Responsibility Gap

Mischa

November 19, 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 it's own decisions based on it's prior experience without the programmer or user knowing what the decision is going to be.

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

This part is very dirty. I don't think I will put it here, because it is too confusing for now. I will see, it it makes sense to include it somewhere else.

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 it's 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 DID YOU EXPLAIN IT MISCHA???? , 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 rigor-

ous account of the term. The necessity of clarifiying the term becomes clear, when we expose it's 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 certain action, attitude or event can be attributed to someone in a sense that allows us to appraise them on it's 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 responsibility. 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.

NOTE: Criticise MATTHIAS FOR HIS CR. IT is not well defined. What is control. Control over future choices, past choices, what about situation where one does not have all the situation. Where does control end and so on.

2.2.1 Strawson

There is an ongoing discussion in philosophy about the existance of determinism and it's impact on free will, and some philosophers deem free will to be closely related to moral responsibility [zitieren]. To discuss this topic is notbla bla nor is it my intention to take a position on this issue. Instead I will try to elegantly sidestep the matter by taking a Strawsonian approach on moral responsibility.

PERHAPS I CAN TAKE THAT FIRST PARAGRAPH OUT? NO NEED TO SPEAK OF DETERMINISM.

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.

BEISPIEL EINFÜGEN!!

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) THE 'TO BE CLEAR' IS NOT AS CLEAR AS IT SHOULD BE.

Strawson proceeds and gives these vicarious reactive attitudes the qualifier *'moral'* and the objects of such reactive attitudes are said to have done something 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 is not 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" (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.

HERE I MUST FINISH THE DIGRESSION AND EXPLAIN REFERENCE THE INITIAL QUESTION: WHAT ARE WE RESPONSIBLE FOR?

FIND MORE INTERPRETATIONS AND PUT THEM INTO CONTEXT

SHOEMAKER SAYS: MORAL RESPONSIBILITY IS, AT LEAST IN PART, BEING OPEN TO A CERATAIN RANGE OF MORAL RESPONSES (PAGE 11)

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 [...]" (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.

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

STRAWSONS ACCOUNT DISMISSES THE METAPHYSICAL VALUE OF MR. INSTEAD IT IS SUGGESTED THAT MR HAS ONLY SOCIAL AND PSYCHOLOGICAL SIGNIFICANCE

UNDER STRAWSONS ACCOUNT THERE IS NO SUCH THING AS MORAL RESPONSIBILITY. THERE IS ONLY THE SOCIAL NEGOTIATION OF INTERPERSONAL ATTITUDES IN A SOMEWHAT DIALECTIC MANNER. THERE IS ONLY CAUSAL RESPONSIBLITY AND THE REACTIVE ATTITUDES ATTATCHED TO IT. (HOLDING RESPONSIBLE) AT THIS POINT IT IS A QUESTION THAT WE SHOULD ASK IN A MORE SOCIOLOGICAL AND PSYCHOLOGICAL SETTING RATHER THAN A PHILOSOPHICAL

ONE.

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 *real* 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).

TAKE OUT THIS NEXT PARAGRAPH. IT IS TOO SPECULATIVE. WHAT ABOUT THE WHOLE THING FROM BEFORE? i MIGHT CHANGE THE WHOLE SECTION, TO BE MORE GENERAL ABOUT PROPERTY VIEW RESPONSIBILITY. THEN I CAN INCLUDE SHOEMAKER AND SCANLON AND NOT GO IN SO MUCH DETAIL. I believe, Strawson would answer to this, that his account of our moral practices does not claim to be fair or fulfill all the demands we might have towards said practices. Nor is it necessarily internally consistent. In this sense, it is not normative but largely descriptive. In practice, when blaming someone, we do not distinguish between our percetion of their attitude and their *real* one. I suspect, that even if Strawson would acknowledge the differnce between being held responsible and being responsible, he would say that the latter notion has no practical relevance in our moral practices.

We can take Strawsons account of responsibility as a description for when people are *held* responsible in Smiths sense. 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

[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

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 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

interpretation is correct the word "reason-giving" as it is used by Shoemaker is a bit more fitting (Shoemaker 2011, p. 23).

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.

HERE I START MAKING JUMPS IN THE TEXT. THERE ARE A LOT OF WHOLES THAT NEED TO BE FILLED

2.3 What are the machines we will be talking about?

In my writing I will use such phrasings as artificial intelligence (AI), autonomous system, machine and robot almost interchangeably. It is thus useful to define what I mean when I use these terms and, looking at the huge actual and possible variety that these technologies occur in, to find a way to classify them. So this will be our working definition that unifies all of them:

A human artifact that is capable of making autonomous decisions and act upon them.

