Working title

Mischa

June 29, 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.

But first, let us examine the traditional way 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* [1].

"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."[1, p.4]

Thus, according to the instrumentalist definition, technology is something that is inheritly connected to humans and it's connection to morality and ethics can be found in the quality of the *human* end that it tries to achieve and in the way it is used by *humans*. Instrumental Theory:

[2]: Computers are moral entites but not moral agents For most of the time human technology has been used as a tool. The user or manufacturer has the responisibility for what the tool does. This has been working so far quite well and the instrumental theory was a valid and accepted extension of our moral understanding. We are now in a time, where new technologies, that exhibit more and more autonomy challenge this stance, that they are mere tools. How should we handle des question of responisibility with this new technology? Is there are responisibility gap?

Introduction of new technologies: ML

1.1 The Dirty Problem

According to Matthias [3] 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." [3, 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 theroy 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.

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

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?

[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 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

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

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

[3] 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.