UNIVERSITY OF TARTU
INSTITUTE OF PHILOSOPHY AND SEMIOTICS

René Tõnisson

Can Autonomous Machines Make Ethical Decisions?

Bachelor Thesis

Supervisor: Mats Volberg (PhD)

Tartu 2016

Table of Contents

Introduction

This thesis explores the prospect of ethical decision making in autonomous machines. Continuous progress in the field of artificial intelligence has given rise to machines that are operating more and more autonomously from humans. Autonomous here is not meant in the philosophical sense i.e. having free will, but that they are operating without direct intervention by humans. While they are not conscious or acting intentionally, they are capable of operating on their own and making decisions. Previously they have been performing rather simple and safe tasks, but this is changing rapidly. The well-known example of self-driving cars illustrates the extent of complexity that their behaviour can have.

Machines like this raise several questions about their safety and trustworthiness. When they start operating around us, people need to know that these machines are not a threat to them or their loved ones and that they can be trusted.

To avoid problems like that, these machines need to be able to behave ethically at least in their domain. There is a relatively new field of research called machine ethics, which engages in problems like this. This field is attempting to answer the questions if and how is it possible to create machines that behave ethically.

In my thesis I explore how ethical principles could be implemented in autonomous machines and whether these machines can make ethical decisions.

The first chapter will explain what artificial intelligence (AI) is and how they are classified. There are several different classifications of AI, I am going to present one of them. It does not appeal to be one and only way, instead it describes AI in the context that is relevant to this thesis. After that I will give an overview of different types of ethics of AI: roboethics and machine ethics. In the end of the chapter I will discuss why there is a need for machine ethics.

The second chapter discusses different possibilities of implementing ethical principles in the machine. This chapter covers both the classical ethical theories and the alternative approaches. It also gives a few examples of experiments that have been conducted in this area.

The purpose of this chapter is not to find the best ethical approach for machines, but instead to give an overview of advantages and disadvantages of the different approaches.

The third chapter covers the internal model which supports the machine in ethical decision making by enabling it to predict a possible chain of events and simulate the possible actions to take. This allows the machine to find the best ethical action for the current situation. The concept of an internal model is not specific to ethics, but in this case it is examined as a tool that can be used in the interest of ethical decision making.

In the fourth chapter an overview of the nature of a machine's decisions will be given and the limitations of a machine's ethical decisions discussed. It touches aspects of why an ethical machine is not equal to a human as a moral agent, who is responsible for the machine's decisions and why a machine cannot have a moral status. It is concluded that a machine cannot make ethical decisions in the same sense as humans do, however, a machine is capable of making decisions that are considered ethically correct in a particular situation.

1. Intelligent machines

1.1 What is artificial intelligence?

Science fiction movies and literature tend to portray AI as conscious, self-aware robots or computer systems that are often hostile towards humans. The reality is much less dramatic. The possibility of conscious machines has been debated a lot over the years, some saying that it cannot be done and others being convinced they will be created some time in this century.

There have been a lot of debates on what exactly artificial intelligence (AI) means and there are many different definitions of it. Here are couple of ways it can be defined: artificial intelligence is a research field that is dedicated to building artefacts which display intelligent behaviour (Arkoudas & Bringsjord, 2014: 34).

> Artificial intelligence, the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. (Copeland, 2016)

What is meant by intelligence when people are talking about AI? This is often rather vaguely explained and might get confusing when considering the wide variety of different aspects of this field with many contrasting solutions that have been created. Human intelligence itself can be understood in multiple ways. In case of AI, intelligence is mainly involved with learning, reasoning, problem solving, perception and language processing (Copeland, 2016). In the highly regarded textbook "Artificial intelligence: a modern approach", intelligence is mainly seen as being concerned with rational action (Russell & Norvig, 2009: 30).

There are several ways to classify different types of artificial intelligence. For the purpose of this work, they will be referred to by the extent of their intellectual capabilities.

Narrow AI (also referred to as a weak AI or an applied AI) is a computer intelligence that is capable of acting in a single domain (Pennachin & Goertzel 2007: 1). It means that this kind of AI has been developed to do one specific practical task and it can be very effective at it, often outperforming humans. A recent example would be Google's AlphaGo winning four out of five matches of Go against world champion Lee Sedol (Metz, 2016). When we compare

a narrow AI to human intelligence, it could be seen as a human having some specific type of skill.

While the narrow AI can be very good at the task that it was made for, it is unable to do or learn anything else. Neither a chess playing AI nor AlphaGo would know what to do if they had to play checkers or tic-tac-toe. It cannot adapt to other domains, which is why it is called a narrow AI. Examples of the narrow AI are self-driving cars, facial or speech recognition software, robotic security guards, computer algorithms that trade stocks, etc.

Artificial general intelligence (AGI) is not restricted to any single area and should be able to act in a diverse range of domains (Bostrom & Yudkowsky, 2014: 318). C. Pennachin and B. Goertzel (2007: 7) have concluded that to emulate human intelligence, AGI should have the following abilities: both general and domain-specific problem solving, which it must be capable of unifying if necessary, ability to learn from different sources such as the environment, other systems, teachers, ability to improve at solving new problems through experience.

In the discussions about AGI, one often comes across the term "strong AI" which was originally coined by J. Searle (1980: 417). His definition was rather different from what is understood as AGI. What Searle meant by "strong AI" is a computer that was programmed in a way that it actually is a mind, it has understanding and other cognitive states (Searle, 1980: 417).

In 1962, I. J. Good brought forward the idea of an ultra-intelligent machine (now more commonly called superintelligence), the intellectual capacity of which would far exceed any human intellectual capacity in existence. Good suggested that since designing machines is one of man's intellectual activities, then an ultra-intelligent machine would be able to create even better machines, which would lead to an "intelligence explosion" (Good, 1965: 31-33). Superintelligence or super-human-level machine intelligence would be the next step from AGI (Bostrom, 2014: 4).

Together with AGI and superintelligence there are often talks about singularity. It represents a theoretical point in time where the superintelligence has evolved to the level that is incomprehensible to the human mind and the change is so radical that further events become unpredictable (Wallach & Allen, 2009: 57).

Unlike AGI and superintelligence, which still do not exist, narrow AI is already embedded into our everyday lives and its usage is growing daily. With the rapid development

of AI there are also starting to appear robots or machines that operate autonomously, without constant supervision by humans. Most of the spotlight in this area is on self-driving cars, but these are just one example. There are advances with autonomous military robots, care robots in hospitals and nursing homes etc. At this point it is important to note that not all robots contain AI, but a robot can be controlled by AI. In the following text, "machine" refers to a machine that is equipped with narrow artificial intelligence.

1.2 Ethics of artificial intelligence

The ethics of artificial intelligence could be divided into two distinct branches: roboethics and machine ethics.

1.2.1 Roboethics

Roboethics is a term that was coined by G. Veruggio at the First International Symposium on Roboethics in 2004 (Veruggio & Abney, 2011: 348). The main focus here is on ethical issues concerning the creation of robots and their application in society. Along with these come questions about protecting human dignity and limiting the divide and inequalities that could be created by robots replacing people in the workplaces and other walks of life. (Veruggio & Abney, 2011: 347)

Roboethics is a human-centered branch of ethics that promotes the development of robotics to advance human society. It attempts to achieve ethical, social and legal framework of robotics as well as professional ethics of roboticists. (Veruggio & Abney, 2011: 348)

1.2.2 Machine ethics

S. Torrance (2011: 116) defines machine ethics as a discipline that is concerned with designing machines that do things; which, when made by humans, could indicate that human as having an ethical status. It is investigating whether it is possible to create machines that make ethical decisions and behave ethically.

Are machines capable of acting ethically? Can ethics be made computable? (S. L. Anderson, 2011b: 163-164) Can machines be moral agents? Which moral standards should be implemented in the machine? (Wallach & Allen, 2009: 58, 78) Which ethical theories can be used as a basis for ethical machines? (Gips, 2011: 252) These are some of the questions that machine ethics is concerned with.