I need somebody to make a connection between moral responsibility and (moral) agency PLEASE! AHHHHH

While I will not go into any specific implementational or engineering details about the connected technologies, we shall try to understand what that means in the context of the topic on a conceptual philosophical level. The notion of autonomy is emphasised and it implies a certain kind of agency. Catrin Misselhorn gives a gradual and multidimensional definition of agency. Introduce autonomous power by Hellström. And speak of autonomy in the context of Allen and Wallach. Give examples of systems whose actions have moral implcications.

Allen and Wallach introduce a gradual understanding of autonomy. There are things with low autonomy and things with high autonomy, and there is a full spectrum of autonomy in between.

Catrin Misselhorn Agency:

When can a certain behaviour be considered an action? What is an agent? There are natural and artificial entities that are considered agents. They cover a broad spectrum of behaviour. So the notion of agency must be gradual and multidimensional. Agency is an interplay of many factors. The two most important factors (according to Misselhorn) are rationality and autonomy (the ability to initiate one's own behaviour).

AI, Robots, Autonomous systems and machines will be used interchangably. Introduce the scale of allan and wallach morality-autonomy. Introduce strong ai, weak ai. AGI

3 Can we Bridge the Gap?

The story sofar:
We have defined moral responsibility to be an openess 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 based on a social negotiation, 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 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", 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. This seems to me like a fair account even beyond autonomous robot warfare. It seems that the three targets (programmer, user, machine) are intuitively where to look for responsibility and we will mainly focus on them.

Before moving on, I shall present a brief argument, that will help us structure the assessment of the question:

The debate around the responsibility gap is centered around the argument that when an autonomous AI does something, neither the programmer nor the user can reasonably be considered morally responsible. If then, the AI cannot be considered responsible as well, we are faced with a sitiuation where something has happened and no one can be considered responsible. I would like to now go back and lay out the arguments of those who say that the problem exists in the first place; those who say that we will not be able to deal with the problem of responsibility ascription in the same way we are used to by subscribing to the above mentioned instrumental theory; those who say that the involved humabns are not responsible and nor is the machine.

MY ARGUMENTATION MUST BE CLEAR HERE. DO PEOPLE UNDERSTAND WHAT I AM GOING TO? WHAT AM I GOING TO IN THE ABOVE PARAGRAPH: THE QUESTION ABOUT THE QUESTION ITSELF. WHAT MOTVIATES US TO THINK OF THE RG: HUMANS ARE NOT RE-

SPONSIBLE IN THE SAME WAY THEY USED TO BE. LET US EXAMINE WHO SAYS THAT.

3.1 Why humans cannot be responsible

Let us reconsider Matthias' argument for the responsibility gap. Matthias 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"".

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.

Matthias does not really consider the machine to be an appropriate locus of responsibility. I assume that the machines he talks about are just that: machines, tools, systems to be used by humans. They cannot have responsibility.

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", 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* it's actions on it's own (Sparrow 2007, p. 70). 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.

'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

In taking the claim of

IF THE GAP PERSISTS, IT IS ARGUED THAT WE SHOULD CEASE THE DEVELOPMENT OF SUCH AUTONOMOUS ROBOTS (SPARROW/Robots should be slaves etc.). THERE IS ALSO AN ARGUMENT TO BE MADE THAT WE AS A SOCIETY TAKE THE RISK OF HAVING THEM BECAUSE OF THE BENEFITS THEY PRESENT (Hellström)

I think that, what irritates us, what motivates us to speak of a responsibility gap, is the introduction of autonomous machines as

I shall structure my examination of the three

QUESTIONS: 1) WHAT IS THE AUTHORS MAIN POINT? 2) WHAT KIND OF AUTONOMOUS SYSTEM ARE THEY TALKING ABOUT? 3) WHAT IS THEIR VIEW ON MORAL RESPONSIBILITY? 4) WHO SUPPORTS THEIR ARGUMENT? 5) WHO DISAGREES WITH THEIR ARGUMENT?

3.2 Can machines be responsible?

Arguably, if machines cannot be responsible, we do not need to divert from instrumentalism and deal with the stuff just like we have dealt with it in the past.

The question of can machines be responsible can be answered in four different ways: Yes, No, Maybe and Kinda

3.2.1 Machines cannot be responsible

Who says that and why?

Sparrow: Machines, how ever autonomous they are, cannot suffer, thus they cannot properly be held responsible. Sparrow talks in his essay 'Killer Robots' about Autonomous Weapon Systems, the points he makes about moral responsibility are not specific to that particular type of domain and are, thus, generalisable. Sparrow 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, says he, 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. 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 it's 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)

I THINK I SHOULD PUT THE PAPER ABOUT STATISTICALLY RESPONSIBLE AI HERE; THEY COME TO THE SAME CONCLUSION; BUT THEIR VIEWS ON MR AND CONSCIOUSNESS ARE DIAMETRICALLY OPPOSED. WHICH IS FUNNY.

PUT CONCLUSION ABOUT THE TWO PAPERS AT THE END OF BOTH OF THEIR DISCUSSION!

