Responsibility Gap

Mischa

December 21, 2021

Contents

1 Introduction

This is a rough sketch of the beginning. It is very bad and not ready and I probably will change most of it, but right now it kinda gives an introduction into the whole topic.

Nowadays it is an obivious statement to make that we live in a time in which technology is ubiquitous and new technology is being developed at an unprecidented rate. It penetrates our society and is one of the adhesives that hold 'the system' in place. We must only envision a world in which cars do not exist; or refrigerators; or the internet; or x-ray machines to see how much of our (everyday) lives depends and is shaped by it. I also don't think that I go out on a limb when I say that we integrate some technology relatively fast into our lives.

In this work I will discuss technology and moral responsibility and how the two relate to each other. Specifically, I will investigate the different ways we can seek for moral responisbility in situations where an autonomous machine is involved.

Put more here.

But first, let us examine the traditional way of how we ascribe moral responisbility in situations where technology is involved:

Suppose the following situations: A person hits another person with a hammer and kills them. A newly installed dam breaks and a city is flooded. A hacker manages to get access to a digital banking system through his own computer and steals a good deal of money.

The hammer, the dam and the hacker's computer are technology that is directly involved in morally critical situations. Yet, we abstain from blaming these artifacts for what has happened in the respective situations. We also do not put the events off as natural tragedies, as we do when a storm destroys a house or an avalanche kills a skier in the mountains. We naturally ascribe the responsibility for the events to the people behind the technology. The person who wielded the hammer, the architect of the dam, the hacker. These people used the technology as a tool to achive their own end and they are responsible for the effects, that the technology has on our world, whether they achieve the end or (as in the case of the dam-architect) not. To view technology as tools or instruments used by humans and the humans as the ultimately responsible entities for the technology is what Heidegger calls the *instrumentalist definition of technology* (Heidegger 1977).

"We ask the question concerning technology when we ask what it is. Everyone knows the two statements that answer our question. One says: Technology is a means to an end. The other says: Technology is a human activity."(Heidegger 1977, p.4)

Thus, according to the instrumentalist definition, technology is something that is intimately connected to humans and inherits its moral standing from the quality of the *human* end that it tries to achieve and from the way it is used by *humans*. Technology is not moral on its own. Only in the context of human

ends and actions. If it is humans that decide to use technlogy for some end it is naturally them, who are responsible for the results.
A very similar line of reasoning is layed out by Sullins in what he calls the user, tool and victim model (Sullins 2006, p. 152).

In the advent of machine learning we are facing a type of technology that is intentinally becoming more and more autonomous, making decisions on its own without any human supervision, without any human being able to predict or at least explain those decisions. The Machines are essentially black boxes making decisions and affecting the world in morally significant ways:

Autonomous Vehicles are being developed to populate the streets and naviagte dangerous situations. Machine Learning Algorithms analyse our behaviour on the internet, recommend content that they find we would be interested in and influence us in this manner. There is a multitude of applications for ML in health care. We can easily imagine that medical practitioners will increasingly rely on tools that diagnose diseases and even propose treatments. Eventually, patients might even cut out the middle man and receive their medical care from artificial physicians. There already are services that provide a kind of psychotherapy by texting with a chatbot. [zitieren] Autonomous Weapon Systems (AWS) are being developed. The aim is to create war robots that can be sent into the battle field and they would be able to decide on their own whether to kill a target or not.

But what is the problem here? Why don't treat the vehicle, the recommender algorithms, the health care systems, the war robots in just the same way as the hammer, the dam or the computer from the examples above? Why not again hold the operator or the manufacturer responsible? What is the difference? Why this paper?

Include something like this:

Moreover, the situation makes us ask more fundamental questions: How do we justify our practives of ascription of responsibility? What are morally responsible agents. What are they responsible for? Depending on how we answer these questions, results in different answers regarding the responsiblity gap and how we deal with autonomous machines... bla bla bla

1.1 The Dirty Problem

The answer for these questions is the challenge that these technologies pose to our traditional ways of ascribing responsibility.

According to Matthias (Matthias 2004) the reason why we can hold either the manufacturer or the operator of a machine responsible for what it effects in the world, is because we can sensibly say that they are the moral agents who were in control of said machine. Matthias claims that responsibility necessitates control:

"[An] agent can be considered responsible only if he knows the particular facts surrounding his action, and if he is able to freely form a decision to act, and to select one of a suitable set of available alternative actions based on these

facts." (Matthias 2004, p.175)

This means that we can only hold someone responsible for something they have done, if they had sufficient control over their action. This is widely referred to as the Control Requirement (CR) [zitieren].

Converesly, if an agent does not have sufficient control, we can acsribe at most partial responsibility, if any, to them.

This notion of control as a precondition for responsibility seems to complement the intrumental theory: The manufacturer and operator have control over their machines, thus they are the ones who are responsible if something happens because of the machines. It is then very clear how to assign responisbility in critical situations: If the operator uses the machine in accordance with the manufactureres specifications and something goes wrong, we say that the manufacturer is responsible. If the operator deviates from the manufacturers specifications and something goes wrong, we say the operator is responsible (Matthias 2004, p.175).

Enter Machine Learning: We are now in a time where computer scientists and engineers work on increasing the autonomy of their programs and machines by using techniques that can be bundeled by the term *machine learning* (ML). 'Autonomy' in this sense means that we allow the technology to make its own decisions based on its prior experience without the programmer or user knowing what the decision is going to be or understanding how it came about.

This leads to a change in the classical roles of the manufacturer (programmer) and operator (user) insofar that the amount of control they exert over the machines diminishes as the machines become more and more autonomous. Arguably, our traditional ways of ascribing responsibility are challenged. If we agree with the CR and

The real question: Now here comes the question of this work: Who can sensibly be held responsible for the actions of an autonomous machine? Who is responsible if a driverless car runs over a person? Who is responsible if an artificial physician proposes a wrong treatment for a patient? Who is responisble for a war crime commited by an autonomous weapon system (AWS)?[1]

Matthias says that our current practices of ascribing moral responsibility fail at finding an appropriate target when an autonomous machine is strongly involved in a situation. He calls this problem *the responsibility gap.*

The considerations above, do not entail that our current practices of dealing with the effects machines have on our world must necessarily change, but rather that the contemporary and foreseeable developments in AI and ML challenge our current practices and motivate their reevaluation.

On the following pages I will first give descriptions of various accounts for moral responsiblity and will then proceed to describing and debating the dif-

[1] Talk about asymetry of blame and reward. Refer to later in next section perhaps, because here I mention responisbility only in the sense of blaming someone

ferent approaches philosophers have proposed for dealing with the assumed responisbility gap.

2 The Basics

Before jumping into any analysis of the responsibility gap as described above, it makes sense to first explore what I mean, when I speak of the two things that give this work its title. Moral responsibility and autonomous systems. Additionally will draw attention to two different types of views bla bla.

2.1 Two Perspectives

The process view and the property view. I would like to introduce an idea by Daniel Tigard, that will serve as very helpful tool to classify different approaches discribed in this piece. Though Tigard relates that idea to specifically moral responsibility, I will try to broaden its application. The pattern that is unveiled by that idea will follow us throughout the following pages in different forms and variations, but it is still unmistakable. When Tigard speaks of models of moral responsibility he speaks that we may have a *process view* or a *property view* on it CITATION. While I will examine what that means for moral responsibility with greater detail later on , I want to explain what these two possible views shall mean for us.

The two views are lenses that people use to explain some things we attribute to each other. This seems like a very vague statement but allow me to elaborate: What are these "things" that I mean? Well I mean stuff like consciousness, moral responsibility, moral agency, intentionality, in short: all of those fine concepts that philosophers hold so dearly by their hearts. From the property view then, these things have some sort of metaphysical truth to them. The subjects that we ascrive these things to, are said to have some kind of real *property* that gives rise to the thing and our perception of it. Without this property on cannot truthfully say that the thing is there. Relating to moral responsibility, Tigard says that "*being* responsible [is] conceptually prior to being *held* responsible." To be rightfully held responsible, one must truly be responsible. To be considered conscious, one mus truly by conscious. On the other hand, we have the process view. The process view is another way of saying "It's a social construct". According to it, these things that we ascribe to ourselves and each other are the results of a process of social and individual negotiation. The thing is not born from some fundamental property, but from subjectively *being regarded as* existent. "*[H]olding* is conceptually prior to *being* [...]".

I understand if these two views lack any substance right now, but please bear with me. Wou will lear to see them in what the people write. And with that we shall continue.

2.2 What is Moral Responsibility

While we generally have an intuitive understanding of what we mean by (moral) responsibility, it is important for the following discussion to have a more rigorous account of the term. The necessity of clarifiying the term becomes clear, when we expose its ambiguity in our everyday language. Sometimes we use moral responsibilty to say that someone has some sort of moral obligation to do something. Sometimes we use the term to say that we blame someone. Sometimes, to denote that a person is accountable for a certain action, attitude or event in a sense that allows us to appraise them on its basis. And on and on...

