

The Upgrade Proposal

Theo Portlock

August 20, 2020

Abstract

In our lives, we try to make decisions that increase the possibility of a moral outcome. We judge others by their morality and their adherence to it. Identifying, understanding, and mitigating the flaws of human thought process and decision making is essential for human progress. The aim of The Upgrade is to create a tool that will make better, and more informed moral decisions than a human. The possibility of succeeding in this creation is low but increases with effort, time, expertise, and help. The possibility of this project achieving that aim is incalculable. However, if it does, the consequences will pave the way for the future of a more moral society.

Contents

List of Figures

LIST OF FIGURES

List of Tables

Chapter 1

Introduction

This chapter will outline the justifications for the current work. It will begin with a philosophical justification and will end with a set of aims pursued in the remaining chapters.

1.1 Morality

Morality is the set of principles used by individuals to distinguish between right and wrong actions. A moral agent is an individual that acts within a moral system. An event is considered morally virtuous (good) with respect to a moral system if its consequences increase the probability of a moral end. An event is considered morally reprehensible (bad) with respect to a moral system if its consequences decrease the probability of a moral end. A moral system can contain multiple ends but often, a combination of these ends weighted by personal importance can be assumed to be an end in of itself.

1.1.1 Majority moral opinion

These ends vary between individuals. When individual moral agents interact, one of three possible combinations can be observed. Firstly, a mutual rejection of moral systems can be defined as the assumption of one another's actions are bad. Secondly, an unbalanced interaction is a consequence of the willingness of only one moral agent to encourage the others pursuit of their respective moral system. Lastly, a mutual acceptance of moral systems is the understanding that the ends described by a moral system is aligned to the point that their continued actions weighted by their estimated probability is similar to those actions taken by the moral system of another individual. As with the legal system, an estimation of common "good" is necessary for collective arbitration of an action. Once a consensus is established, the application of the golden rule, the principle of treating others as you wish to be treated, forms the idea of justice and social contract. The search for this consensus is done at a personal scale during adolescence, or at a collective scale when we vote, complete polls, or protest. Communication of intention can often be reciprocal, that is to say one individual has an understanding of the others understanding of your understanding and so on. An understanding of the consequences of breaking this rule encourages the enforcement of a social contract.

1.1.2 How morality changes over time

The consensus morality can change over time. Judgement of people from different timepoints.

1.1.3 The current state of moral opinion

1.2 Progress

1.2.1 How is society improving?

1.2.2 How is society deteriorating?

1.3 The fallibility of human decision making

1.3.1 Bias

1.3.2 Misinformation

1.3.3 Missing information

1.4 Alternative decision making algorithms

1.4.1 Deep learning

1.4.2 HTM

numenta (?, ?), alpha go, etc

1.4.3 How to align intelligent systems with consensus moral beliefs

1.5 AI safety and responsibility

1.5.1 Is AI a threat to mankind?

1.5.2

1.6 The aims of The Upgrade

1.6.1 Probability of The Upgrade to succeed

Acknowledgements