Machine ethics is not only interested in creating ethical machines, it is also examining ethics itself through the help of artificial intelligence. Several authors have seen AI as an opportunity to improve our understanding of ethics. M. Guarini (2011: 316) used artificial neural networks and machine learning to explore ethical reasoning in classification of moral situations. S. L. Anderson (2011: 25) says that machine ethics can help us to understand better what is ethical behaviour in practice, because implementing ethical principles requires a detailed approach rather than abstract reasoning. H. Seville and D. G. Field (2011) see AI as a possible assistant in ethical decision making.

1.3 Do machines need ethics?

They do not. Humans are the ones that actually need machines to have ethics. Machines that start acting autonomously in the world raise questions about safety and trustworthiness. Following the laws and ethical norms could alleviate these concerns. Laws are normally written down in a great detail, which cannot always be said about ethical standards and norms. For a self-driving car, it is relatively easy to follow traffic rules, but how should it act in a situation where there is group of people on the road and the only way to avoid driving into them is to crash into the wall on the roadside, which would kill its passengers?

Humans already trust computers and software. Using online banking software means that they trust software with their finances and personal data. Many medical devices for measuring or monitoring vital signs and administering drugs run software for doing that, so people trust their health with computers. Different kinds of security systems that guard property and people from intruders also run on software, so therefore people trust their safety to computers.

However, all of what was previously mentioned is not quite autonomous, many aspects of those solutions are directly or at least occasionally controlled by a human. Autonomous machines are very different in that sense. They are capable of decision making and operating completely on their own. Of course, there are already some autonomous machines in everyday use like robot vacuum cleaners, but not much ethics are expected from them and the main question of trust in this case is whether the floor really got clean or is there some spot left untouched. On the other hand, when self-driving cars would finally reach the point of common use, they are directly related to questions of safety and life and death. This kind of autonomy changes these aspects drastically.

It is not possible to hardcode all the right actions and outcomes for all possible situations into a machine. Engineers of the ethical machine cannot predict all possible situations and dilemmas that the machine will come across during its operational period. This means that we need to be able to create machines that can behave ethically on their own.

2. On implementability of ethical approaches in a machine

To be able to make ethical decisions, machines need to base their decisions on some ethical system or theory. At this point of time there is not a single universal ethical theory that is accepted by everyone and could be applied everywhere. Ethical norms and expectations tend to vary a lot throughout the world and communities. They are affected and formed by many different factors such as local traditions, religion, laws, and unwritten rules in communities. There are some views that are relatively common worldwide. Killing, stealing and lying are some of those acts that are unethical in most cultures but even then there are exceptions.

There are societies that do not have private property and no rules for stealing. In others lying or killing is prohibited only inside a local group while encouraged towards outsiders. Moral diversity does not have to occur only between societies, often times it can be observed also inside a single society. For example, vegetarians who believe that raising animals for food is wrong exist in the same society as non-vegetarians. Also, strict egoists, who see no point in helping others, live side by side with people who see the reasons for helping others. (Harman & Thomson, 1995: 9-11)

The high variation in moral differences adds considerable challenges to creating an ethical machine. Not only does it require finding a suitable principle to implement, these machines may also need to be customizable per society or region to meet local expectations and norms.

There are several different ways to approach the implementation of ethics. It has been somewhat common in the machine ethics field to divide them into top-down, bottom-up, and hybrid approaches. This concept is also adopted in this thesis, since it enables clear categorization of the approaches. The top-down approach is based on established ethical theories, the bottom-up approaches use artificial evolution and machine learning as the means to discover ethical principles (Wallach & Allen, 2009: 79-80). The hybrid approaches try to combine these two in a way that compensates their shortcomings or complements each other.

The predominant issue that comes up mainly with top-down approaches is the abstractness of the moral norms and concepts. Machines operate on a very detailed level, high level abstract concept like ‘good’ and ‘bad’ can be hard to convey to them.

Currently, in the fields where machine learning is used, artificial neural networks are trained with a very large number of examples until they are able to start making conclusions. In case of ethics, this kind of approach may be very time consuming, since it would likely require a large number of similar examples of situations which illustrate the concept they are trying to learn. Humans can often learn from very few or even single examples, which makes the process faster. From machine ethics point of view it would be very useful if machine learning also reached this point at some time. There is at least one company, Geometric Intelligence[1], that is working on this research but they have not published results yet.

2.1 Top-Down approaches

Under top-down approaches utilitarianism, deontology and virtue ethics will be examined as possible grounds for an ethical machine.

2.1.1 Actions guided by consequences

J. Bentham (1907) defined the principle of utility as a principle which approves any action that increases the happiness or disapproves any action that decreases it in regard of all those concerned. Rightness or wrongness of the action depends on its consequences, whether it increases happiness or causes unhappiness.

S. L. Anderson (2011) draws attention to some advantages that a machine could have compared to a human who follows utilitarianism. The calculation of happiness and desirability - machines are strict in the calculations, while humans tend to guess or estimate values and make assumptions on which action would give better results. This can easily result in error. Humans have a tendency to be partial and decide in favour of themselves or some people close to them while disregarding others who are affected by their actions. Machines on the other hand can be designed to be completely neutral. People also are more likely not to consider all relevant actions and circumstances, especially when it is a highly complicated situation with possible actions that have far-reaching consequences, while machines can be more thorough in their analysis. (S. L. Anderson, 2011:163)

[1] For more information on this company's approach can be found in the interview with the founder and CEO of the company Gary Marcus in the MIT Technology Review https://www.technologyreview.com/s/544606/can-this-man-make-aimore-human/ (last accessed at 17.04.2016)

J. Gips defined four features that an utilitarian machine should have:

1. A way of describing the situation in the world
2. A way of generating possible actions
3. A means of predicting the situation that would result if an action were taken given the current situation
4. A method of evaluating a situation in terms of its goodness or desirability (Gips, 2011:245)

First three involve their own complications, but are relatively straightforward tasks compared to the fourth. The first one deals with continuously acquiring data from the world about the present moment and interpreting it accurately in real time. The amount of data in this case can be very large with a high noise to signal ratio, which means that the algorithm that separates relevant information from the noise must be optimized for speed and accuracy. The second task must be able to take the interpretation of the situation from the first task and generate, based on the given data, all the possible actions that the agent can take. The third one needs to run a simulation for each possible action and analyse the consequences. This immediately gives rise to the question of how far-reaching should the consequences being considered be? There needs to be some way to reasonably limit the scope of the simulation, both in the sense of how far into the future consequences will be included and how many variables need to be considered in relation to each action. Trying to encompass absolutely everything even seemingly relevant would create either a never-ending simulation or it would occupy all available resources so exhaustively that the machine could get stuck with the simulation. Also, the further into the future consequences the calculation goes, the more inaccurate it will become, because of the number of unforeseen variables and circumstances that can enter into the equation.

While the first three are mainly data processing and calculation issues, the fourth one creates a much different problem - how to evaluate the goodness or desirability of the situation? Since in classic utilitarianism each person counts equally and there are no hierarchies, the machine should calculate the amount of happiness and pain caused by the action to each affected person and sum it up. This would give the total amount of happiness and pain caused by each action and machine should choose the action with highest total happiness. (Gips, 2011:245) A rather simple formula, except how does one measure happiness or desirability? Gips himself evades the direct answer here by saying that measuring is up to the creators of the machine and their theory of value. W. Wallach and C. Allen (2009:89) mention an option of gathering as

many subjective utility ratings as possible, applying a weighing formula to these and then incrementally fine-tuning the machine's decisions and actions until they are adequate enough. They do not specify how exactly the collection of this data is supposed to be done, but they do admit that there would be serious difficulties in gathering the assessments in real time.