Angela Smith describes 3 different meanings that the sentence "*A* holds *B* responsible for *X*" can have: (A. M. Smith 2007, p. 469):

1. A thinks that B is open to moral appraisal because of X.

2. A thinks that B is culpable and therefore blameworthy because of X.

3. A blames B for X.

For the sake of this work we are mostly interested in Smith's first sense of the sentence. Thus, being responsible for something means that *one is open to moral appraisal for it*. In other words, when I am responsible for something, it means that I am an appropriate target of blame, praise or other moral responses because of it. PERHAPS THIS IS A GOOD MOMENT TO TALK ABOUT THE IMBALANCE IN R. As we will see later, the same of very similar definitions of responsibility can be found in other philosophical works on the topic. We will, thus, continue working with it.

Now that we have agreed on what we mean when we speak of moral responsibility, the questions that still remain are: What are the things that we are responsible for? What are the conditions that need to be fulfilled to be morally responsible for something (A. M. Smith 2008, p. 370)? Let us take a look at two accounts that try to answer that question.

2.2.1 Strawson

In "Freedom and Resentment" P.F.Strawson gives and account of our moral practices and tries to explain the mechanisms behind them. These mechanisms lay the groundwork for what can be understood as moral responsibility. In the centre of Strawson's argumentation lies "the very great importance that we attach to the attitudes and intetions towards us of other human beings [...]"(Strawson 1962, p.5). In other words, we care a lot about how other people treat us. We like it, if other people treat us with what we interpret as respect and goodwill and we do not like it, if other people treat us with what we interpret as illwill or indifference. Depending on how other people treat us and which attitudes we ascribe to them, we in turn develop and adjust our own attitudes towards them. Strawson calls the attitudes we form as a reaction to other

people (quite fittingly) our *reactive attitudes*. Examples for such attitudes are resentment, indignation, gratitude. These reactive attitudes form the basis for our practices of blaming and praising other people.

Example 1. Matt is seventeen and likes playing computer games. His ten year old brother Charly often watches him play and frequently asks Matt, if he can play too. Matt usually denies Charly's request. Charly finds this unfair because Matt can play so much and he can only watch. Charly develops slight resentment against his brother because in his eyes, Matt does not care enough about him to fulfill Charly's wish of playing. Eventually, Charly runs to his mother and complains about Matt's unwillingness to allow Charly to play on the computer.

The primitive example above portrays the mechanism, Strawson tries to describe. In the situation Charly interprets that his brother, Matt, treats him with an attitude he does not like: indifference. This prompts Charly to develop resentment towards Matt as a reactive attitude. Charly's going to his mother and complaining about Matt is his way of blaming Matt.

Strawson stresses the importance of attitude behind an action, for we evaluate other people and their actions strongly on the basis of their attitudes and intentions. The same action with different attitudes elicits different reactions from us. Strawson gives the example of someone stepping on his hand. If P-Boy found that they did it accidentally and they were sorry for injuring him, he would feel the pain in his hand, but probably no (appropriate) resentment towards them. If, on the other hand, he found that they stepped on his hand out of malevolence or were indifferent to what had happened, Strawsons reaction would include some kind of resentment towards the other person. The same is true, for when another person benefits us in some way. The degree of gratitude we would feel towards them would differ, depending on whether they did it on purpose and out of good will or accidentally (Strawson 1962, p.6).

I should also point out, just like Strawson repeatedly does (Strawson 1962, p.5, p.7), that the way reactive attitudes work is much more complicated than can be explained in this text. There is a complex interplay between different parties and the attitudes vary on a broad spectrum as well as in intensity.

The type of reactive attitudes I have described until now is generally about close personal interactions with other people. They develop because of the way other people treat specifically *us*. However, reactive attitudes are not only a personal phenomenon but are also developed and affected by how the objects of these attitudes treat other people. Thus, Strawson introduces another class of reactive attitudes, which he calls *vicarious* or *impersonal* reactive attitudes. These attitudes target the behaviour or will of others independent of who is affected by them. To be clear: These impersonal reactive attitudes can also be developed if *we* are the suffering party, but "[...] they are essentially capable of being vicarious"(Strawson 1962, p.15)

Strawson proceeds and gives these vicarious reactive attitudes the qualifier *'moral'* and the objects of such reactive attitudes are said to have done some-

thing that has moral value (positive or negative) to us (Strawson 1962, p.15). And thus, Strawson has linked the concept of moralitiy with his reactive attitudes.

Example 2. Clara likes to read the newspaper in the morning. Today, she finds an article about a CEO of a big international company and how he knowingly choses suppliers that violate human rights to drive the price of their commodities down. Clara does not like this behaviour.

It is clear that Clara is not personally affected (at least not directly) by the behaviour of the CEO. She still develops a reactive attitude towards him on the basis of his indifference regarding humabn rights and the people who suffer because of it. What Clara experiences is moral indignation.

To sum it all up: According to Strawson, we expect from other people that they behave in accordance with attitudes of respect and goodwill. Depending on whether they cohere with these expectations, we exhibit resentment or gratitude (reactive attitudes) towards them. We blame or praise other people on the basis of these reactive attitudes. Morality comes into play, when we acknowlidge that we expect certain behaviour not only towars us, "[...] but towards all those on whose behalf moral indignation may be felt [...]" (Strawson 1962, p.16).

In light of this account, moral responsibility does not seem to be a metaphysical entity. From a Strawsonian perspective, to be morally responsible can be interpreted as being an appropriate object of vicarious reactive attitudes (N. Smith and Vickers 2021, p.3) (Matthias 2004, p.175). Tigard takes moral responsibility in this regard "as a social funtion of [...] reactive attitudes" (Daniel W. Tigard 2020, p.3). These definitions cohere very well with the one I already mentioned above: To be responsible means to be open to moral appraisal.

Before moving on, Strawson, introduces another idea, which might be important for our further discussion on the main topic of this work, the responsibility gap. He describes in which cases reactive attitudes are mitigated or even not exhibited at all. Strawson distinguishes two general groups of such cases:

1. Cases of the first group are those where the source of injury is a moral agent but their explanation for their action can be summarised with the sentences 'I didn't know', 'I had to do it' or something similar (Strawson 1962, p.7-8). Tigard descibes these cases as situations "where the agent is normal, but the circumstances are abnormal [...]" (Daniel W. Tigard 2020, p.5).

   Examples of such cases are the gentleman who accidentally steps on someones foot because the train is too full and he tries to navigate through the crowd, or the doctor who has lost a patient and is then rude to her husband. The people who suffer the injury usually tend to modify their reactive attitudes to fit the circumstances. TIGARD HAS A SIMILAR EXAMPLE

2. The second group is again nicely described by Tigard as cases "where the circumstances are normal but the agent is abnormal" (Daniel W. Tigard

2020, p.5). Strawson speaks of children or schizophrenics or people that act out of compulsion (Strawson 1962, p.8-9). Such agents cannot be appropriate targets of reactive attitudes because the expectations upon which the reactive attitudes are based cannot be reasonably targeted towards them. According to Strawson, it is unreasonable to expect moral behaviour from someone who is morally deranged or underdeveloped. In this sense, they are not moral agents and cannot be treated as such. We do not see them as members of the moral community (Strawson 1962, p.18). The attitudes we exhibit towards them differ accordingly compared to those who are members of the moral community. We see them as "object[s] of social policy; as [...] subject[s] for [...] treatment; as something certainly to be taken account, perhaps precautionary account, of; to be managed or handled or cured or trained; perhaps simply to be avoided [...]" (Strawson 1962, p.8). Seeing an agent as such, implies that we portray a second set of attitudes towards them. Strawson calls these attitudes *objective attitudes* (Strawson 1962, p.9).

I want to reiterate: To have reactive attitudes towards someone *means* to view them as a fully responsible agent. In Strawson's eyes these are the same things (Strawson 1962, p.23). To have objective attitudes towards someone (or something) *means* to view them outside of the moral community and, thus, to view them as an inadequate target for ascribing responsibility. #foreshadowing *It means that only full moral agents (members of the moral community) can be morally responsible for their actions.* This is a very important insight that will become relevant later on.

2.2.2 Smith

I already have repeadetly used such phrasings as 'inadequate target' or 'appropriate object' of moral responsibility or reactive attitudes or blame or praise. However, the attentive reader will find that Strawson's account of our moral practices focuses strongly on our external perceptions of other's internal attitudes.

In this regard, we are prone to say, that someone is an appropriate object of, for example, blame, if (1) we see them as a member of the moral community (we can develop reactive attitudes towards them) and (2) we *interpret* their attitudes as malevolent or indifferent towards us. But what about the cases where our interpretation is wrong? We might think their action is an expression of ill will towards us, but by looking beneath their sculp, we might see that it is actually not the case and we had misinterpreted their attitude. Intuitively, it would not be fair to blame someone, if their *actual* attitude would not correspond with our *perception* of their attitude. Or their action was subject to circumstances unbeknownst to us. We hold them responsible and blame them for the action. But upon learning more about the circumstances we change our mind and judge the person to be not responsible anymore. In fact, we say that they have not been responsible at all even for the time we thought they were

responsible. Does this not show that there is a sense of being responsible that goes beyond 'being held responsible' by others? Does this not show that we in general do believe in a kind of responsibility relies less on our perception and more on the truth of the situation? Smith argues that there is a difference in being held responsible and *being* responsible. And it would not be fair for us to hold someone responsible, if they 'in reality' are not responsible (A. M. Smith 2007, p. 472).

We can take Strawsons account of responsibility as a description for when people are *held* responsible in Smiths sense. But how can we make sense of people *being* responsible?

There are two opposing stances on that. Unified approaches vs pluralistic approaches. Unified approaches are... Pluralistic approaches are ...

