Responsibility Fap

Mischa

August 21, 2021

Contents

1 Introduction

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. Iphone in 2007 and a couple years later everybody has a smartphone bla bla bli blub

Traffic lights are installed and programmed, perhaps superficially supervised and occasionally maintained and updated. peepee

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 not (in the case of the dam-architect). 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* (**heidegger1977technology**).

"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."(**heidegger1977technology**)

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.

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 the decisions and even without any human being able to explain the 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, even 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.

1.1 The Dirty Problem

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 implies 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**)

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 refered 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 complements the intrumental theory nicely: 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**)

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

Current practices in Artificial Intelligence (AI) aim at increasing the autonomy of machines and their software [zitieren]. '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. In other words, they reduce the amount of control the programmer or user has over the machines. This is primarily the case for technology that uses machine learning to acquire a certain behaviour. [Passt hier eine beschreibung von machine learning rein?] At the same time the actions of such technologies have more and more impact on the people surrounding them [zitieren].

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

[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

2 What is Moral Responsibility

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 moral responsibilty. 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. PUT SOMETHING HERE. ANGELA SMITH DESCRIBES THE DIFFERENT MEANINGS OF "A HOLDS B RESPONSIBLE FOR X"

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.

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 [...]"(**Strawson1962**). 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's attitudes towards us (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 (**Strawson1962**).

I should also point out, just like Strawson repeatedly does (**Strawson1962**), 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"(**Strawson1962**) 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 (**Strawson1962**). 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 [...]" (**Strawson1962**).

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 (**SmithVickers2021**) (**Matthias˙2004**). Tigard takes moral responsibility in this regard "as a social funtion of [...] reactive attitudes" (**Tigard˙2020**). 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 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 (**Strawson1962**). Tigard descibes these cases as situations "where the agent is normal, but the circumstances are abnormal [...]" (**Tigard˙2020**).

   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**). Strawson speaks of children or schizophrenics or people that act out of compulsion. 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 (**Strawson1962**). 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 [...]" (**Strawson1962**). Seeing an agent as such, implies that we portray a second set of attitudes towards them. Strawson calls these attitudes *objective attitudes* (**Strawson1962**).

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 (**Strawson1962**). 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

2.2 Shoemaker

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 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? And it would not be fair for us to hold someone responsible, if they 'in reality' are not responsible (**Smith˙2007**).
*Oh hey what is that? I think that is determinism creeping in the background! Oh no, I hope nobody will notice it! *MISCHA USES HIS SPECIAL ATTACK:* ***GLOSSING OVER STUFF THAT MIGHT BE VERY IMPORTANT****

Can you feel it's presence? It is still at distance, yet comes closer the more we seek for truth. It regards us patiently as it is the end boss and it waits for a fight that may never halt. And the fight might never halt; not because we are equal adversaries, but because we are too weak to know when it is over.

It's breath is cold and merciless. It knows. But as the great minds, whose work is the basis of the words you read, I shall continue to ignore this final question and move on. And I ask the same of you. With the words of the russian writer Mikhail Bulgakov: Follow me, reader!

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 it is not even 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. THIS IS SOMETHING THAT I WANT TO SAY, BUT BETTER: Being a descriptive account of our moral practices, we can take it as it is standing alone, but it is also capable of leaving enough space to place a normative theory inside of it to complement and it. Shoemakers account of responsibility is such a theory, that can be placed into strawsons account.

Angela Smith mentions the difference between being held responsible and being actually responsible

Still, we find other philosophers seeking for a more rigorous and satisfying account of 'appropriate' in this context. We have encountered one such approach in Matthias' control requirement (**Matthias˙2004**), which we will dicuss in a later section. @FUTURE_MISCHA: HAVE YOU REALLY DISCUSSED IT? DO YOU HAVE ENOUGH MATERIAL TO DISCUSS IT? I also want to introduce an additional approach, proposed by David Shoemaker (**Shoemaker˙2011**). DA SHOE tries to give an account of what it means to be truly responsible. He proposes three distinct types of responsibility: Answerability, Attributability and Accountability.

Fro mthis it is easy to interpret that moral responsibility rests solely on other's impression of one's will (though this is not a necessary conclusion or even something that Strawson argues for). In this regard, we might want to say that this seems too one-sided. If Strawsons account depends so heavily on attitudes, it shall not neglect the blamee's 'true' attitudes.

based on our perception of their attitude, if thi

According to Strawson, Moral Responsibility must not be a metaphysical entity, but rather manifests as a result of human nature and our social practices.

With this in mind, instead of asking 'What is moral responsibility?', the better (and certainly easier) question to ask is: What does it mean to be morally responsible?

In "Freedom and Resentment" P.F.Strawson explains that reactive attitudes. Bla bla We expect a certain behaviour from other people and depending on wether they cohere with these expectations we exhibit resentment or gratitude towards their behaviour[Mehr ins detail gehen]. Some philosophers say that responsibility is the property that allows us to appropiately target an agent with gratitude or resentment for something they have done [zitieren/umschreiben das sind nur dreckige sätze]. Bla bla.

In light of Strawsons refusal to see moral responsibility as a metaphysical entity, 'What is moral responsibility?' is perhaps the wrong question to ask. The better (and certainly easier) question is: What does it mean to be morally responsible (for something)?

So what are the cases, in which we appropriately say that an agent is responsible. Explain CR again. Explain when an agent is excused and when they are exempted from being held responsible.

Extend the model of responsibility to shoemakers 'accountability, answerability and explainability model'

When we talk about moral responsibility, we must probably first explore

what we mean by that term. Specifically we need to answer two central questions:

1. In which cases can somebody be held responsibly?
2. What does moral responsibility entail?

For the sake of a focused and productive argumentation I will, for the duration of this entire work, assume that the concept of moral responisibility is important and is necessary for a functioning and ethical society (mRINFES) without providing an argument for this assumption. Questioning this assumption would, I believe, fill a whole other bachelor's thesis and likely even more. In the sense of this assumption, I will also ignore the debate around free will and how it is connected to moral responisbility.

The Control Requirement More complex models of responsibility

Moral Agency

3 Can we Bridge the Gap

Who are the candidates?: The manufacturer, the user, the machine If the machine is responsible does it imply moral agency/ we must develop reactive attitudes. -¿ What are the conditions for developing reactive attitudes towars machines (Statistically responsible AI Vickers and Smith) Shoemaker has another example about the aliens, that might eb fitting here

Essentially: How do machines fit into these frameworks Upper bound - lower bound of moral agents

Yes we can: Here is how Instrumentalism 2.0 There is a moral risk in using unpredictable machines and the users/manufacturers that use them accept this risk and are (implicitly) accepting the responsibility. Analogy: There is a risk in using medical drugs because of the side effects. Machine Ethics Hybrid responisibility

No, we cant: Here is why:

4 Real World Problems

4.1 Autonomous Weapon Systems

4.2 Healthcare

4.3 COMPAS

5 Discussion

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