Essay on Ethics
DD2380, Artificial Intelligence

Magnus Arvidsson, 930131-2170
magarv@kth.se

October 5, 2016

With artificial intelligene being a topic of very high interest again, and with it gaining a lot of attention it can be easy to forget that as the algorithms get smarter we need adress the problem of what role these new entities will get in a modern society. Artificial intelligence clearly has the potential to make most of our lives easier and more convenient but it comes at a cost.

As humans we expect our peers to have some sense of what is right and what is wrong. In a game of chess it is often pretty easy to define what will actually constitute the right move. While maybe not immediately obvious we can often argue logically and accept it as a consequence of the rules. This is also a natural way for computers and algorithms to handle problems. The issue is that, in taking decisions we are often faced with another kind of right and wrong, namely that of ethics and moral, take for example the trolley problem. The answers for these problems is eventually mostly based on what we feel is right, but feelings is not something inherently built into a computer, hence we must agree on a way to handle this.

Many moral and ethical dilemmas have been considered by philosophers and they have often come up with theories about how to act and how to argue when faced with these. Thus it may be tempting to simply hard code one into a computer and trust that it will do the correct thing. The problem is then that while seemingly applicable for one dilemma, often varying a small part of this would often make a human change her mind. Take for example the variation of the trolley problem where instead of a lever, you have the possibility to push a large man onto the tracks to stop the rampaging trolley. Many say that they would certainly pull the lever, thus dooming a person laying on the parallel track but not be prepared to push someone to their death, in order to save the persons on the original track. This ability to change our mind is not something we would hope for a computer with hard coded ethics to have since we cannot possibly foresee every situation it may encounter.

The conclusion then is that moral problems are complicated and even as humans you will eventually confront a situation when there is no clear answer. None the less, acting is often a necessity. The question we must then ask ourselves is if it is even possible to encode moral thinking in a language a computer can understand. If we don't even know what is the best decision, how could we have the ambition to code any behavioral mechanisms into a machine. One possibility is to implement a system that learns as it goes along, trying it best to adapt to the behavior it has seen before, but then another question

arises. Who should have the right to train these agents? There are certainly undesirable behaviors that we would not like them to adapt. Furthermore we would also not like this right to be too restricted since there is a possible incentive to make the process somewhat democratic.

Wordcount: 539