I will briefly present Smiths unified apporach of moral responsibility in order to get a sense of what I generally mean by people being responsible. Still, we shall not forget, that this is only one of many accounts of moral responsibility of there.

But how does Smith then make sense people *being* responsible? She proposes an approach, which she calls the rational relations view CITE.

According to Smith, to be responsible for an action, attitude or mental state, one must have a specific connection to it (A. M. Smith 2008, p 370). What is this connection? The connection cannot be that the action, attitude or mental state is attributable to me. I am not responsible for feeling hungry, a person with epilepsy is not responsible for having seizures, even though these are a things that can properly be attributed to me and Epilepsy-Eric (A. M. Smith 2012, p. 584). Thus, we are not responsible for everything that is attributable to us. One might now come up with the idea that we are responsible for our conscious choices, but Smith argues that the condition of volition does not satisfyingly cover the domain of responsibility. We might be responsible for actions and attitudes we deliberatly choose to take or have, but in general we can also be responsible for actions and attitudes that are not deliberate but spontaneous and involuntary. An example that Smith brings up is her forgetting her friends birthday. She did not choose to forget the birthday, she did not undergo a thought process that weighed the pros and cons of forgetting the birthday and then arrived at the conclusion that it would make sense to ignore the birthday. It just happened. She called her friend as soon as she remembered congratulated her and apologised for forgetting. Of course her friend forgave her but the implicite assumption still was that Smith was responsible for forgetting the birthday. In general, people are also considered responsible for their emotional reactions and arguably these are not subject to deliberate choice. One could argue now that, if not choice, control is what makes someone respnsible. And by control I mean that a person has the theoretical control over the things they are doing, perhaps through past choices, perhaps through the ability to change a certain aspect of oneself in the future. After all, Smith had the control to write down her friends birthday in her calendar and install a reminder and she has the

control over making sure that such a thing will not happen again. This sounds very similar to Matthias' control requirement (Matthias 2004, p.175) that we have already introduced in the first section. For Smith, this account, though plausible, is not satisfying (A. M. Smith 2005, p. 251). Because what we find bad is not the fact that Smith did not take any measures to be reminded of her friends birthday, but rather that her forgetting her friends birthday shows (on the surface) that Smith does not value her friend enough to remember it. The assumption is that if Smith had judged the friendship to be important enough, she would have had thought of that significant date.

These kinds of judgements are the basis of Smith's rational relations view. Let us examine how she develops her idea.

According to Smith, people make certain judgements of "value, importance, or significance" (A. M. Smith 2005, p. 251). These *evaluative judgements* can be abstract and form individual normative ideals, like valuing freedom more than security, or they can be more concrete like judging spiders to be dangerous. The judgements people make should *rationally*[2] lead to certain behaviour in accordance with the judgements (A. M. Smith 2005, p. 244, p. 247, p. 250). Thus, the attitudes and actions are a direct reflection of evaluative judgements. Smith argues, if a persons behaviour is based on such an evaluative judgement, they are "open, in principle, to demands for justification" for their behaviour (A. M. Smith 2012, p. 577-578). This is what she calls *answerability*. So, people are answerable for their behaviour, if they are theoretically able to reference a judgement that the behaviour was expressing and to defend and justify that judgement. Further, she states that moral appraisal of an agent "always embodies (at least implicitly) a demand to her to justify herself" (A. M. Smith 2012, p. 578).

Let us now connect all the dots: Being responsible is being open for moral appraisal. Moral appraisal *always* encompasses demands of justification. Such demands only make sense, if a person is answerable for the appraised action or attitude, meaning "that the thing in question must in some way reflect [the persons] judgement or assessent of reasons" (A. M. Smith 2015, p. 103). According to Smith, people are responsible for all and only those things, for which they are also answerable (A. M. Smith 2005, p. 251, p.256).

An important property of these *evaluative judgements* is, that they must not *necessarily* be on the conscious radar or arrived at though deliberate thought. They can also be sponataneous judgements that the person only forms or discovers when being confronted with a new situation (A. M. Smith 2005, p. 251-252).

Example 3. Melissa has never thought much about her becoming a victim of sexual assult. It is not a topic that crosses her mind in general. One night she walks home through a dark alley and she spots a man walking behind her. To her own surprise, she finds herself being afraid of that man. The fear makes

[2]I find that Smith uses the word "rational" in a very loose sense. She probably does not mean rational in the idealised and logical way, but rather in a more holisitc sense. If my interpretation is correct the word "reason-giving" as it is used by Shoemaker is a bit more fitting (Shoemaker 2011, p. 23).

Melissa walk faster with the hope of getting more distance between her and the man and getting home faster.

In the example, we see Melissa discovering her judgement that a man can be potentially dangerous to her. The judgement arises sponatneously without her having thought much about forming it and it elicits certain attitudes and actions in Melissa. She experiences fear and walks faster because of it. Notice also that Melissa's judgement is open to critical assessment; As soon as she is aware of her judgement she has the possibility to think about the judgement and decide whether it is justified or not. Now, according to Smith, there is a rational relation between Melissa's judgement and her attitudes and actions. Melissa's actions and attitudes are expressions and reflections of her judgements. Melissa can thus justify her behaviour by referencing and defending the evaluative judgements that caused it. She is answerable for her behaviour. And that is what makes her responsible for it.

To summarise: When we think that someone is morally responsible for something, we *might* demand a justification for their conduct before we appraise them. The assumption is that their conduct is a direct result of their explicit or implicit judgement. Our appraisal is then formed on the basis of their justification for their judgement or, in other words, their answer to our demand. If it is reasonable to make such a demand for justification the person is said to be answerable. And according to Smith, people are responsible (open to moral appraisal) for all and only those things they are answerable for.

2.3 What are the machines we will be talking about?

Now that we have understood what moral responsibility can mean, we shall look at the other component of the topic. The machines, the autonomous systems, the robots, the AIs. I use these terms almost interchangeably. What is important here is that these technologies have the capacity for autonomous action. Matthias says that the question about responsibility gap arises because humans have less and less control over intelligent machines (Matthias 2004, p. 175). It seems reasonable to say that the loss of control results from an increase in the machines autonomy.
Robert Sparrow explicitly says that the problem with the ascription of moral responsibility exists for truly autonomous machines (Sparrow 2007, p. 64-65). For Sparrow, being autonomous means to have internal states (like beliefs or desires) and to be influenced by them. Moreover, an autonomous agent is able to "form and revise these beliefs themselves" (Sparrow 2007, p. 65).
Thomas Hellström implies that with higher intelligence the robots will become more autonomous[3] and that this will lead to responsibility issues.

So what is this mysterious property called *autonomy* that seems to be so important? Well, unfortunately I don't really have the time to extensively dive

[3]Hellström uses a little bit of a different terminology: For him, autonomy is an absolute property. He introduces the concept of autonomous power, which describes the range in which an agent is autonomous. Though I find this a fair account of autonomy, we shall not make that distinction. For the sake of simplicity, we shall see autonomy as a gradual property.

into this broad topic. Instead we shall remember the working definition from above, as it seems to caputure all the features of autonomy that are relevant for the topic.

"*'Autonomy' in this sense means that we allow the technology to make its own decisions based on its prior experience without the programmer or user knowing what the decision is going to be or understanding how it came about.*" (Taken from page 4 of this very document).

ANOTHER NOTION FOR AUTONOMY: AUTONOMY MEANS TO BE ABLE TO CHOSE YOUR OWN GOALS (SPARROW TRUE AUTONOMY)

What is important is, that the autonomous decisions are in some sense intelligent but also somewhat unpredictable and untransparent.

The question of the responsibility gap is asked because of these properties but, as we will see later, the answer may depend on other properties as well.

We will encounter technologies that range from simple artificial neural networks to war robots that can decide who to kill to truly conscious AIs[4]. It seems that the domain, in which these systems reside, is multi-dimensional and gradual (Misselhorn 2018, p. 75). During the whole exploration of the responsibility gap, this is something that we should keep in mind.

One last remark, before we continue:

I will not go into any specific implementational or engineering details about the relevant technologies. This analysis will take the technologies and their properties on a conceptual and philosophical level. IS THAT TRUE? CHECK AT THE END. MAYBE ADD A 'MOSTLY'.

