Working title

Mischa

July 21, 2021

Contents

1 Introduction

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* (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.

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, 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 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, 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

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.

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.

There is an ongoing discussion in philosophy about the existance of determinism and it's impact on free will, which in turn appears 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 argumentations lies "the very great importance that we attach to the attitudes and intetions towards us of other human beings

[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

[...]"(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 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. The same action with different attitudes elicits different reactions from us. Strawson's example is about 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 at this point, just like Strawson repeatedly does in his paper(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 an complicated 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 the behaviour or will. To be clear: These impersonal reactive attitudes can also be developed if *we* are the suffering party, but if somebody else was the victim, we would develop the same reactive attitudes towards the target.

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

Strawson proceeds and gives these 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.

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 humand rights and the people who suffer because of it.

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)

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 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.

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