Smith/Vickers: Machines can only be responsible, if they are conscious and their morality is similar to ours

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[3], 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 community?

I THINKT THIS ABOVE IS A GOOD PARAGRAPH

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 it's members. Based on these properties, we judge which expecatations, demands and reactive attitudes are fair to have and which not.

Example 1. 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 it's 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

[3]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.

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 it's 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."[4] (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. That is the issue which motivated Strawson 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

[4]Italics taken from the original text

interpersonal attitudes we have towards one another. Imposing the requirement of a will, undoes Strawsons whole work[5]

END?

[5]In my humble opinion...

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 or inner life. As long as that is not given, we face a responsibility gap.

Two questions come to mind. First: Can we perhaps still imagine an AI that does not have the demanded qualities but can still be considered responsible. Second: Is it true that we face a responsibility gap, if the machine cannot be responsible? What about the other involved players?

PLACE TIGARD HERE AND SULLINS AND COECKELBERG

Johnson: Machines cannot be full moral agents bla bla, I am not sure what she says Bryson: Robots should be slaves. Otherwise we will use them to evade responsibility Marino/Tamburrini: Implicit neglection of the option Champagne/Tokens: Implicit neglection of the option

3.2.2 Machines can be responsible

Discuss artificial agency **Tigard: different types of responsibility/SmithVickers would probably disagree, because the machines would have a different morality than humans and that is not morality Coeckelberg: It all depends on what position in society they will have.**

Sullins: Robots can be functionally responsible. If the only way to make sense of a robots behaviour is to ascribe some responsibility to it, then it is responsible.

Sullins approach is a bit different insofar as he says that being able to be responsible is a precondition for the machine to be an autonomous agent, not the other way around. He writes that a robot can be responsible, if ascribing responsiblity to it is the best way to explain it's behaviour. 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. As he cynically remarks: "The machine may have no claim to consciousness, [...], or any of the other somewhat philosophically dubious entities we ascribe to human specialness"(**sullins2006robots**). 'Belief' is a functional term to describe something that motivates one to solve moral problems in a certain way (**sullins2006robots**). WRITE ABOUT SMITH ANS VICKERS OBJECTION

4 Real World Problems

4.1 Autonomous Weapon Systems

4.2 Healthcare

4.3 COMPAS

5 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 it's 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

6 Conclusion

7 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

8 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

[Hei77] Martin Heidegger. *The Question Concerning Technology and Other Essays*. Translated by William Lovitt. Garland Publishing, Inc., 1977.

[Mat04] Andreas Matthias. "The responsibility gap: Ascribing responsibility for the actions of learning automata". In: *Ethics and Information Technology* 6.3 (2004), pp. 175–183. DOI: 10.1007/s10676-004-3422-1. URL: https://doi.org/10.1007%2Fs10676-004-3422-1.

[Sho11] David Shoemaker. "Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility". In: *Ethics* 121.3 (Apr. 2011), pp. 602–632. DOI: 10.1086/659003. URL: https://doi.org/10.1086%2F659003.

[Smi05] Angela M Smith. "Responsibility for attitudes: Activity and passivity in mental life". In: *Ethics* 115.2 (2005), pp. 236–271.

[Smi07] Angela M Smith. "On Being Responsible and Holding Responsible". In: *The Journal of Ethics* 11.4 (Jan. 2007), pp. 465–484. DOI: 10.1007/s10892-005-7989-5. URL: https://doi.org/10.1007%2Fs10892-005-7989-5.

[Smi08] Angela M Smith. "Control, responsibility, and moral assessment". In: *Philosophical Studies* 138.3 (2008), pp. 367–392.

[Smi12] Angela M Smith. "Attributability, answerability, and accountability: In defense of a unified account". In: *Ethics* 122.3 (2012), pp. 575–589.

[Smi15] Angela M Smith. "Responsibility as answerability". In: *Inquiry* 58.2 (2015), pp. 99–126.

[Spa07] Robert Sparrow. "Killer robots". In: *Journal of applied philosophy* 24.1 (2007), pp. 62–77.

[Str62] Peter Strawson. "Freedom and Resentment". In: *Proceedings of the British Academy, Volume 48*. 1962, pp. 1–25.

[Sul06] John P Sullins. "When is a robot a moral agent". In: *Machine ethics* 6.2006 (2006), pp. 23–30.

[SV21] Nicholas Smith and Darby Vickers. "Statistically responsible artificial intelligences". In: *Ethics and Information Technology* (Apr. 2021). DOI: 10.1007/s10676-021-09591-1. URL: https://doi.org/10.1007%2Fs10676-021-09591-1.

[Tig20] Daniel W. Tigard. "There Is No Techno-Responsibility Gap". In: *Philosophy & Technology* (July 2020). DOI: 10.1007/s13347-020-00414-7. URL: https://doi.org/10.1007%2Fs13347-020-00414-7.