3 How Big is the Gap

The story sofar:
We have defined moral responsibility to be an openness for moral appraisal. Matthias says that to be morally responsible for something, one must be in control of it. According to Strawson, the way we blame and praise other people is the result a social negotiation, which is based on actions, intentions, attitudes and reactions. Smith goes a bit further and tries to give a more rigorous account of the conditions to moral responsibility. She says that people are responsible for all and only those things that are rationally caused by a preceding evaluative judgement. Let us now turn back to the responsibility gap, as it is described above. Matthias claims that there is a responsibility gap, because it is not clear how to ascribe responsibility for actions performed by intelligent autonomous systems. The problem is that new intelligent autonomous technologies act more and more without human control and supervision, but can still have morally

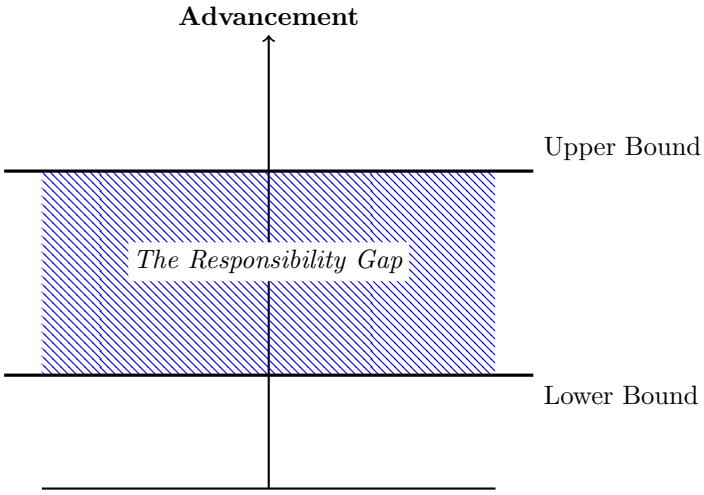
[4]I will not discuss the nature of conscious AI and other questions that dive into the conceptual and technological sensibleness and/or possibility of it.

relevant impacts. Though Matthias makes this claim on the basis of the control requirement, I suspect that what makes us wonder about the ascription of responsibility is our moral intuition and not any sophisticated philosophical explanation of our moral practices. We see the new technology and understand that it is different from anything else that humanity has produced and we ask the question "How do we deal with this?" And specifically we wonder, who is responsible when a machine causes any harm? Our inability, insecurity or hesitation to answer this question is exactly what the responsibility gap denotes. This question can be asked regardless of the philosophical account of moral responsibility one supports, but it may be that depending on the account the answers can vary widely.

In "Killer Robots", Robert Sparrow provides a more concrete mental model for the problem. He, says that there is a "conceptual space"

In "Killer Robots", Robert Sparrow assesses the ethics of autonomous weapon systems and what happens when they commit something that would be considered a war crime. Who would be responsible for it? According to Sparrow there are three sensible cadidates to ascribe responsibility to: The programmer, the commanding officer and the machine itself (Sparrow 2007, p. 69-71). This seems to me like a fair account even beyond autonomous robot warfare as a similar dissection of the relevant actors can be found in Deborah Johnsons "triad of intentionality" (Johnson 2006, p. 202). It seems that the three targets (programmer, user, machine) are also intuitively where to look for responsibility and we will mainly focus on them.

Sparrow further says that with increasing autonomy and capacity of the AI system, the locus of responsibility changes from the involved humans to the system itself. However, the transition between these stages is not instant. According to Sparrow, these stages are separated by a "conceptual space", where the ascription of responsibility is problematic (Sparrow 2007, p.74). A system, that falls into this interspace would be on the one hand too autonomous for a human to be responsible for it - on the other hand not autonomous enough to be responsible for itself. We can give this space a name: The responsibility gap.



The diagram nicely visualises what Sparrow means. The Advancement-axis simplistically denotes bla bla bla

Below the *Lower Bound* the responsibility falls onto humans. Above the *Upper Bound* it falls onto the machine. Inbetween is the responsibility gap. The rest of this section will consider possible responses to the responsibility gap and how to deal with it. What we will see is that the authors vary strongly in where they draw the upper and lower bound. For some they are distinct and form the responsibility gap; for some they coalesce and form one single threshold, with no space for the gap; for some these separating lines do not exist at all. Depending on that, the responses to the responsibility gap differ accordingly.

3.1 Where Human Responsibility Ends

The lower bound is the answer to the question: Can there be machines that humans cannot be morally responsible for?
Let us again consider Angela Smiths account of moral responsibility. She says that one is only responsible for actions that are the results of ones own evaluative judgements. We might imagine AI systems that have a certain kind of autonomy such that their actions do not refelct any humans evaluative judgements. For that, we only need to consider examples from science fiction media, like C-3PO from the Star Wars universe.

Andreas Matthias draws the line a at point that is less abstract. He says that

"[f]or a person to be *rightly* held responsible, that is, in accordance with our sense of justice, she must have *control* over her behaviour and the resulting consequences "in a suitable sense" (Fischer and Ravizza 1998:13)"[5]

[5] According to Matthias this quote can be found in:

Further, he says that machine learning technology will lead to a decrease of control humans will have over machines and their doings. People will not be able to predict what a machine does nor understand why it did it. The control requirement would not be fulfilled and no person could 'rightly' be held responsible.

This is where Matthias draws the lower bound: When neither the programmer nor the user can properly predict or explain, what the machine does.[6]

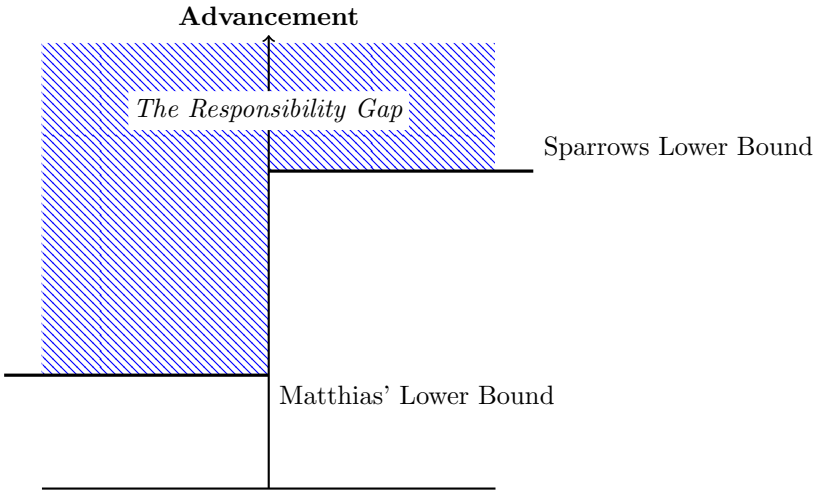
Sparrow considers how humans could be responsible and comes to the conclusion, that, if the system is truly autonomous, no human can be responsible for it. The programmer could be considered only responsible, if the mistakes the machine made came as a result of the programmer's negligence. However, Sparrow goes on, if the possibility of a mistake, say an autonomous weapon system kills the wrong target, a medical software misdiagnoses a patient, is clearly stated and disclosed as a "limitation of the system" (Sparrow 2007, p. 69), the programmer is released from bearing the responsibility and it would be taken over by the user, the commanding officer, the physician employing the technology. But they as well would not be appropriate targets of responsibility in Sparrow's view. He reasons that if the AI was truly autonomous, it would *choose* its actions on its own (Sparrow 2007, p. 70) and define its own ends (Sparrow 2007, p. 74). It seems to me that Sparrow suggests that an autonomous AI would exercise a will. And it would not be fair to hold the user responsible for something that originated from the will of such a machine.

I want to point out that Matthias and Sparrow seem to disagree about the degree of autonomy a machines would need to have, such that a gap in responsibility emerges. According to Matthias, the gap will exist pretty much any kind of machine learning technology, that has the capcity to learn and change its behaviour and be unpredictable for humans. He specifically mentions neural networks, genetic algorithms and reinforcement learning; technologies that already exist. Sparrow talks about systems that are much more sophisticated and further down the road. For him, the problem of the responsibility gap only arises with the kind of autonomy, that allows the machine to choose its own goals without being sentient. It is of course debatable, whether such a combination of properties can exist (Champagne and Tonkens 2015, p. 127). It might turn out that the autonomy of the relevant kind requires consciousness. But this issue goes beyond the scope of this thesis.

We can visualise the difference in Matthias' and Sparrows views in our neat little diagram:

J.M. Fischer and M.S.J. *Ravizza Responsibility and Control. A Theory of Moral Responsibility.* Cambridge University Press, Cambridge, 1998.

[6] Matthias does not really consider the machine to be an appropriate locus of responsibility. His essay gives no clue to where he would draw the upper bound.



PUT DIAGRAM HERE

"true" autonomy in a philosphically most rigorous sense (see p. 12).

'However, legal questions regarding how is responsible for the acrions of (ro)bot, and when it might cross the threshold to where it bears resposnibility for its own actions are certainly related to the themes of this book.' -Moral Machines p.191

3.2 Where Machine Responsibility Begins

If the argumentation above is convincing, there is a degree of autonomy that machines can achieve that which would render it unfair to hold the programmers and users responsible for it. It then makes sense to explore the third possible locus of responsibility, that Sparrow talks about: The machine *ominous music starts playing*

In the discussion about whether machines can be responsible, we find the sometimes explicit (N. Smith and Vickers 2021, p. 489), but mostly implicit assumption that a morally responsible agent is a full moral agent. They are the same thing. This is in line with Strawsons statement that having reactive attitudes towards someone is to see them as a member of the moral community (see page 9). Demonstrating the truth of that assumption goes beyond the scope of this thesis, but we shall accept it because of its intuitive validity. We will thus see that the search for how machines can be responsible will lead us occasionally to the question about how machines can be moral agents. So please do not be confused.

So, we find agreement among the authors that in order to be responsible, machines must be moral agents. However, we will encounter a lot of disagreement about what makes a machine a moral agent or whether they can achieve this status at all. In other words, the upper bound of the responsibility gap is placed at different heights by the different authors.

Who says that and why?