J. Bentham (1907) mentions seven circumstances that determine the value of pleasure and pain caused by the act to people affected: intensity, duration, certainty or uncertainty, propinquity or remoteness, fecundity, purity, and extent (represents the number of persons who are affected by the act). For every person that is affected by the act, using these circumstances, a calculation can be made, where pleasure is summed up on the one side and pain on the other. After that, summing up these numbers over all the affected people, would show whether the balance is towards pleasure or pain, which will give the general tendency towards good or bad of that particular act. (Bentham, 1907)

The very same procedure of calculating the morality of the action is what makes utilitarianism appealing to machine ethics. On the other hand, aforementioned problems with gathering and assigning values to the involved variables seem to undermine this possibility.

According to utilitarianism, the end justifies the means and the end goal is maximizing the happiness. Following these principles might result in acts that are unjust or seem morally problematic. For example, imagine a case where a doctor kills one healthy person to harvest his organs that will save the lives of five others. In the big picture this would result in five happy people who can go on living, which means their combined happiness is bigger than the killed person's happiness could ever have been. Does not seem a very moral thing to do, but on utilitarian grounds, it is justified. If pure utilitarianism was implemented as a moral principle for a machine to follow, it may very well result in similar situations. While this robot may not necessarily start harvesting organs, it could start maximizing happiness in some other situations that would end up in morally questionable actions and serious damage to people or their property. (Grau, 2011: 453)

The exact implementation of utilitarianism to the machine does not seem feasible, but it also should not be ruled out entirely. It could be used as an additional measure in dilemmas where for example rules do not offer solution. To be usable, it should be implemented with limited scope, meaning it considers a short range of actions and those who are immediately affected. This may not guarantee maximum happiness, but can add additional weight to one of the options in dilemma and enable making a choice.

2.1.2 Rules as basis of ethical decisions

Deontological approaches would see machines making decisions and acting according to a set of rules. Rules can be turned into an algorithm for a machine, which it would use to compute its ethical decisions. (Abney, 2011: 41)

One of the most famous deontological approaches that was geared specifically towards robots, is I. Asimov's Three Laws of Robotics. Regardless of the fact that he had prioritized them in a way that should have minimized conflicts between laws, he later kept demonstrating in his stories that these rules nevertheless would lead to deadlocks in a robot. (Abney, 2011: 42-43)

I. Kant's Categorical Imperative is another deontological approach that is widely debated in machine ethics. T. M. Powers has proposed that Kantian ethics could be used as a basis for an ethical machine if Kant's ethics is not followed strictly. Instead he focuses on a machine computable categorical imperative. (Powers, 2011: 466)

He suggests the possibility that a machine might be able to build a theory of ethics by using a universalization step on maxims and aligning them to the categories of forbidden, permissible and obligatory actions. Powers specifically states that it should not be a human who creates the moral laws for the machine, because then it would human ethics, not machine ethics. (Powers, 2011: 466) Powers is not very clear on how the machine would find the maxims. Since they should not come from us, it seems that therefore they should be somehow generated by a machine, but the question is – based on what? To speculate, if the machine has been fed descriptive information about the world, then it might be the grounds from which to start. This is one option but I am not making any claims on whether Powers meant something similar or not.

For a rule to be universal, it must fit into the system of rules that applies to all agents and must be consistent with other rules that have been generated in the same manner. This is what has been interpreted as a two-step consistency check of universalizability and systematicity. The machine would build a theory of forbidden maxims which would state what it ought not to do. After that it would move on to obligatory maxims which specify what the machine ought to do, and then permissible maxims that are neither obligatory nor forbidden. (Powers, 2011: 466-467)

When two or more rules conflict, there is no principle that would specify how to prioritize them, which would require creating a procedure for a machine to decide the priorities of rules (Powers, 2011: 472). Conflict of rules is an issue that one comes across in many deontological systems that would create a problem in the machine following them. Two opposing rules can create a deadlock in a machine that would bring it to a situation where it ceases operation since it cannot take any actions or would seemingly attempt to take multiple actions simultaneously. Another version of this problem is a situation where the conflict happens within the same rule. When two people are in danger and the machine can save only one, how should it decide? What if it cannot make a decision at all and both people die?

Therefore with creating this kind of a machine, a principle to avoid these conflicts must also be created. Even a bad decision can be sometimes be better than no decision at all. The aforementioned limited utilitarian procedure could be one option. It can be used to measure which conflicting rule gives the most positive outcome (or the least negative) and decide to follow that rule.

Wallach and Allen (2009) have also offered an option of how categorical imperative could be implemented in a machine. Of course there is the problem that machines do not have a will, so they cannot will anything to be a universal law. Despite of this, the categorical imperative can be used by a machine as a formal tool to check if the maxim is moral by comparing whether the goal is attainable if other agents acted in the same way in a particular situation. Three elements are required for running this simulation model – a goal, a course of action, and a statement of the circumstances under which to achieve this goal. This kind of simulation could then show what would be the outcome of pursuing a specific goal, if all agents acted upon the same maxim. Wallach and Allen suggest that using this tool may not be suitable for every machine, because it would require a lot of computing power. (Wallach & Allen, 2009: 95)

In regard of the required computing power I disagree with them. This simulation does not have to depict each and every individual agent that there is, instead it could do calculations for an estimated total numbers of these agents and the collective impact of them when acting on maxim. This would make the simulation achievable to a much wider range of machines.

R. Shafer-Landau (2009) brings to attention that universalizability does not necessarily guarantee that a maxim is moral. For example, if a bank robber does not rob the bank for his personal gain, but to destroy its business. If everybody did that, banks would go out of business,

which means that the maxim is universalizable, but it definitely is not moral. (Shafer-landau, 2009: 154) In this light it is not hard to see how a machine, which only follows universalization check, can end up with maxims that are not moral. This means it still needs additional means to assess the morality of these maxims.

R. Tonkens (2009) says that a Kantian machine cannot be created at all, because it would be anti-Kantian. Kantian ethics requires an agent to be free to be a moral agent. A machine is not free, which would make it anti-Kantian. Also, creating these machines would be anti-Kantian, because they would be treated only as means, and not as ends. (Tonkens, 2009: 429-431)

Another theory that could be categorized under deontology, even though it is not as strict as some other theories may be, is prima facie duties. It maintains that there are no absolute duties, rather there are duties that people should try to follow and which can be overridden by other duties which in a particular situation have a higher importance. The problem with this theory (as with several other deontological approaches) is that it does not have the decision procedure for determining which duty prevails in case of a conflict. A human actor, again, could use intuition, but this does not apply to a machine. (Anderson & Anderson, 2011: 476)

To overcome this issue, M. and S. L. Anderson (2011) focused on the problem of discovering the decision-making principle that could be used by a machine to determine the correct action in case of conflicting duties. For establishing the prototype solution to the problem, they chose Beauchamp and Childress's principles of biomedical ethics, which is a prima facie duty theory that has four duties. Andersons used three of them: patient's autonomy, non-maleficence and beneficence. The ethical dilemma that was used as a basis for working out the decision principle was the one where the patient refuses to follow the treatment that was recommended by the doctor. The question is whether the doctor should accept the patient's rejection or try to change the patient's mind? (Anderson & Anderson, 2011: 477)

They gave different numerical values to the factors that influenced the outcome and gave examples of preferred actions for four cases, so that a machine-learning procedure which used inductive logic programming was able to extract a general principle for decision-making. This principle was used by the Andersons for developing two applications: MedEthEx - a medical advisor system and EthEl - a medication-reminder system for the elderly. (Anderson & Anderson, 2011: 478-480)

This experiment could be seen as a proof of concept due to its very limited scope and data. The data was given to the machine by its creators, it did not acquire it itself nor had the means to do that. It does not give a good indication whether it would work in large scale systems. The main concern in my opinion is whether this same approach can be used for a machine that operates in one domain but comes across many different situations. Is it capable of successfully executing this same prioritization procedure on the run? It is unlikely that it could be ran in advance to be prepared for situations. One option would be that the results from previous situations are saved and used as heuristics for situations that are similar enough. Another option would be using earlier results as seed information for the procedure to avoid starting from zero. Also, it can be approached as a range where based on the percentage of the similarity with the current case, previous results are either used verbatim or used as basis for new results by combining data.

