UNIVERSITY OF AMSTERDAM

# The benefits of moral machines

## The advantages and challenges of making a moral machine

May 8, 2019

*Student:*
Tim Stolp

*Tutor:*
Mara Fennema

*Lecturer:*
Wouter van den Bos

# Contents

# 1 Introduction

In an increasingly more diverse society the quest to create moral machines is being pursued more and more. Practical applications such as the replacement of judges in criminal courts and self driving cars deal with situations where fair treatment of all people is a necessity. To meet this requirement it is important to define morals, but morals are not a rigid concept. Morals change over time and heavily depend on culture, personal development, and many other factors. [1] Therefore it is important to have a machine that can learn from new developments in society. The clear solution to this problem is the usage of artificial intelligence. But even for artificial intelligence it is required to give it some bases of morals. To figure out the philosophical questions about morals scientists have tried to find the answer in neuroscience. This report will go over the challenges of creating moral machines and the possible solution of using artificial intelligence combined with recent theories of morality in neuroscience.

# 2 Challenges

There are many situations where moral machines could be used. Some situations bring more challenges than others. The replacement of judges is one of the easier problems to solve as current judges need to abide to the laws of the country. These laws are usually very clearly laid out and thus could also be implemented relatively easy in moral machines. A more difficult case would be in self driving cars. Situations will occur when the car will have to choose between the lives of different people during an accident. When it comes to figuring out the answer to such problems it quickly becomes clear that there is no right answer. Many people have various different opinions, so there will always be people that disagree with the choice the car eventually would have to make. [1]

# 3 Possible solutions

A step towards a possible solution would be to create a wide variety of moral systems from which these machines could act. Theses moral systems would mostly be defined by the location and the culture of the society in which the moral machine is used. Using this approach would increase the relative amount of people that would agree on one answer in moral dilemmas. This is a somewhat democratic approach to ensure the happiness of the majority.

A different solution would be to come together and discuss the base rules of morality. This has been done by the Ethics Commission on Automated Driving. This commision has agreed on a few rules such as the priority of life over property and automated driving must in total cause a reduction in accidents. [2]

# 4 AI and morals

To practically be able to implement any type of solution it is necessary to make sure that the moral machine can learn from and adapt to many different situations. Artificial intelligence, and specifically machine learning, allows us to create such systems. With machine learning it is possible to have moral machines that constantly learn new things given a necessary amount of data. This data could come from many different places in society. Starting with previous records of moral dilemmas and the decisions that were made one could make a moral machine that is up to date with the present's morals. To make sure that the moral machine could continue to adapt is to provide it with information about various human behaviour such as voting behaviour, social media comments, etc.

A different approach to teach a moral machine is to look at neuroscience. Animals learn through specific processes in the brain. These processes share a lot of similarities with the way machine learning works. This makes neuroscience a likely candidate when it comes to trying to create machines to replace a human's moral judgment.

# 5 Moral neuroscience

In recent years the connection between morality and neuroscience has been researched extensively. Some findings indicate that morals might be innate to the brain. During experiments where brain activity was monitored during moral tasks it was shown that specific regions in the brain light up. Morality can not be assigned to just one region in the brain like with many other types of tasks. Many regions that have to do with high order behaviour types work in combination to create what seems like a moral network. [3] More research has to be done with different people to figure out what parts overlap when dealing with morality. This could possibly lead to some kind of objective truth to morality across humans.

# 6 Conclusion

Moral machines are a necessity to some technological advancement. There are some big challenges to overcome before these machines could be implemented. Artificial intelligence could be the solution to most challenges. Where artificial is lacking looking at neuroscience could provide us with solutions to finally create the first moral machine.

# References

[1] Bonnefon et al. “The Moral Machine experiment”. In: (2018), pp. 59–64.

[2] *Ethics Commission on Automated Driving presents report.* `https://www.bmvi.de/SharedDocs/EN/PressRelease/2017/084-ethic-commission-report-automated-driving.html`. Accessed:2019-05-08.

[3] MF Mendez. “The neurobiology of moral behavior: review and neuropsychiatric implications.” In: (2009), pp. 608–620.