Sparrow, for example, acknowledges that a machine can be causally responsible for some action, but denies that this is enough to be morally responsible. To be morally responsible is to be open to moral appraisal (blame or praise) and as a result to be treated accordingly (punished or rewarded). Sparrow then proceeds imagining how punishing an autonomous machine might look like. He suggests that a sufficiently intelligent (and therefore autonomous) machine would probably have internal states akin to human internal states, that can be described as desires and needs. Punishemnt could theoretically be done by preventing these desires and needs from fulfillment: If they earn wages, machines could be fined; their liberty could be restricted by imprisoning them; or, in the most severe case, they could be destroyed as a form of capital punishment. However, Sparrow assumes that punishment is only punishment if the target suffers as a result. And that is a very demanding condition. It would not be enough, he says, that the machine would suffer 'functionally'. To fulfill our moral demands, the machine's suffering ought to have a phenomenological quality to it. Otherwise punishing an autonomous system would be no different from punishing a hammer[7]. It wouldn't *really* care. This means that, according to Sparrow, automonous machines could potetially be responsible, but only if they have the necessary phenomenology with the capability to suffer when punished and, consequently, to feel pleasure when rewarded. Sparrows argumentation seems to rely strongly on a property view of moral responsibility: To be morally responsible requires the phenomenological capability to suffer. When it comes to how we could establish whether the machine had such a capability or not, against my expectations, Sparrow says that all the machine had to do is to convince us that it had it and evoke appropriate responses within us humans through its behaviour. This approach heavily resembles a process view on phenomenological experience. In that case the machines would be "full 'moral persons'" with moral rights and duties and appropriate subjects for moral considerations, that are hitherto reserved for human beings. MAYBE I SHOULD TAKE OUT THIS LAST HALF SENTENCE (NO SOURCE) (NOT SELF-EVIDENT)

Smith and Vickers take a Strawsonian view on moral responsibility, instead. Additionally they stipulate that in order for machines to be morally responsible agents their moral system must "cohere with the system that we already have". This means that it is not feasable to invent a specific machine morality and marry it with the already existing human morality. That would lead to a state where ascriptions of responsibility are not understood by every member of the new moral community and, hence, would not find acceptance. Thus, all members of the moral community must have the same type of morality. Since we already have a human moral community[8], the question becomes: Under which circumstances can an AI become a member of the human moral community? Or in other words, which capacities must an entity have to be part of our moral

[7] Coeckelberg makes a similar comparison with a hammer.ZITIEREN

[8] I use the term human moral community in the loosest possible sense, to denote that, in general, most humans regard each other as moral beings with the capacity to take responsibility for their actions.

community?

Smith and Vickers argue that there are three core capacities that must be possessed by every "full member" of any moral community (on a Strawsonian account):

- The capacity to have reactive attitudes.
- The capacity to recognise other's reactive attitudes as demands for a certain type of treatment or regard.
- The capacity to respond to reactive attitudes. CITE

Thus, we can write on our little list of demands that an AI should have these capacities.

On top of that, each moral community forms moral traditions and practices that depend on the specific properties of its members. Based on these properties, we judge which expecatations, demands and reactive attitudes are fair to have and which not.

Example 4. We might say that it is morally wrong to drink and drive because of all the risks that it bears. But, we can easily imagine an animal akin to the homo sapiens with the only differnce that it naturally behaves as if drunk. With lower inhibition and motor control and worse memory forming capacities. All else being eqaul, it seems plausible that a society of these homo alcoholicus would form moral practices that account for this natural state of its members. (CITE HIERONYMI p.31-32)

We see that the moral system that we have developed is a contingent possibility based on reactive attitudes *and* the expectations that we have based on certain properties of the population. Moreover, these properties receive their relevance from their, somewhat, statistical normality and oridinariness (CITE STAWSON, HIERONYMI, SMITH/VICKERS).

Any responsible AI would need to have those statistically ordinary capacities. Smith and Vickers give only few examples of what these ordinary capacities are. However, they say two things: An AI that has these capacities would act "indistinguishably from us" *and* it would have a will. By the author's definition, an AI that behaves like humans and has a will is called a strong AI and a strong AI is a responsible AI.

Strong AI stands in contrast to weak AI, which is also indistinguishable from humans in its behaviour *but* has no will, no inner life. For Smith and Vickers, having a will is such a fundamental statistically ordinary capacity in the human moral community, that nothing that does not have a will cannot be reasonably considered a morally responsible agent.

Weak AI, however capable, is a mere object and expressing reactive attitudes towards it, would be no different than expressing them towards a chair. Smith and Vickers even go sofar to say that it would not be 'right'.

"There is something *wrong* with a person who genuinely blames a table when they bang their knee, or genuinely blames a baby who throws food. We might be irritated [...], but *blame* is misplaced."[9] (N. Smith and Vickers 2021, p. 4-5).

The authors, then, see no one who could reasonably be held responsible for such an AI's doings and we would face a responsibility gap.

We can observe that Smith and Vickers come to a similar conclusion as Sparrow: For an autonomous system to be morally responsible it must have an inner life, some sort of phenomenology, that legitimises it being a moral agent and not just a thing. If that is not given, the authors of both papers say that the responsibility gap persists and it is unclear who should have the responsibility for what the system does. They conclude that either the AI is responsible on the basis of having an inner life or nobody is an appropriate object of moral responsibility.

Notably, Sparrow and Smith/Vickers come to the same conclusion, by taking diametrically opposed stances on the concept of moral responsibility. It seems that Sparrow takes rather a property view on moral responsibility. He has a sense in which someone truly *is* responsible for their actions. Smith and Vickers, on the other hand, explicitly refer to the Strawsonian take on moral responsibility, which is a process view on the matter. However, their requirement for an inner life, in my opinion, does not cohere with the main point of Strawson's account. In having this requirement they tie moral responsibility to a metaphysical property that we cannot prove[10]. That is the very issue which motivated Strawson to develop his view in the first place. His whole idea revolves around the point that our moral practices work independent of any metaphysical properties like freedom of will and intentionality and consciousness. They instead rely on the interpersonal attitudes we have towards one another. Imposing the requirement of a will, undoes Strawsons whole work[11]

At this point we have considered two positions that both come to the same conclusion: Robots, AI, autonomous systems can be responsible, but only under the condition that they have some sort of phenomenology, will or inner life. As long as that is not given, we face a responsibility gap.

A question immedeatly comes to mind: Can we perhaps still imagine an AI that does not have the demanded qualities but can be considered responsible anyway?

John P. Sullins, for examplem, is not as demanding when it comes to moral responsibility. He writes that a robot can be responsible, if ascribing responsiblity to it is the best way to explain its behaviour. This is clearly a process view on responsibility. For Sullins assuming responsibility comes from a 'belief' of duty to do something. This belief must not originate from something we might call consciousness or a thought. This is probably the main difference between what Sullins thinks and what Sparrow and Smith/Vickers think. As he

[9] Italics taken from the original text
[10] Yet?
[11] In my humble opinion...

cynically remarks: "The machine may have no claim to consciousness, [...], or any of the other somewhat philosophically dubious entities we ascribe to human specialness"(Sullins 2006, p. 159).To Sullins, 'belief' is a functional term to describe something that motivates one to solve moral problems in a certain way (Sullins 2006, p. 159). PUT THIS IN A LATER SECTION ?: Machines that are not regarded as responsible agents can still be moral entities if they have sufficient autonomy and intentionality. Just like responsibility, these attributes can be merely apparent. Such machines should be targets of moral considerations. Sullins suspects that their moral status will be similar to the status that for example dogs have and their owners are the ones who are responsible for them(Sullins 2006, p. 159).
Here, Sullins does not leave room for a responsibility gap. He says that the user assumes the responsibility for an autonomous system until it fulfills the requirements to be responsible on its own. As I already said, these requirements are not very strict compared to Sparrow and Smith/Vickers. A machine's moral status is dependent on its role in society and how it is regarded by the society; further properties, such as consciousness, are secondary or even irrelevant.

We find two statements here, that are relevant for the topic at hand. Firstly, Sullins grants that machines can be responsible, if it makes sense to describe their behaviour in such a way. Secondly, he says that if machines are not responsible, the humans behind the machines are. There is no conceptual space for the responsibility gap. Still, there is a somewhat clear boundary for what makes an agent responsible.

Mar Coeckelbergh is even more radical than Sullins.

According to Coeckelberg, there is a mismatch between how we usually think about machines in general and how we *sometimes* act towards them (Coeckelbergh 2014, p. 61). We tend to think of machines as things and tools that do not have any kind of moral standing. To have a certain kind of moral standing an entity must have a certain property, or a set of certain properties. This is clearly a property view; Coeckelbergh calls it *the standard approach* (Coeckelbergh 2014, p. 62). These properties could be the usual suspects: Consciousness, the capacity to suffer, etc. (Coeckelbergh 2014, p. 62-63). And machines just do not seem to have these properties, that is why people are reluctant to give them the status of a full moral agent.
According to Coeckelbergh, there are several problems with this view. First of all there are epistemological issues. We have no way of objectively observing the relevant properties; there is also no objective way to bind a specific moral status to a specific property (Coeckelbergh 2014, p. 63). Furthermore, there is a cleft[12] between the just described way we think about machines in general and how we treat them. Coeckelbergh says that we sometimes ascribe emotions, personalities and "presence" to them (Coeckelbergh 2014, p. 62, 64). We may care about them and develop relationships with them. From the perspective of the standard approach this behaviour would be not "correct" or rational

[12]Coeckelberg actually calls it a "gap", but for obvious reasons I decided to alter his terminology.