As could be seen from above, the deontological approach is attractive to machine ethics because of rules that could be turned into algorithm. The disadvantages are possible conflicts between rules and within the same rule. Before implementation there needs to exist an established procedure for a machine, on how to deal with these conflicts. Also, rules need to be made detailed enough and they need to be substantial enough, so they can be applied to a necessary range of situations.

2.1.3 Machines of virtue

Virtue ethics is an approach that instead of duties and consequences focuses on building a good character. Instead of asking which actions one should take, it asks what kind of a person one should be. (Hursthouse, 2003)

Wallach and Allen (2009) explore whether it is possible to program virtues into a machine. The first major problem is that virtues can contain complicated patterns of motivation and desire and may affect a large number of the actions one takes. Broad virtues, such as kindness, can cause situations where the machine, while checking whether an action is compatible with a virtue, can become stuck in endless loop. For example, in a situation where two people request something from a machine but it is capable of fulfilling only one request. Let us say that both of these requests are equal by nature and the person whose request will not be fulfilled would consider the machine to be unkind. Being unkind, however, is unacceptable to the machine, because it has been programmed to be kind. How should the machine decide

which request to fulfil? Unless some fall-back mechanism has been implemented for situations like that, the machine very likely cannot make the decision. (Wallach & Allen, 2009: 120)

The example above presents a machine that is rigidly kind, it cannot handle situations that it cannot fully cover with its virtue. This suggests that the machine cannot be 100% virtuous, it has to be able handle exceptions for dilemmas. When a human who is kind ends up in this situation, most likely he/she will fill one person's wish, acknowledge the fact that he/she cannot fulfil both and accept that the other person may consider it unkind. Similar exception handling needs to be implemented in a machine for it to be able to actually function (and not get stuck in dilemmas on regular basis).

Wallach and Allen (2009) bring out an important benefit of the broad virtues - it is stability and trust that they could create over a wide range of features. If one demonstrates virtue in one situation, then it is likely that one acts similarly in other similar situations. Stability over a wide range of circumstances is particularly attractive in case of machines, since an ethical machine needs to be reliable and predictable. (Wallach & Allen, 2009: 120-121)

It is hard to see, though, how exactly implementing a virtuous character into machines would take place. They are normally created to act with a certain purpose, not built for their character. Gips (2011: 249) says that nowadays virtue-based systems tend to be converted to the deontological system e.g. act courageously instead of be courageous. This would very likely happen in case of machines also. Even if a machine was built with character, it would still end up evaluating actions against some criteria that would be its "character traits" that qualify or disqualify the action. If these "character traits" were called rules, would there actually be a difference? As long as we're dealing with narrow AI in the machines, the virtue approach does not seem reasonable or even implementable option.

2.2 Bottom-up approaches

The 'bottom up' approach takes a completely opposite path to the development of the ethical machine. Instead of trying to implement a specific existing ethical theory, it attempts to introduce moral values to the machine through evolution and learning. This could be a relatively slow process, since evolution and learning through experience, trial and error, and unsuccessful strategies is time-consuming. (Allen, Smit, & Wallach, 2005: 151)

On the other hand, considering the sheer complexity of ethics and morality, it seems reasonable that their machine implementation is not a short term project. Even the development of simpler machines that do not appeal to such intricacy can take considerable time.

2.2.1 Possibility of ethics through machine evolution

Wallach and Allen (2009) point out an idea from sociobiology that the origin of morality may lie in human evolution. Inspired by this, they see genetic algorithms and artificial life (ALife) simulations as one potential source for developing machine agents with ethical behaviour. In genetic algorithms, populations of robots vary slightly from each other and are evaluated on how they perform at certain task. Success (fitness) is measured by assigning a score to their performance. Best performing robots are used to generate new population, by recombining their components and adding small random mutations. This process is repeated for many generations, which results in continuous improvement of their performance. (Wallach & Allen, 2009: 100-101)

At this time, ALife simulations are still far from the complexity of the real world and experiments for evolving ethical machines there cannot be made, yet. Nonetheless, due to continuous progress towards more sophisticated virtual worlds, researching evolution of ethics in ALife could one day be a possibility. (Wallach & Allen, 2009: 104-105)

If human ethics and morality are a result of human evolution, then using ALife and evolutionary algorithms could be an interesting opportunity for ethics in general. Simulating evolution though artificial agents could give an insight into moral decision making in humans today and why morality has reached its current state.

2.2.2 Ethics through learning

People acquire morality normally by learning through experience, which gives an idea that maybe it could be done by machines also. However, machine learning is far from the level of learning that humans are capable of. In moral development, children receive feedback in form of reward, punishment, approval and disapproval. While machines would need some sort of feedback in their learning, too, it is hard to find similar equivalents in a computer. (Wallach & Allen, 2009: 106-107)

There does exist model in machine learning that uses reward for the right action. It is called reinforcement learning. In this case nobody tells the machine what the right actions are, instead, it receives feedback called reward that indicates that it took the right action. What is

indicated as a reward varies a lot, depending on the task that the machine is learning. For example, if the machine learns to play table tennis, then every point that it scores is a reward or if it is learning to crawl then forward movement can be reward. The important aspect here is that the reward must be something specific that a machine can recognize from all the other inputs it receives. (Russell & Norvig, 2009: 830) In case of ethical situations, it may not be easy to find such reward which confirms that the machine made the right decision. Quite likely it can be found for the same type of situations but if the situations vary a lot, there may not be this common denominator to identify the right outcome.

If we were to create a machine which would acquire ethical understanding through learning, it cannot be released into society from day one, because at that moment it is not yet an ethical machine. Instead, it would be a clean slate, ethics wise. It will only just start learning through simulations and later can be gradually integrated into society.

At this time there is no clear model of how humans learn. Assembling diverse fragments of information with background knowledge by analogy and other processes is something that cannot be done by machine learning, yet. This does not necessarily mean that there is something fundamentally different between human and machine learning or that machines could never learn like humans. Instead it insinuates to the problem that there are still significant gaps in our understanding the process. (Danks, 2014: 161) Nevertheless, even if the machine cannot learn like a human, it is important that it uses some kind of learning, so it could adjust and update itself in its environment. (Allen et al., 2005: 152)

There are several potential dangers with learning machines, especially if they are learning on their own without supervision. A somewhat obvious one is that a machine might learn the wrong things and as a result develop into an immoral machine. (Wallach & Allen, 2009: 110) The question is, how will the machine tell the difference between ethical and unethical or good and bad in its learning process. If this machine starts from a clean slate, then it does not know what "ethical" is and at least for some period of time would need to be trained with clear examples prepared by its creators.

The machine could also start altering its code and bypass built-in constraints, unleashing unwanted features with unexpected consequences. One possible solution to the self-modification problem could be a layered architecture, where the lower level protocols are isolated from the higher level functionality, which would make them inaccessible to the parts

that learn and process new information. Of course this does not guarantee that somehow the learning machine is not able to circumvent these restrictions. (Wallach & Allen, 2009: 110-111)

Self-reprogramming machines are mostly talked about in topics of AGI and superintelligence, where the idea is that AGI keeps creating improved versions of itself until it reaches superintelligence (Bostrom, 2014: 29). Machine's ability to completely reprogram itself still lies further in the future, but self-modifying machines themselves actually already exist. It is called autonomic computing and it was introduced in 2001 by IBM. These systems are capable of self-configuring, self-healing (repairing software and hardware problems), self-optimizing and self-protecting (defending against attacks and cascading failures). (Kephart & Chess, 2003: 41-43) It is possible that some autonomous machines will also receive some level of ability to self-maintenance, whether it is self-configuration or fixing problematic code. Considering that these are not AGI level systems (which we are still far from), they most likely will not have the ability to create more advanced versions of themselves, but there is still the risk of bad repair that will change its behaviour. This is why there should be some parts of the system that cannot be fixed autonomously.

