Working title

Mischa

June 7, 2021

Contents

1 Introduction

Instrumental Theory 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 [1] 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." [1, p.175]

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.

Current practices in Artificial Intelligence (AI) aim at increasing the autonomy of machines and their software [zitieren]. 'Autonomous' 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].

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 appropriatley find 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?
2. What does moral responsibility entail?

The answer to the second question can easily be regarded as a functional definition of moral responsibility. 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

[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

society (mRINFES) without providing an argument for this assumption. Questioning this assumption would, I believe, fill 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

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

Yes we can: Here is how Instrumentalism 2.0 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. 1

AWS Autonomous Weapon System. 1

CR Control Requirement. 1, 2

mRINFES Moral Responsibility is Important and Necessary for a Functional and Ethical Society. 2

References

[1] 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.

[2] Stuart Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 2010.