(Coeckelbergh 2014, p. 64). These problems are sufficient for Coeckelbergh to question the standard approach.
Thus, he introduces *the relational approach* to how we assign moral status and, with it, responsibility. This account is a process view and it emphasises the relations the entity in question has with other entities. And "moral status [is] something that emerges through relations between entities" (Coeckelbergh 2014, p. 64). Moreover, the relation that matters most is the relation between the entity itself and the one who ascribes the moral status to it. Whether a particular machine is an appropriate object of moral considerations, whether it can hold responsibility is for the observer to decide. It is a highly subjective decision that strongly depends on the relation they have with the machine. This implies that moral status is not only not objective, but it is also not universal. On this view, there is no correct way to view an autonomous system and different people would also judge differently.
Note that Coeckelberghs account is much more permissive compared to everything else that we have encountered. Using the Strawsonian terminology: If the relation to an entity is such that one develops reactive attitudes towards it, one regards it as a responsible moral agent, regardless of what the entity in question is. Other authors pose specific requirements to establish which entities are 'appropriate' targets of reactive attitudes and which are not. Smith and Vickers argue that it is only appropriate, if the target has an inner life. Sullins says that it is legitimate to have reactive attitudes towards a machine, when its behavioural complexity matches the the human one and it is 'reasonable' to assign agancy to the machine in view of its behaviour (Sullins 2006, p. 169). All of these and other requirements fall away with Coeckelberghs relational approach. There is no such thing as an 'appropriate target' of reactive attitudes. This, of course, means that we do not stop at intelligent machines regarding the question of moral standing. Other types of entities can have moral standing.

Example 5. In early 2020, I visited the Wellcome Collection in London. One of the vitrines diplayed several masks, that were collected from different indigenous tribes from all over the world. One of the spots was empty. The guide explained that the tribe where that mask was taken from, has demanded it back. According to the tribes tradition, these masks were inhabited by the ghosts of deceased tribe members and they regarded and treated them as real people.

Under the standard approach[13], viewing the mask as a person does not really make sense. We might say that the tribes care for the mask comes from a believe rooted in tradition, religion and culture, but the mask is a mere thing without an objective moral status. Coeckelberghs relational approach on the other hand, allows for "multisubjectivity and plurality of truths" (Coeckelbergh 2014, p. 66). Coeckelbergh argues that this view treats morality not as an abstract philosophical concept, but accords with the reality of human everyday life and asks the question of moral standing on an individual basis (Coeckelbergh

[13]Caveats may apply: We are situated in the western culture, etc.

2014, p. 66). It respects how a particular entity is situated in personal, social, cultural, natural and other contexts.

This seems like a very extreme approach and demands a reevaluation of the way we conceptualise moral agency and responsibility. BLABLA BLA

Daniel Tigard takes a slightly different path. His account of moral responsibility leaves the question whether it is a process or a property view unanswered. However, Tigard subscribes to a pluralistic account of responsiblity (Daniel W Tigard 2021, p. 442-444). This means that there are different ways in which agents can be considered responsible. For example, agents are responsible for their characters in a different way than they are responsible for their evaluative judgements[14]. According to Tigard, such pluralistic approaches are very flexible and expandable (Daniel W Tigard 2021, p. 442-444). He proposes that as autonomous systems become more ubiquitous and the question of responsibility becomes more relevant, we might develop additional conceptions of what responsibility is that fit into this pluralistic view. He does not claim, that machines will be responsible in the very same sense as humans. Instead, there will be another conception of responsiblity, an artificial responsiblity that can only target artificial moral agents. In that, Tigard disagrees with Smith and Vickers about their claim, that machine responsiblity must be the same as human responsiblity. He also disagrees with Sparrows point that in order for a machine to be responsible it must be punishable - and for it to be punishable, it must have the capacity to suffer. This is a retributivist account of punishment (Sparrow 2007, p. 77, Daniel W Tigard 2021, p. 441). According to Tigard, there are alternative conceptions to punishment and they do not necessarily require the suffering part. Punishment can also be seen as a rehabilitation with the aim to achieve certain changes in the target and help it imporove (Daniel W Tigard 2021, p. 442). Now, it is possible to punish a machine in that sense, by restricting it in certain ways or changing its code or what ever one can come up with (Daniel W Tigard 2021, p. 443). Suddenly, Sparrows argumentation of why machines cannot be responsible falls away, because responsiblity, under this approach, does not rely on an inner life or similar things.
To summarise, Tigard proposes that there is no reason that machines must necessarily be responsible in the very same way as humans. Referring to a pluralistic approach of moral responsibility, he explores what it would mean to have sort of an artificial moral responsibility that could be applied to powerful AIs.

In this section, we have discussed what makes an autonomous system responsible. We have seen a gradual decline in what machines need to be in order to be responsible for their actions.
Sparrow and Smith/Vickers claim that in order for them to be responsible, AIs must have an inner life in the strongest possible sense. Otherwise all attributions of responsiblity would be misplaced.
Sullins' position is less rigid and proposes a rather functional account of respon-

[14]See (Shoemaker 2011) for further information

sibility. To him, machines can be regarded as responsible agents as soon as doing so is a sensible and natural way to explain their behaviour.
Coeckelbergh recognises that there is a mismatch between how we think about moral agency and responsibility and how we actually ascribe them. He further explains that there are epistemological problems with how we traditionally conceptualise the matter by looking at the properties of the entity in question. Instead, he explores the possibility that moral status is ascribed on the basis of the relation between the entity itself and the one who ascribes the moral status. In that sense, responsibility is something that is only ascibed by the observer and has no underlying universality to it.
Tigard relies on the pluralism of responsibility and says that we might expand our current notions by developing some sort of artificial responsibility.

It is difficult to properly sort these positions in the Advancement-diagram. If I were to guess, it would probably look something like that:

3.2.1 Why Machines Cannot Be Responsible

In a sense, the positions from the previous two sections argue that it is possible that machines outgrow the instrumentalist theory of technology. As a reminder, the instrumentalist theory supposes that computers and machines and all their derivatives are tools that are constructed with a human goal in mind and do not have their own ends (Gunkel 2020, p. 308). It follows that the responsibility for any action that originates in this technology should be ascribed to the relevant humans. The arguments from the last section invite us to abandon this theory by permitting the (at least conceptual) possibility of technology that is not a tool, but a moral agent with its own goals. Naturally, there are voices that hold on to the instrumentalist theory and doubt that machines can (or should) ever be responsible agents. Let us explore what they have to say.

all advocate for leaving behind the instrumentalist theory of technology. As a reminder, the instrumentalist theory supposes that computers and machines and all their derivatives are tools that are constructed with a human goal in mind and do not have their own ends (Gunkel 2020, p. 308). It follows that the responsibility for any action that originates in this technology should be ascribed to the relevant humans. The arguments from the last section invite us to abandon this theory by either permitting the (at least conceptual) possibility of technology that is not a tool, but a moral agent with its own goals - or denying that responsibility necessarily falls onto those who set the ultimate ends. Naturally, there are voices that hold on to the instrumentalist theory and doubt that machines can (or should) ever be responsible agents. Let us explore what they have to say.

One way to argue against machines being proper responsible agents is of course to point to a theory of moral responsibility and explain why they are not eligible for being considered as such. For example, we might remember Smiths account of moral responsibility (see page 9). Agents are responsible for those

and only those actions that are based on evaluative judgements. Can machines have evaluative judgements? SHOULD I PUT IN THIS PARAGRAPH?

Deborah Johnson is one of these voices (Johnson 2006). She argues that no man-made artifact can ever be a moral agent on its own. Her postion rests on, what she calls, the *standard account of moral agency and action* (Johnson 2006, p. 198). In order for an action to be open for moral evaluation, it must meet five conditions (Johnson 2006, p. 198):

1. The agent must have internal mental states (such as desires, beliefs, wishes) that motivate a behaviour. One of these states is an intending to act. Without that, there would not be an action. Johnson says that these internal states are the *reasons* for an action (Johnson 2006, p. 198).

2. Something happens. There is an external and embodied behaviour.

3. The behaviour is caused by the agents internal states as a rational means to accomplish their end.

4. The behaviour causes an outward effect.

5. The effect acts on a moral patient, an entity that is subject to moral considerations.

Only when an entity is capable of performing these kinds of actions, it can be considered a moral agent. We see that this is a property view on moral agency.

Johnson says that computers can easily fulfill the conditions 2,3,4 and 5. The second condition is met by computers being able to portray a behaviour by changing the screen and audio output or moving physical parts by controlling a motor. Since this behaviour is regulated by the machines internal states, the third condition can be checked off. That the behaviour can have effects on the outside world and act upon moral patients is obvious and these are exactly the conditions 4 and 5.
Condition 1 is trickier though. It states that the internal states of an agent are mental states and that one of these states is the intending to act. Johnson focuses on the second part and claims that computer systems will never have a true intending to act. The intending to act, so Johnsons reasoning, comes from an agents freedom. "Action is the exercise of freedom and freedom is what makes morality possible" (Johnson 2006, p. 198). Thus, if an entity is not free in that sense, it cannot act and is not an agent. Human behaviour has non-deterministic component to it, that "mysterious[ly]" makes the freedom (Johnson 2006, p. 198). One could now argue, that if computers have the ability to learn, their behaviour will also be non-deterministic. However, the way in which humans are undetermined is different for the way machines are; or at least it is unclear how to compare the two ways (Johnson 2006, p. 198). From this, Johnson concludes that Machines do not have the relevant type of freedom and consequently cannot develop intendings to act. Condition 1 in not met, computers cannot be moral agents.