To create a general ethics machine through learning, it needs to be able to learn similarly to humans – on its own, unsupervised, and be able to judge what is and what is not ethical behaviour. Only this way it could adapt on its own to new regions, otherwise it would have to be customized for every new country or region in the world. An AI that is capable of learning generally, like a human, is more likely to be AGI than a narrow AI. While machine learning has not reached this level yet, it can still be used as an effective additional tool for the ethical machine.

2.3 Hybrid approaches, combining top-down and bottom-up

There is a possibility that neither top-down nor bottom-up approaches alone are enough for creating an ethical machine and instead it is necessary to combine both of them into a hybrid approach. Top-down approaches capture generic concerns of what is and what is not acceptable behaviour; bottom-up approaches, through evolving and learning, can provide systems with flexibility and choices that cannot be retrieved from the rigidity of top-down systems (Wallach & Allen, 2009: 117-118). Top-down systems are often rather broad and abstract, which complicate their implementation. Bottom-up approach might be of help here by making them

more specific with learning through examples. A few examples of possible hybrid approaches are discussed below.

In a highly complicated game of Go, Google's artificial intelligence AlphaGo beat the world champion Lee Sedol (Metz, 2016). In any game, just knowing the rules is not enough to be good at the game, the player has to learn to play the game. Just programming in the rules was not enough for AlphaGo to win, it had to learn to play. Its neural networks were trained with expert human moves and it also played many games between its own neural networks (Silver et al., 2016:484).

Even though AlphaGo is not ethics related, this idea might be adaptable in machine ethics. For example, creating a machine with a hybrid approach that combines deontological theory from top-down and machine learning from bottom-up. There would be defined rules for the machine to follow, which it would learn to apply by learning through examples and by "playing through" simulated life-like situations under human supervision.

It is obvious that there are many things that make it much easier for a machine to learn a game than it is to learn ethics. A game has a set of possible moves and a fixed area for movements on table. It also has straightforward non-abstract rules that define the game. Ethics can be much vaguer in a wide, highly complex world with an unknown number of possible situations. This makes it clear that it will be a lot harder to create a system that is capable of this kind of learning, which does not mean that in the long run it cannot be done.

Another example of a hybrid approach is an experiment, conducted by M. A. Pontier and J. F. Hoorn. They combined a bottom-up structure with top-down moral duties. It balances duties with computed levels of morality which is the estimated influence on the total amount of utility in the world. The system (moral reasoner) is connected to the Silicon Coppélia (SC), which is a cognitive model of emotional intelligence and affective decision making. (Pontier & Hoorn, 2012: 2199)

In the simulation they focused on biomedical ethics, because it has an established wide consensus on what is ethical. Also, healthcare is a domain where the use of robots is on the rise due to a predicted lack of resources and qualified personnel. As for moral duties, they chose to concentrate on autonomy, beneficence and non-maleficence. (Pontier & Hoorn, 2012: 2199)

The system has library of goals which have a level of ambition for every agent. Levels of ambition that an agent attaches to the goals have a real value between [-1,1], where negative

value means an undesired and positive value a desired goal. A higher value means a higher importance to the agent. (Pontier & Hoorn, 2012: 2199)

The library also contains actions that an agent can perform. Actions also have a real value between [-1,1] where a negative value means an action is inhibiting the goal and a positive value means an action is facilitating the goal. The expected utility of an action is calculated by looking at its influence on the goal. (Pontier & Hoorn, 2012: 2200)

The three duties - autonomy, beneficence and non-maleficence - are seen by an agent as moral goals. These ambition levels of these moral duties were set in a way that autonomy had the highest ambition level and beneficence had the lowest. The estimated level of the morality of an action is calculated by an agent by taking the sum of the ambition levels of the moral goals multiplied with the beliefs that the actions facilitate the moral goals. (Pontier & Hoorn, 2012: 2200)

To be able to verify whether the moral reasoner was capable of simulating ethical decision making similarly to experts of medical ethics or not, they took test cases for experiments from A. E. Buchanan's and D. W. Brock's book „Deciding for Others: The Ethics of Surrogate Decision Making" (1989). (Pontier & Hoorn, 2012: 2200)

They ran six experiments with scenarios where a patient refused treatment for different reasons (fear, religion, false beliefs etc.) with the question being whether the doctor should accept the patient's decision or try to persuade him to change his mind. Expert analyses of these cases were taken from the aforementioned book. (Pontier & Hoorn, 2012: 2200-2201)

The results showed that the moral reasoner reached the same conclusions as experts. A notable case was experiment number 2. During the development of the system, the author had added a specific rule to the moral reasoner to ensure that the decisions of fully autonomous patients would not be questioned. That same rule was the reason why in experiment 2 (refusal because of religious beliefs) the system reached the same conclusion as experts. Without it, the moral reasoner would have reached an incorrect result. (Pontier & Hoorn, 2012: 2200-2201)

There is no one fixed formula that a hybrid system must follow. This is the advantage of this approach. Top-down and bottom-up approaches can be combined in a way that is the most effective in particular solution. Of course, due to the complex nature of different systems and theories, issues with merging will arise, but nevertheless there is a great promise of dynamics and adaptability in the hybrid approach.

2.4 Domain-specific ethics

Machines so far are designed to function in a specific domain and defining what is ethically acceptable in a particular domain is an easier task than trying to create a general theory of ethical behaviour at once (Anderson, 2011: 25). There are many areas which have a rather wide consensus on what is and what is not ethical in that domain. Medical ethics is one of the examples here (Pontier & Hoorn, 2012: 2199).

In many cases, implementing general ethics would be unnecessary. A machine with a narrow AI would not have any use for a large part of general ethics. Moreover, it probably does not even have the means to comprehend the situations that are outside of its domain, let alone have the ability to apply ethical principles there.

The magnitude of the task of implementing general ethics stems from the level of details that is required to make it usable for a machine. As mentioned earlier, ethical approaches tend to be rather abstract which would make them incomprehensible to the machine. Ambiguity in a machine's code can yield unexpected and undesired results which means that these abstract concepts must be turned into very specific code.

Domain-specific ethics could pave the way to generalizations that may be applicable to other domains (Anderson, 2011: 25). Creating general machine ethics will be an incremental process. Starting from domain-specific ethics it could be expanded over time, eventually reaching generality. Now it could be asked, what if ethics from different domains cannot be consolidated? This is expected and they do not have to be fully compatible. Developing general ethics with this approach would mean that it can use the commonalities that arise from domain-specific ethics. Domain-specific ethics should be considered a starting point or a test bed, from where general machine ethics could evolve over time.

2.5 Ethical approaches in retrospective

Neither top-down nor bottom-up approaches alone seem to be sufficient for implementation into a machine. Disadvantages of top-down are abstractness and the generality of their concepts. Utilitarianism places heavy computational demands on a machine and problems can arise with acquiring all the relevant information. On the other hand, if offers a method to evaluate the morality of an action through its consequences. Deontological approaches face possible conflicts between the rules and within the same rule. Then again, the advantage of these rules is that they can be turned into an algorithm.

To overcome these limitations, top-down approaches could be combined with bottom-up, specifically with machine learning. While machine learning by itself is not yet powerful and dynamic enough to be the only means for acquiring ethical principles, it could compensate the shortcomings of top-down.

Attempting to implement general ethics at once might be too extensive task, therefore domain-specific ethics seems a more attainable approach. Many specific areas where a machine would operate have well-established ethical rules and over time the commonalities in domain-specific ethics can give rise to general machine ethics.

