Crowdsourcing Moral AI in Self-Driving Cars

Duke University

Jeremy Fox

December 4, 2016

Abstract

The thesis concerns the problem of moral decision making in self-driving cars. I examine some of the key philosophical issues involved in programming cars to make moral decisions and propose crowdsourced morality as a solution. I begin by surveying the problems inherent in this approach and explore random sampling as a possible solution. Finally, I conclude by suggesting further research avenues in this area.

Dedication

To mum and dad

Acknowledgements

I want to thank Professors Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, and the rest of the Moral AI group for their continued help in writing this thesis.

Contents

Chapter 1

Introduction

The current state of artificial intelligence mandates the need to imbue machines with the ability to make moral decisions. Autonomous vehicles (AVs) roam our streets (citation), algorithms allocate kidneys to the sick (citation), and within the next few years, we may see the arrival of AI used to target drone strikes (citation) and even sentence criminals (citation). Thus, it is imperative that we as a society take steps towards programming and regulating these machines in ways that preserve our moral values – failure to do so could be the difference between life and death. (cheesy, I know. Need to fix).

Of equal importance to programming *what* moral decisions AI will make is *how* we program it to make these decisions. Given that AI systems will soon make decisions that are currently made by government officials, we must consider who programs the AI, and how they are held accountable. Are these systems controlled by private organizations, governments, or individuals? How will we control them if they do not function the way we anticipate?

MIT's Iyad Rahwan recently explored this question and characterized it as the need to develop society-in-the-loop algorithms capable of extending the social contract assumed between citizens and their governments to an algorithm.

In my thesis, I explore the implementation of society-in-the-loop algorithms in AVs via voting. Why voting? A well-established method of connecting people to their rulers, voting is backed by centuries of political theory and has, depending on your viewpoint, enabled moderate to extreme success for the societies that rely upon it as a method of political decision making. Since voting has facilitated large-scale societal decision making in the past, it seems natural to ask if it might do so now, in the case of deciding how moral AI systems will act. Why AVs? They provide an excellent domain to study – the self-driving car trolley problem can be defined fairly easily, the algorithms that control the vehicles will affect society as a whole, and we as a nation have taken no significant steps towards deciding how this problem will be solved.

Chapter 2

Review of Related Research

2.1 Social Choice Theory

Here, I briefly review the basics of social choice theory as they relate to my thesis. For a more in-depth understanding of social choice theory and its applications in computer science and AI, I refer readers to the Handbook of Computational Social Choice [?].

2.1.1 Social choice theory basics

Voting informally seeks to aggregate the preferences of a group of people. More formally, if we have N voters and M candidates, we define a *ballot* as a ranked ordering over alternatives by a voter, and a *voting rule* as a function that maps from the set of ballots to a single ranked ordering.

Often, in social choice theory, we are handed a set of ballots – the question is how to choose the best voting rule to aggregate them.

2.1.2 Votes as noisy perceptions of correct rankings

Although many people view vote aggregation as a way of selecting the candidate that best agrees with the preferences of voters, this is not the only model of voting. In 1785, Condorcet proposed a model of social choice in which a correct ranking existed over all the candidates, and our votes were noisy perceptions of this correct ranking [?]. In Condorcet's model, voters were more likely than not to make a correct ranking rather than an incorrect ranking, they cast their votes independently of one another, and they were all equally likely to be correct. In this model, the question becomes: Which voting algorithm chooses a ranking that has maximal probability of being correct? Condorcet's Jury Theorem tells us that majority rule is the best decision function to use in this case.[**?**]

Theorem 2.1.1 (Condorcet Jury Theorem)**.** *Assume we have N voters voting over 2 alternatives for which their exists a correct ranking. Each alternative has an a priori chance of .5 of being correct, and each voter has a probability $.5 < p_i \leq 1$ of being correct. Let P_N denote the probability that majority vote over the groups preferences gives the correct decision.*

$$\lim_{N \to \infty} P_N = 1$$

This model can be extended to cases in which voters do not have equal probability of being right [**?**, **?**].

Theorem 2.1.2 (The Bayesian Optimal Decision Rule)**.** *If we have a dichotomous decision with each choice having an a priori probability of .5 of being correct, and voters cast their votes independently of each other, then weighted majority vote maximizes the probability of being correct, with weights w_i given by:*

$$w_i \propto \log \frac{p_i}{1 - p_i}$$

2.2 Artificial Intelligence Morality

In his *I, Robot* series, Isaac Asimov recognized the problem of programming morality into machines and explored the ways in which humanity's attempts to solve this problem could go awry.[**?**]. Today, AI morality is no longer the worry of science fiction – the potential benefits and dangers associated with artificial intelligence have been recently been highlighted by major news outlets[] and even the White House[**?**, **?**]. However, as the One Hundred Year Study on Artificial Intelligence recently acknowledged, the societal impacts and safety of AI are currently under-researched and under-funded.[**?**].

Chapter 3

Chapter Three Title

Chapter 4

Chapter Four Title

Chapter 5

Conclusion

Appendix A

Appendix Title