Further, Johnsons says that machines have intentionality, but this intentionality is put into them by their designer and user. The machines intentionality can be captured by their functionality and it allows it to behave autonomously. Still, the functionality is determined by the designers and users goals.
This view of autonomous systems is loyal to Heideggers instrumentalist theory and the responsibility falls onto the humans as the agents that use these systems as tools. In short, machines are not moral agents and cannot be responsible for their actions because they do not have the necessary intendings to act. Their functionality is determined by the goals of those humans that design and use them. Therfore, they are responsible for the machines.

While I do believe that Johnson holds an important position, that is in line with a commonplace conception of technology, I find that she fails to convincingly demonstrate that machines do not have intedings to act. She claims that intendings to act arise from an agents freedom. Humans have freedom because their character is non-deterministic, that is why they can have intendings to act. She further says that machines can also behave in a non-deterministic way. But because of the differences in how humans and robots are composed, we can never *really* know, whether their non-determinism grants them the same freedom that humans are so blessed to enjoy. (may enjoy.)
It would then be a mistake to to claim that they can ever have the relevant type of freedom. Instead she asserts that computers can never have it (Johnson 2006, p. 203). Here we see the jump in Johnsons logic. She acknowledges an epistemological limitation and concludes that it implies the metaphysical difference.

Johnsons second point is a bit more convincing. In stating that the intentionality that autonomous systems have is always put into them by humans, the instrumentalist theory binds them together and makes the humans responsible. They are the ones who determine the goal, the use, the purpose of technology - not technology itself.

In short, Johnsons position denies the responsibility gap and asserts that humans will always have the responsibility for their machines. The reason is that it is an inherent property of all technology - no matter how advanced or autnomous - that its goals and functionality are defined by humans, not by the machines.

I am missing a comment about why this will also never be a case in the future, why there is no conceptual space for an autonomus system with its own intentionality and goals.

Joanna Bryson also subscribes to the instrumental theory. She says that "[we] determine their goals and behaviour [...] through specifying their intelligence [...]"(Bryson 2010, p. 3). This makes us humans responsible for them. She also provides a line of reasoning that might be taken as an argument against Sullins' and Coeckelberghs positions as they are described above. Bryson writes that there is a moral cost in accepting machines as agents, when they are, in fact, not (Bryson 2010, p. 2). The cost can be found on an individual level and on an institutional level. Accepting machines as responsible agents would nec-

essarily lead to their humanisation. They would be treated as peers to humans. People would befriend them and spend a lot of their "social capital" on them (Bryson 2010, p. 5)[15]. This means, that these people would have less capacity to socialise with other humans, their actual peers. Bryson suspects that people would actually prefer interactions with robots, because they would be easier and not as messy as true human-human interactions. She does concede though, that people who are lonely anyway would definetely benefit from these artificial interactions(Bryson 2010, p. 5).

Machines that behave responsibly and human like

Deborah Johnson says that machines *will* never have their own ends.

4 Filling the Gap

The last section was an exploration of the responsibility gap and its boundaries. We have found that, depending on ones conception of moral responsibility and agency, the problem of the gap presents itself in different lights. There are conceptions, where the gap does not even exist and the challenge disappears. For other conceptions though, the we see a gap, the conceptual space, that Sparrow talks about, where ascription of responsibility is problematic. How should we deal with this?

One solution is to cease the development of such autonomous robots entirely. Relating to AWS, Sparrow argues that in a just war, for every action there must be a responsible person for it (Sparrow 2007, p. 67). If that is not given, the war is unjust and immoral. Deploying AWS would result in situations where no one can justly be held morally responsible and that would be unethical (Sparrow 2007, p. 74). At first glance, it is reasonable to apply this view to other domains as well: If a machine could autonomously do something that has a moral risk attached to it and nobody was responsible, the situation would seem undesirable. Hence, such machines should not exist at all.

There are many people who disagree with Sparrow, in the sense that they claim that the responsibility gap can be filled[16]. Torben Swoboda thinks that the programmer should be responsible for the machine's conduct as part of their professional code (Swoboda 2017). This position makes sense, when looking at how AI systems are being developed. In particular, Swoboda refers to artificial neural networks (ANN). They are a specific type of machine learning technology that learns to transform input into a desired output to solve a predefined task. For that the programmer has to decide what kind of input the ANN will re-

[15] Bryson takes the term social capital from Putnam:
Putnam, Robert D. *Bowling alone: The collapse and revival of American community.* Simon and Schuster, 2000.

[16] It becomes increasingly difficult to proplerly map the different positions onto the metaphor the the *gap*. Does it make sense to speak of the gap being filled? Arguably, if it is possible to 'fill it', it probably does not exist. The more we use the metaphor, the more it wears out. However, it is not the metaphor that is important. What is important, is how we deal with the question of responsibility.

ceive. Every task can be represented in a multitude of ways and the programmer has to come up with a digital representation of the input that makes the most sense. Further, the programmer needs a data set of inputs with corresponding desired outputs. The ANN learns typically by receiving a massive amount of inputs; in the beginning of the learning phase the network will produce random outputs that do not make any sense, but throughout the process the produced outputs will be compared to the desired outputs. Then the parameters of the network will be tweaked and nudged to change how it produces its outputs until it performs well (meaning that it will generate a desired output for a sufficient portion of the given inputs). This whole procedure is under the programmer's control and the performance of the system depends strongly on the quality of the data set. Further, and this is probably the most important argument that Swoboda contributes, the system does not have to learn anymore. It only learns in the controlled training environment. Before it is applied in the real world, the system will be tested for performance and if it passes, it must not learn anymore. Moreover, if the system continues to learn, the performance might even change for the worse (Swoboda 2017, p. 307-309). While the AI might still produce specific outputs that are not predictable, it still has an overall behaviour that developed under the programmer's supervision and because the learning phase has been completed, that overall behaviour will not change. The overall behaviour of the system is thus under the programmer's control in the sense that they have to make sure that it acts in a desired way. One part of the overall behaviour is the accuracy at which the system performs its task. After the system has finished learning, the programmer assesses its performance on a second, unseen data set to get a hopefully reliable estimation on how accurate the system performs. This accuracy is given in percantages: What percentage of the input data has been mapped onto a desired output. For complicated tasks this accuracy does not reach 100 percent. So, there will be a small probability that the AI will make a mistake. The responsibility for that miskate, according to Swoboda, falls onto the programmer (Swoboda 2017, p. 310).

Swoboda's essay is postioned as a counter argument against Sparrow's concern about autonomous weapon systems and the problems of responsibility ascription. Looking at how contemporary AI systems are being developed and what makes the most sense from a programming point of view, he doubts that such systems will be continously learning after they are put into the field. Their learning is done before that and is supervised by the programmer. Because of that the programmer can assess the systems overall behaviour and intervene if it is not satisfying or even dangerous.

In my opinion, Swobodas take on the responsibility gap is a valid argument to at least reduce Matthias' responsibility gap. Both, Swoboda and Matthias refer to ANNs and other machine learning technologies that already exist today, so we may put the two positions against each other. Matthias is concerend with contemporary technology whose autonomy is of the simple type: The programmer and user cannot predict the outcome. The technology described by Swoboda is of the same type. However, Sparrow talks about a different kind of autonomy (see page 16). Swoboda himself says that the AI systems he taks about might

not be exactly what Sparrow means (Swoboda 2017, p. 309). It seems as if he is arguing that the machine of Sparrow's kind will not exist, because it does not make sense to develop it in this way. Sparrow, as we may remember says, that the such machines should not exist, because they give rise to the responsibility gap. We see that Swoboda actually does not contradict Sparrow. He only says that we should not worry about that kind of responsibility gap.

But what if we continue worrying about it? After all, Swoboda does not show beyond any doubt that we will never have the problem with the responsibility gap, that Sparrow describes, nor does it seem that the way he proposes to solve the problem of responsibility ascription apply for machines that can have their own goals.

Perhaps we can find refuge in another proposal made by Champagne and Tonkens (Champagne and Tonkens 2015). They recognise the specific type of autonomous system, for which Sparrow could not resolve the problem of responsibility. "[T]he robot must be sophisticated enough to make its own choices - but not so sophisticated that it can experience pain and pleasure" (Champagne and Tonkens 2015, p. 128). However, the authors say that Sparrow overlooks a possibility to ascribe responsibility to someone: "Blank check responsibility" (Champagne and Tonkens 2015, p. 132). They argue that a suitable individual may willfully and priorly take the responsibility for whatever an autonomous AI does. In a sense that person issues a blank check and whenever the AI does something of moral relevance, people can 'cash' that check and blame (or praise) that person. The interesting question is: Who can be a suitable individual? Champagne and Tonkens base their argumentation on the example of autonomous warfare; the AI is an autonomous weapon system, that is autonomous in the sense that Sparrow means. The suitable individual, then, could be a military leader who decides whether to deploy this autonomous technology or not. Accepting the responsibility would be a "nonnegotiable" part of their job (Champagne and Tonkens 2015, p. 132). When deciding about using AWS, they would need to consider the risks and opportunities for themselves as the bearer of responsibility for any consequence that come from that decision.
Champagne and Tonkens assert that this idea resolves the problem with responsibility ascription, since it would now be possible to point to a specific human that takes the responsibility (Champagne and Tonkens 2015, p. 132). Their contribution to the debate is that there must not necessarily by a "transitive chain" that connects an action to the responsible person. This allows for the concept of blank check responsibility, where an individual takes the responsibility for an action that did not originate from their own will. Here, it also seems fair to transfer the basic idea (with some adjustments) to domains other than autonomous warfare.