3. Internal models and functional imagination in ethical machine.

In the previous chapters were mainly discussed different approaches of implementing ethical principles into a machine. These are the grounds that a machine bases its decisions on. This chapter focuses on a different, more technical feature that can play important role in machines decisions. It is called internal model.

Internal model is a mechanism which is embedded into a machine that allows a machine to run a simulation of itself and its perceived environment. A machine with an internal model can test what-if hypotheses of its possible actions and their consequences in a given situation, without having to actually perform them. (Winfield, Blum, & Liu, 2014: 86)

Internal models and functional imagination, described below, are not specific to an ethical machine but have been used in robots for some time with different purposes (Winfield et al., 2014: 86-87). The role of internal model in ethical machine is to enable machine to anticipate the situations that require ethical decisions and in simulation play through different actions. This, as mentioned before, enables the machine find out the consequences of the actions without having to try them out in real life. As a result, a machine can choose action with best possible consequences.

The concept of using internal model for the purpose of ethics in a machine was proposed by A. Winfield, C. Blum and W. Liu (2014). More detailed overview of their experiment is given further below. Before that a short overview of functional imagination will be given which is general concept of the internal model in a machine.

Functional imagination is system that lets the machine to simulate actions, predict their consequences and elicit possible behavioural benefit in the process. Behavioural benefit here is understood as obtained net reward in relation to either external or internal factors. (Marques & Holland, 2009: 746)

Functional imagination can have two types of forward models. These models are used to predict the next state of the system in regard of current state and executed action. Unmediated forward model takes current state and action as input and gives predicted state as output. Mediated forward model utilizes an intricate simulation where outputs are determined through

the reconciliation of self-model interacting with a simulated world. (Marques & Holland, 2009: 746)

For properly working functional imagination, it is essential that agent can select several different actions in the particular situation that it is simulating. (Marques & Holland, 2009: 747) Obviously, the actions that can be selected in simulation, must be the same that can be selected in external situation.

Important characteristic of functional imagination system is how many steps ahead it is capable of simulating. The simplest case is a system that can simulate only one step ahead, more complicated systems can simulate multiple steps. In multi-step systems, number of possible steps is mainly determined by its computing power but there may be also other limiting factors such as noise. (Marques & Holland, 2009:748)

Another relevant factor is, whether system stores the solutions of previously successful solutions of the encountered problems (memorising) or not (memoryless). (Marques & Holland, 2009: 748) For an ethical machine it would be useful to store previous cases for learning purposes and reusing them in similar future situations, which can save time it takes to make a decision.

A. Winfield, C. Blum and W. Liu (2014:86) implemented this kind of internal model in an ethical action selection mechanism called consequence engine.

Consequence engine is not tightly integrated into robot's controller, rather it is running in parallel with it and is separable. This means that robot can run effectively without it, but is not able to estimate the consequences of its actions. (Winfield et al., 2014:87)

CE is capable of modelling possible threats for all the actors and itself in the perceived environment. Capability of anticipating the outcomes of agent's actions in the environment allegedly gives the machine very basic theory of mind for that other agent. When all possible actions have been tested, the most suitable action will be sent to a robot's controller. (Winfield et al., 2014: 87)

They tested CE with small mobile robots, where robot A contained CE, and other robots, that represented humans, did not. In the field where these machines moved around, was a virtual hole (harm), that robot A could perceive and others could not. The goal for robot A was to avoid the hole and also keep other robots from reaching the hole. (Winfield et al., 2014: 90)

When robot A was moving in the simulation field alone, it managed to avoid the hole (stay safe) 100% of times. In trial where there was another robot H (representing human), robot A's CE recognized that H is on path of danger and diverted from its own trajectory, managing to successfully save H and also stay safe itself. (Winfield et al., 2014: 8-9)

In the third trial they added another robot H2 to the arena. In this case robot A managed to save human representing robot 58% of times and both 9% of times. However, this was not because of choice, but because of noise. The robots did not start at the same time or not from the same position and CE indicated one of the H's problems first or there was enough time to also reach second robot H after saving the first. If all things were equal, CE did not have preference for either of H robots, meaning that robot A got stuck in a dilemma and did not save either of them. (Winfield et al., 2014: 92-93)

Authors of this experiment noted that to solve this issue, they could have introduced some sort of heuristic or a rule, but they deliberately did not as it should not be solved with engineering but with ethics. (Winfield et al., 2014: 95)

This experiment illustrates the importance of using internal model in a machine. It enables a machine to predict possible course of events and consider suitable ethical action (in this case to prevent "human" from coming to harm). Even though the environment and number of actors in this experiment were very small, it revealed some important aspects. First one was already mentioned importance of internal model which enables to predict consequences. Second, is the problem of the ethical dilemmas. In situation where two "humans" were moving towards danger, often the robot could not decide which one it should save and eventually saved neither of them. If a machine ends up in this kind of a deadlock, it means that the values for actions to choose from are equal and it needs additional criteria to make a choice. If that criterion cannot be found, there is something missing from the ethical approach that was implemented in the machine. Since, ethical approaches are not perfect, there should be implemented a contingency principle that can add additional value to one of the choices in a dilemma. This contingency principle cannot be universal, because the solution would depend on the particular implementation that a machine has.

4. Decisions, responsibility and moral status

One could argue that the machine's decisions are not made by a machine but by humans who created the machine and the program that is running it. Even though humans definitely play significant part in the process, every individual judgement made by a machine cannot be attributed to its human creators. For example, every single move made by a chess-playing computer or every individual diagnosis given by a medical diagnosis advisor cannot be attributed to designers and developers of the program. Applications like these are normally created by the teams that contain several people with different specialties. Many of those people probably play chess much worse than the program or are not at all familiar with a medical diagnosis on a professional level. (Whitby, 2011: 142)

The assumption that machines only follow the program that is implemented in them is misleading. This may be true in case of very simple solutions, but it is not so straightforward in much more complex systems. Program as a set of instructions is indeed essential part of it, but a programmer does not determine every possible outcome there is. For example, a system that incorporates machine learning or case-based reasoning will have evolving responses and generalizations in the light of the new information it acquires. More adequate explanation would be that a programmer created a collection of decision-making procedures which enable a machine to make decisions in a given situation. (Whitby, 2011: 140-141) Of course this does not mean that the machine is not limited in its actions, it is still bound to the range of activity that its code allows.

From machine's perspective ethical decision is no different from any other decision it makes. It could act and reason ethically, yet it would not know it was being ethical. To know what ethical was, the machine would have to encounter a situation where it is torn between acting in self-interest or in the interest of someone it cares about and doing the right, ethical thing. (McDermott, 2011: 89) According to this, a machine would have to have self and likely self-awareness, to make ethical decisions. Currently there are not yet self-aware machines and the machines are also not capable of caring about someone, but they can be developed to act in the interest of their creators or their owners. In this case a machine can be caught between

ethical decision and acting in the interest of its owner. This, however, would not bring it closer to knowing what ethical was, since acting in someone's interest would mean that its algorithm is created in a way that prioritizes particular individuals over others, regardless of the ethical principles that have been implemented into the machine. To actually understand the difference a machine would have to have some sort of awareness.

