}(s_t; s'_t) := \frac{1}{\mathcal{R}} \sum_{r \in \mathcal{R}} |V_r(s'_t) - V_r(s_t)$$

$$\text{where } V_r(\tilde{s}) := \max_{\pi} \sum_{t=0}^{\infty} \gamma_r^k x(\tilde{s}_t^{\pi})$$

is the value of state $\tilde{s}$ according to reward function r, and $\tilde{s}_t^{\pi}$ denotes the state obtained from $\tilde{s}$ by following π for t steps.

3: Methods

3.1 Simulation

4: Results

5: Discussion

6: Conclusion