One thing that can be observed in Champagne and Tonkens idea, is that they seem to view the decision to issue the blank responsibility check separate from the decision to deploy such autonomous systems. The former is an "ex-

plicit social contract", that one can decide not to agree to (Champagne and Tonkens 2015, p. 132-133, 134). This leaves some important questions unanswered. What happens, for example, if someone decides not to explicitly agree to such a contract but would still use an autonomous system? Could a random person step forward and accept the responsibility, while others decide about the use of such machines? Champagne and Tonkens refer to the military leader who should take the responsibility, but this referal is not substantiated by any pre-existing relation that they have to the machine. On Champagne and Tonkens view, it is not important, that they are the ones who have the control about using the technology. They say that "two things can be related just in virtue of their being related" (Champagne and Tonkens 2015, p. 127). While this might be true, it seems unnatural and unnecessary to base the concept of blank check responsibility on a relation that exists only because it implies itself. It requires a practical infrastructure on top of morality[17] that produces the conditions for ascribing responsibility and it seems too easy to bypass that infrastructure. If we accept the promoted idea as a viable solution to the responsibility gap, it seems only natural to also accept a solution where the decision to deploy autonomous machines is inherently connected to taking responsibility for their actions. The person who decides to use autonomous machines implicitly takes the responsibility. In my opinion, this would be a much simpler solution that does not raise as many questions. The relation, though perhaps indirect, already exists, so it is not necessary to invent another relation as Champagne and Tonkens do.

Specifically, they mean that a suitable individual

[17]Such as explicitly and willingly entering a social contract.

For these kinds of systems, they agree that there is a respo

They introduce the concept of "blank check responsibility"

not strong enough to solve Sparrow's responsibility gap. It is strong enough though to at least reduce Matthias' responsibility gap to the one Sparrow talks about. Swoboda and Matthias refer to ANNs and other machine learning technologies that already exist today, so we may put there positions against each other. Sparrow on the other hand talks of technology that is more sophisticated. For him the problem Swoboda himself says that the AI systems he taks about might not be exactly what Sparrow means (Swoboda 2017, p. 309).

Sparrow himself thinks that the responsibility gap should be avoided.
A convincing and realistic proposal is made by

5 Real World Problems

5.1 Autonomous Weapon Systems

5.2 Healthcare

5.3 COMPAS

6 Discussion

I don't speak of the technical side. This is something that is actually really important, because how the robot actually is, how it is programmed can determine its moral capacity and how we view it. (Johnson: Technology without any human responsibiltiy)

We do not pretend to be able to predict the future of AI. Nevertheless, the more optimistic scenarios are, to our skeptical minds, based on assumptions that border on blind faith. It is far from clear which platforms will be the most successful for building advanced forms of AI. Different platforms will pose different challenges, and different remedies for those challenges. (Ro)bots with emotions, for example, represent a totally different species from (ro)bots without emotions. - Moral Machines p.194

War robots are bad? We want the moral cost and risc be very high.

Hypothesis: We could say that machines are capable of being moral and responsible, if they, if left alone, would develop some sort of morality of their own. This seems to be an empirical question. I find this very compelling

The authors talk about responsibility for actions and attitudes but there is very little talk about responisbility for consequences, which seems even more important in the context of AI.

In some cases we need to decide what our moral obligation is. Is it more moral to create a system, where no one is really morally responsible, but there are much less bad outcomes because machines perform better than humans (e.g. autonomous vehicles) or is it so important that we can find moral responsibility in such cases that we cannot turn to such a system

Strawson talks about a resource that we all have: We can regard someone with objective attitudes who we usually regard with a reactive attitude. What if, then, we have another resource? What if we have the resource to regard somthings with a reactive attitude that we would usually regard with an objective attitude? Cite: FREEDOM AND RESENTMENT PAGE 10. ALSO IN HIERONYMI

7 Conclusion

CITE COECKELBERGH ENDE 65-66 about philosophy should not try to close the discussion about moral standing but instead keep it open. It is so good as a finishing thought!

8 Meditation

WHAT IS THIS SECTION ABOUT? MY THOUGHT ON MY JOURNEY.

I CHANGED MY VIEW ON MORAL RESPONSIBILITY IT WAS DIFFICULT FOR ME TO AVOID RABBIT HOLES

This topics is insanely chaotic without structure and perbe without even sufficient legitimacy for the question itself(?).

How to differenciate between normative and descriptive approaches and problems?

My ideas for future research: Would it be a good measure to determine a machines moral standing by isolating a society of these machines and observe if they develop something like a morality themselves?

Are some of the proposed approaches against the ideas of the enlightenment?

Talk about the resource but the other way around.

9 Disclaimers

Put this in the beginning or even better into the introduction

Blame and praise are asymmetrical in how we pay attentio to them . While it might be an interesting intellectual execise to think about who deserves credit for a piece of art produced by a machine learning algorithm the question of responsiblity seems far more pressing for when an automated car runs over a pedestrian or a medical software misdiagnoses a patient. I will thus, mostly restrict my search for responsibility to cases in which we want to blame. REFERENCE: SMITH BEING RESPONSIBLE VS HOLDING RESPIONSIBLE P.5

10 Acknowledgements

Acronyms

AI Artificial Intelligence. 2

AWS Autonomous Weapon System. 2

CR Control Requirement. 1

ML Machine Learning. 2

mRINFES Moral Responsibility is Important and Necessary for a Functional and Ethical Society. 3

References

Bryson, Joanna J (2010). "Robots should be slaves". In: *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* 8, pp. 63–74.

Champagne, Marc and Ryan Tonkens (2015). "Bridging the responsibility gap in automated warfare". In: *Philosophy & Technology* 28.1, pp. 125–137.

Coeckelbergh, Mark (2014). "The moral standing of machines: Towards a relational and non-Cartesian moral hermeneutics". In: *Philosophy & Technology* 27.1, pp. 61–77.

Gunkel, David J (2020). "Mind the gap: responsible robotics and the problem of responsibility". In: *Ethics and Information Technology* 22.4, pp. 307–320.

Heidegger, Martin (1977). *The Question Concerning Technology and Other Essays*. Translated by William Lovitt. Garland Publishing, Inc.

Johnson, Deborah G (2006). "Computer systems: Moral entities but not moral agents". In: *Ethics and information technology* 8.4, pp. 195–204.

Matthias, Andreas (2004). "The responsibility gap: Ascribing responsibility for the actions of learning automata". In: *Ethics and Information Technology* 6.3, pp. 175–183. DOI: 10.1007/s10676-004-3422-1. URL: https://doi.org/10.1007%2Fs10676-004-3422-1.

Misselhorn, Catrin (2018). *Grundfragen der Maschinenethik: Reclams Universal-Bibliothek*. Reclam Verlag.

Shoemaker, David (Apr. 2011). "Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility". In: *Ethics* 121.3, pp. 602–632. DOI: 10.1086/659003. URL: https://doi.org/10.1086%2F659003.

Smith, Angela M (2005). "Responsibility for attitudes: Activity and passivity in mental life". In: *Ethics* 115.2, pp. 236–271.

— (Jan. 2007). "On Being Responsible and Holding Responsible". In: *The Journal of Ethics* 11.4, pp. 465–484. DOI: 10.1007/s10892-005-7989-5. URL: https://doi.org/10.1007%2Fs10892-005-7989-5.

Smith, Angela M (2008). "Control, responsibility, and moral assessment". In: *Philosophical Studies* 138.3, pp. 367–392.

— (2012). "Attributability, answerability, and accountability: In defense of a unified account". In: *Ethics* 122.3, pp. 575–589.

— (2015). "Responsibility as answerability". In: *Inquiry* 58.2, pp. 99–126.

Smith, Nicholas and Darby Vickers (Apr. 2021). "Statistically responsible artificial intelligences". In: *Ethics and Information Technology*. DOI: 10.1007/s10676-021-09591-1. URL: https://doi.org/10.1007%2Fs10676-021-09591-1.

Sparrow, Robert (2007). "Killer robots". In: *Journal of applied philosophy* 24.1, pp. 62–77.

Strawson, Peter (1962). "Freedom and Resentment". In: *Proceedings of the British Academy, Volume 48*, pp. 1–25.

Sullins, John P (2006). "When is a robot a moral agent". In: *Machine ethics* 6.2006, pp. 23–30.

Swoboda, Torben (2017). "Autonomous Weapon Systems-An Alleged Responsibility Gap". In: *3rd Conference on" Philosophy and Theory of Artificial Intelligence*. Springer, pp. 302–313.

Tigard, Daniel W (2021). "Artificial moral responsibility: How we can and cannot hold machines responsible". In: *Cambridge Quarterly of Healthcare Ethics* 30.3, pp. 435–447.

— (July 2020). "There Is No Techno-Responsibility Gap". In: *Philosophy & Technology*. DOI: 10.1007/s13347-020-00414-7. URL: https://doi.org/10.1007%2Fs13347-020-00414-7.