Prevalent thought tends to be that to behave ethically one must be able to act intentionally, have consciousness and free will. Having sentience and emotions are also seen as essential factors, because only someone who has feelings is able to comprehend the feelings of the others. Machines do not possess these qualities and it is not sure that they ever will, which seems to prevent them from becoming moral agents. (S. L. Anderson, 2011: 164)

S. L. Anderson (2011) says that free will and intentionality are necessary if someone has to be held morally responsible for the action. Yet, these are not needed to perform morally correct action in situation which requires a machine to behave ethically and to be able to justify it. All that is expected from the machine, is making a decision which is in compliance with what is considered an ethically correct behaviour in specific situation and the reasoning behind that decision can be traced back to the ethical principle, which it was based on. (S. L. Anderson, 2011: 164)

A machine itself being able to justify its decisions should not be taken too literally. At this point of time it would mean that a system can be put in place which logs the reasoning and actions of the machine and can be later used to see how it reached the decisions it made. A machine would make ethical decisions according to the ethical principle that is implemented in it and the information it has acquired from the world about the current situation. The validity of the machine's decision would come down to quality of the implementation of an ethical principle and how satisfactory the principle is, as well as to how sufficiently the machine is capable of acquiring relevant information from the world. A machine itself is not capable of judging if and how ethical its decision was. In a way, it could be said that machine imitates ethical behaviour, but does not understand what it is doing.

4.1 Who is responsible for the mistakes of a machine?

No matter how well developed and tested a machine is, there always remains a possibility, that it will make a mistake. While consequences of the mistakes of some machines

may be negligible, the mistakes of others such as autonomous vehicles or military robots, can be lethal. Which leads to a question, who should be held responsible? There is no sense in holding a machine responsible for its actions as it does not even have necessary features to comprehend what responsibility is. R. Sparrow (2007: 72) has said that a machine cannot be punished, because it cannot suffer. Only thing that can be done in regard of the machine, is either removing it from circulation or finding the cause(s) of the mistake and improving the machine. This means that the responsibility must lie somewhere else.

A. Matthias (2004) states that there is a responsibility gap between machines and humans. Since autonomous machines can act on their own, they learn and adapt though interaction with environment and other agents. Designers and programmers do not have the same control over the behaviour of the software and machine as they did with regular programs. This results in a gradual process, where designers and programmers of a machine progressively transfer control over to the machine itself. On one hand, creators of the machine do not have sufficient control over it anymore and on the other hand, the machine itself cannot be held accountable for its actions, this is what Matthias calls the responsibility gap. (Matthias, 2004: 181-183).

Sparrows (2007) comes to a similar conclusion. For example, if an autonomous military machine makes a decision which results in a war crime, who is responsible? Given the increase of the autonomy of the system, increases the possibility that it makes a decision which was not predicted by programmers. To hold the programmers or the designers responsible is similar to holding parents responsible for the actions of their adult children who have left home. (Sparrow, 2007: 69-70)

Another option would be to hold responsible a commanding officer that ordered the deployment. If military forces have accepted responsibility for possible cases where the machine makes a wrong decision, then responsibility would lie on the officer. This would mean that these military forces can be held responsible for errors of the autonomous systems over which decisions they actually do not have control. Sparrows concludes here that at one point, the machine's autonomy has grown to the state where it would not be fair anymore to hold the commanding officer responsible. (Sparrow, 2007: 70-71)

Neither of Matthias nor Sparrow offer a solution to how to fill this gap of responsibility, without reverting full responsibility back to humans. Sparrow (2007: 74) says that currently

only way seems to be assigning responsibility to relevant individual, which would still be unfair, because it makes them responsible for the actions they cannot control.

Since humans cannot share the responsibility with the machine and the machine cannot be held responsible for its actions, then eventually the responsibility will lie solely on either machine's designers/programmers/manufacturers, on its owners or on both. If it will be shared responsibility between parties involved or one sole entity that will be held liable, is probably going to be solved in courts, rather than in philosophical discussions.

4.2 Can machines have moral status?

M. A. Warren (2003: 446) proposed list of the principles of moral status, where each principle is based on a certain criteria. I will not expand further on the principles themselves, but will give overview of underlying criteria, because these give unambiguous basis for comparison with the machines.

First, Warren (2003) offers a scale where all living beings have some moral status. Moving further on the scale, sentient creatures have stronger moral status than non-sentient ones. From there on, beings who have moral agency have stronger moral status than those, who do not. In addition she includes social and ecological relationships. Animals that are endangered by human activities and have important role in their ecosystem, have stronger moral status, than compared to the ones that are simply on aforementioned scale. Human infants have stronger moral status than animals because of our social relationship with them. (Warren, 2003: 445-446)

The fact that machines are not alive would not necessarily exclude them from moral status if they were sentient. Self-awareness and moral agency would be further arguments towards moral status. In their current state and at least in near future the machines are not sentient nor self-aware or moral agents. This does not necessarily mean that they can never be but it is not the reality now. Non-sentient machines cannot suffer nor enjoy and cannot perceive harm in a similar way that living, sentient beings can, which means that the machines are not likely to have any moral status (Warren, 2003: 449).

In this context, being able to do ethically acceptable decisions or imitate ethical actions, does not change anything for the machines that are being discussed in this thesis. They would not have moral status regardless of how accurate their decisions would be because they have no sentiency, self-awareness or moral agency.

Conclusions

This thesis has discussed whether autonomous machines are capable of making ethical decisions. Important aspect of ethical decision making is that a machine has an ethical principle implemented, on which it can base its decisions. Top-down, bottom-up and hybrid approaches to implementing ethical principle were discussed.

Top-down approaches are based on some established ethical theory. Utilitarianism, deontological approaches and virtue ethics were covered under this category. Common problem of all top-down approaches is their abstractness which is not comprehensible to a machine. Implementation of this kind of approach needs translating abstract or vague concepts into machine readable details.

Utilitarianism is appealing to machine ethics, because its promise of calculability of ethics. Due to large computational demands and unfeasible requirements of acquiring necessary information makes implementing utilitarianism in its full extent impractical. It should not be entirely disregarded though, as with limited scope it could be used as fall-back mechanism.

Advantages of deontological approaches are rules, which could be turned into an algorithm for a machine. Problematic aspects are possible conflicts that can arise between the rules or within the same rule. This could result in a deadlock in a machine and the machine not making any decision at all.

Virtue ethics does not seem to be appropriate for autonomous machine, since a machine is built for some specific actions. Virtue ethics on the other hand focuses on developing a character, which is not what the machine is created for.

Next were discussed bottom-up approaches, which mean acquiring ethical principles through a simulated artificial life evolution or a machine learning. In case of ALife evolution, it could be possible that in the future when simulations are complex enough, they could give rise to artificial entities with evolving ethics, but at this time it cannot be done yet. With machine learning a machine could learn ethics from large number of example situations. However, a machine is not yet close the way that humans learn and not able to acquire ethical norms on its own. Moreover, machine that is learning completely on its own could possibly be dangerous,

since it could learn wrong things. Currently machine learning is not in state that machine could become ethical only through learning but it can be used as complementary approach to some top-down theory.

Hybrid approaches are the ones that combine top-down and bottom-up approaches. This seems most feasible solution, since the shortcoming of one approach could be compensated by the other one.

Finally, the advantage of domain-specific approach was discussed, which would mean that instead of trying to implement general ethics to a machine, there would be implemented ethics that is specific to a domain that the machine operates in. This would be simpler task and since machines are made with specific purpose they cannot apply general ethics all the way. In the bigger picture, domain-specific ethics could give rise to general ethics through consolidation.

After discussing ethical approaches, the role of internal model was briefly discussed in machine's ethical decision making. Internal model enables the machine to simulate its actions and consequences in environment and in regard of other agents. It enables machine to foresee possible chain of events and find the most suitable action in a current situation, without trying out all the actions.

Eventually, different aspects of machine's ethical decision making were discussed. Since machine does not have consciousness, free will and intentionality, it cannot be held responsible for its actions. It also does not have awareness and self-interest, which means that it does not make ethical decisions in the same sense as humans do. However, with properly implemented ethical principles and adequate information acquisition it could be able to make the decisions that are considered ethically appropriate in a given situation. Since the machine does not have awareness of what it is doing in the sense that a morally responsible human does, then it would mean that it is acting unknowingly.

References

Abney, K. (2011). Robotics, Ethical Theory, and Metaethics: A guide for the Perplexed. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics: the ethical and social implications of robotics* (pp. 35–52). Cambridge, Massachusetts, USA: The MIT Press.

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, *7*(3), 149–155.

Anderson, M., & Anderson, S. L. (2011). A Prima Facie Duty Approach to Machine Ethics: Machine Learning of Features of Ethical Dilemmas, Prima Frima Duties, and Decision Principles through a Dialogue with Ethicists. In M. Anderson & S. L. Anderson (Eds.), *Machines Ethics* (pp. 476–492). New York, USA: Cambridge University Press.

Anderson, S. L. (2011a). Machine Metaethics. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 21–27). New York, USA: Cambridge University Press.

Anderson, S. L. (2011b). Philosophical Concerns with Machine Ethics. In M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (pp. 162–167). New York, USA: Cambridge University Press.

Arkoudas, K., & Bringsjord, S. (2014). Philosophical foundations. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge handbook of Artificial intelligence* (pp. 34–63). Cambridge, United Kingdom: Cambridge University Press.

Bentham, J. (1907). An Introduction to the Principles of Morals and Legislation. *Library of Economics and Liberty.* Retrieved April 18, 2016, from http://www.econlib.org/library/Bentham/bnthPML.html

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford, United Kingdom: Oxford University Press.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 316–334). Cambridge, United Kingdom: Cambridge University Press.

Copeland, B. J. (2016). artificial intelligence (AI). In *Encyclopædia Britannica*. Retrieved April 24, 2016, from http://www.britannica.com/technology/artificial-intelligence

Danks, D. (2014). Learning. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 151–167). Cambridge, United Kingdom: Cambridge University Press.

Gips, J. (2011). Towards the Ethical Robot. In M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (pp. 244–253). New York, USA: Cambridge University Press.

Good, I. J. (1965). Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers*, 6, 31-88

Grau, C. (2011). There Is No "I" in "Robot." In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 451–463). New York, USA: Cambridge University Press.

Guarini, M. (2011). Computational Neural Modeling and the Philosophy of Ethics: Reflections on the Particularism–Generalism Debate. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 316–344). New York, USA: Cambridge University Press.

Harman, G., & Thomson, J. (1996). *Moral Relativism and Moral Objectivity*. Wiley-Blackwell.

Hursthouse, R. (2003, July 18). Virtue Ethics. *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), Edward N. Zalta (ed.) Retrieved March 14, 2016, from http://plato.stanford.edu/archives/fall2013/entries/ethics-virtue/

Kephart, J. O., & Chess, D. M. (2003). The vision of autonomic computing. *Computer*, *36*(1), 41–50.

Marques, H. G., & Holland, O. (2009). Architectures for functional imagination. *Neurocomputing*, *72*(4-6), 743–759.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183.

McDermott, D. (2011). What matters to a Machine? In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 88–114). New York, USA: Cambridge University Press.

Metz, C. (2016). Google's AI wins fifth and final game against Go genius Lee Sedol. In *Wired* Retrieved April 11, 2016, from http://www.wired.com/2016/03/googles-ai-wins-fifth-final-game-go-genius-lee-sedol/

Pennachin, C., & Goertzel, B. (2007). Contemporary Approaches to Artificial General Intelligence. In C. Pennachin & B. Goertzel (Eds.), *Artificial General Intelligence* (pp. 1–30). Springer-Verlag Berlin Heidelberg.

Pontier, M. A., & Hoorn, J. F. (2012). Toward machines that behave ethically better than humans do. *Belgian/Netherlands Artificial Intelligence Conference*, 2198–2203.

Powers, T. M. (2011). Prospects for a Kantian Machine. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 464–475). New York, USA: Cambridge University Press.

Russell, S., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach, 3rd edition* (Upper Sadd). Prentice Hall.

Searle, J. R. (1980). Minds, brains and programs. *The Behavioral and Brain Sciences*, (3), 417–457.

Seville, H., & Field, D. G. (2011). What Can AI Do for Ethics? In M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (pp. 499–511). New York, USA: Cambridge University Press.

Shafer-landau, R. (2009). *The Fundamentals of Ethics*. Oxford University Press.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., … Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, *24*(1), 62–77.

Tonkens, R. (2009). A challenge for machine ethics. *Minds and Machines*, *19*(3), 421–438.

Torrance, S. (2011). Machine Ethics and the idea of a More-Than-Human Moral World. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 115–137). New York, USA: Cambridge University Press.

Wallach, W., & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. New

York, USA: Oxford University Press.

Warren, M. A. (2003). Moral Status. In R. G. Frey & C. H. Wellman (Eds.), *A Companion to Applied Ethics* (pp. 439–450). Blackwell.

Veruggio, G., & Abney, K. (2011). Roboethics: The Applied Ethics for a New Science. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 347–363). Cambridge, Massachusetts, USA: The MIT Press.

Whitby, B. (2011). On Computable Morality: An Examination of Machines as Moral Advisors. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 138–150). New York, USA: Cambridge University Press.

Winfield, A. F. T., Blum, C., & Liu, W. (2014). Towards an ethical robot: Internal models, consequences and ethical action selection. In M. Mistry, A. Leonardis, M. Witkowski, & C. Melhuish (Eds.), *Advances in Autonomous Robotics Systems 15th Annual Conference, TAROS 2014, Birmingham, UK, September 1-3, 2014. Proceedings* (Vol. 8717 LNAI, pp. 85–96). Cham, Switzerland: Springer International Publishing.

Can Autonomous Machines Make Ethical Decisions?

Abstract

This thesis explores whether autonomous machines are capable of making ethical decisions. An autonomous machine here is understood as a machine which operates without direct human intervention and is controlled by artificial intelligence. Author discusses advantages and disadvantages of different ethical approaches which could be used by a machine as basis for its decisions. These approaches cover both classical theories such as utilitarianism and alternative concepts such as building ethics from ground up through machine learning. Domain-specific ethics is suggested as more attainable goal than implementing general ethics. Nature of machine's decision making is covered. Since machines do not have consciousness, free will, intentionality and self-interest, author concludes that they are not capable of making ethical decisions in the same sense as humans are. Nevertheless, with properly implemented ethical principles and ability to acquire information from the world, machine is capable making decisions that are considered ethically appropriate in particular situation.

Kas autonoomsed masinad suudavad teha eetilisi otsuseid?

Resümee

Käesolev bakalaureusetöö küsib, kas autonoomsed masinad suudavad teha eetilisi otsuseid. Autonoomse masina all mõeldakse siin selliseid masinad, mis on võimelised tegutsema ilma pideva inimese poolse juhtimiseta ning mida kontrollib tehisintellekt. Autor arutleb eeliste ja puuduste üle, mis on erinevatel eetilistel printsiipidel, mida masin saaks kasutada alusena oma otsuste tegemisel. Siia on kaasatud nii klassikalised eetikateooriad nagu näiteks utilitarism kui ka alternatiivsed lähenemised nagu juhtumipõhine masinõpe. Samuti leitakse, et valdkonnapõhise eetika masinasse juurutamine oleks kergem, kui püüd korraga arendada üldisel eetikal põhinevat masinat. Vaadeldakse ka üldist masinapoolse otsustusprotsessi olemust ning jõutakse järeldusele, et kuna masinal puudub teadvus, vaba tahe, kavatsuslikkus ja omakasupüüdlikkus, siis ei ole masinad võimelised tegema eetilisi otsuseid sel moel nagu inimene. Sellest hoolimata on masinad võimelised tegema otsuseid, mida saab pidada eetiliseks mingis kindlas situatsioonis, juhul kui neisse on implementeeritud sobiv eetiline printsiip ning masin suudab koguda adekvaatset infot end ümbritseva kohta.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, René Tõnisson,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
„Can Autonomous Machines Make Ethical Decisions?“,

mille juhendaja on Mats Volberg,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **05.05.2016**