ETHICS OF ARTIFICIAL INTELLIGENCE:
MORAL RESPONSIBILITY OF SELF-DRIVING CARS AND SEX ROBOTS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF SOCIAL SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

M. CEM ÖZMEN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
THE DEPARTMENT OF PHILOSOPHY

SEPTEMBER 2019

Approval of the Graduate School of Social Sciences

Assoc. Prof. Dr. Sadettin Kirazcı
Director (Acting)

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Prof. Dr. Ş. Halil Turan
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. M. Hilmi Demir
Supervisor

Examining Committee Members

Assoc. Prof. Dr. A. Fevzi Zambak (METU, PHIL)

Assoc. Prof. Dr. M. Hilmi Demir (METU, PHIL)

Assoc. Prof. Dr. Murat Arıcı (Selçuk Uni., PHIL)

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Mehmet Cem Özmen

Signature :

ABSTRACT

ETHICS OF ARTIFICIAL INTELLIGENCE:
MORAL RESPONSIBILITY OF SELF-DRIVING CARS AND SEX ROBOTS

Özmen, M. Cem
Master, Department of Philosophy
Supervisor: Assoc. Prof. Dr. Mehmet Hilmi Demir

September 2019, 93 pages

This thesis analyzes the ethical impacts of the Artificial Intelligence (AI) applications. AI applications are used in many areas to make daily life more comfortable and efficient. They are used for cleaning houses, taking care of old or sick people, working at dangerous jobs replacing human beings, giving medical advisory, preventing fraudulent situations in finance, etc. Similarly, the usage of sex robots, self-driving cars, translation tools, image and emotion recognition applications, etc. are expected to increase in the near future. Even today, many of AI applications have commenced to be used widely in social and business life. Rapid development of AI technologies, especially after the World War II, led to many social and ethical risks and problems for the world. Any analysis of these challenges needs a philosophical basis for a sound development of these applications and their reliable use in society. In this thesis, the ethical impacts of the Artificial Intelligence applications are investigated by focusing on two specific AI applications: self-driving cars and sex robots. These two AI applications are also analyzed from the perspective of moral responsibility. I conclude my analysis with the following claim:

Since AI applications do not have conscious abilities, free will, and autonomy as in the same sense of the ones seen in the human beings, they cannot be held morally responsible.

Keywords: Moral responsibility, AI application, consciousness, free will, autonomy

ÖZ

YAPAY ZEKA ETİĞİ: SÜRÜCÜSÜZ ARAÇLAR VE SEKS ROBOTLARININ AHLAKİ SORUMLULUĞU

Özmen, M. Cem
Yüksek Lisans, Felsefe Bölümü
Tez Yöneticisi: Doç. Dr. Mehmet Hilmi Demir

Eylül 2019, 93 sayfa

Bu tezde Yapay Zeka (YZ) uygulamalarının ahlaki etkileri incelenmektedir. İnsanların günlük yaşamını daha rahat ve verimli sürdürebilmeleri için birçok alanda YZ uygulamaları kullanılmaktadır. Evlerin temizlenmesi, yaşlı ve hastaların bakımı, tehlikeli işlerde insan yerine çalışma, tıbbi destek verilmesi, finans sektöründe dolandırıcılık olaylarının engellenmesi gibi birçok alanda YZ uygulaması kullanılmaktadır. Benzer şekilde seks robotları, sürücüsüz araçlar, çeviri uygulamaları, görüntü ve duygu algılama uygulamaları vb. gibi birçok uygulamanın kullanımının da yakın zamanda yaygınlaşması beklenmektedir. Birçok YZ uygulaması, daha bugünden sosyal ve iş yaşamında yaygın olarak kullanılmaya başlanmıştır. YZ teknolojilerinin özellikle 2. Dünya Savaşı'ndan sonraki hızlı gelişimi, dünyada birçok sosyal ve ahlaki risk ve probleme de yol açmıştır. Bu uygulamaların sağlıklı bir şekilde gelişimi ve toplumda güvenli bir şekilde kullanımı için söz konusu güçlüklerin felsefi bir zeminde incelenmesi gerekmektedir. Bu tezde iki özel YZ uygulaması olan sürücüsüz araçlar ve seks robotları üzerine yoğunlaşılarak YZ uygulamalarının ahlaki etkileri incelenmektedir. Bunun sonunda analiz, aşağıdaki iddiayla sonuçlandırılmaktadır:

İnsanlarda görüldüğü anlamda bilinçli yetenekleri, özgür iradeleri ve bağımsızlıkları olmadığı için YZ uygulamaları, ahlaki olarak sorumlu tutulamazlar.

Anahtar Kelimeler: Ahlaki sorumluluk, YZ uygulaması, bilinç, özgür irade, bağımsızlık

“The true enemy of good isn't evil, but fear. Evil will battle good, but fear will corrupt it.”

(quoted by Jim C. Hines)

To the people who see ethics as the reference of their life, and pay the price for keeping morality alive for the future of world,

ACKNOWLEDGMENTS

The author wishes to express his deepest gratitude to his supervisor Assoc. Prof. Dr. Mehmet Hilmi Demir, for his guidance, advice, criticism, encouragements and insight throughout the research.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

In this thesis, the ethical impacts of the Artificial Intelligence (AI) applications will be investigated. I will also examine two specific sample AI applications, i.e., self-driving cars and sex robots from the perspective of moral responsibility. Autonomy, consciousness and moral status features of the AI applications will be taken as the basis for the discussion. While doing this Aristotelian ethics, Kantian ethics, Utilitarian ethics and other general ethics approaches will be taken as reference in order to assess the moral status of AI applications. Machine ethics and robot rights issues will also be overviewed within the discussion.

Artificial intelligence can be broadly defined as the intelligence formed by software programs in machines, which can be seen as similar to the natural intelligence shown in humans and other animals. In practice, intelligence is attributed to artificial entities when those entities simulate human-like functions such as “learning” and “problem solving” features. AI applications could be classified based on their functionality as weak (i.e. undertaking only limited tasks) or strong (i.e. fully imitating human beings). However, when AI is mentioned, it is generally used in the meaning of weak AI, for today, since no AI application that imitates human beings with full functionality has been manufactured, yet. Therefore, unless stated otherwise, I will use it as in the current meaning, which mentions the weak (narrow) AI, in the rest of the thesis.

Artificial intelligence applications are used in many areas to make everyday life more convenient and efficient. They are used to clean houses, take care of old or sick people, to work at dangerous jobs replacing human beings, to give medical advisory and healthcare services, to prevent fraudulent situations in finance, etc.

Similarly, the usage of sex robots, self-driving cars, translation tools, image and emotion recognition applications, etc. are expected to increase in the near future. Even today, many of AI applications have commenced to be used widely in social and business life.

Use of artificial intelligence has increased in recent years. Development of Internet technologies and increased innovation led AI to be used widely in daily life. These advances, in addition to the existing socioeconomic and ethical impacts of technology, brought AI to the focus of several new discussions. In addition to the existing Internet-based risks (i.e. data security, misinformation, isolation, etc.), AI-specific problems are expected to be faced with in the near future. In general, the main challenges and issues regarding the AI usage can be listed as following: AI moral responsibility and robot rights; unintended consequences; human-AI interaction; safety and security of AI applications; inequality and unemployment rising and singularity (control of a complex intelligent system).

As the use of AI applications are rapidly increasing, all societies are trying to understand and provide solutions for these emerging issues, as early as possible. Discussions among the related parties, particularly ethicists and philosophers, a number of declarations signed by famous scientists, measures taken by some governments are initial steps for this goal.

Despite such risks, AI applications also bring some positive outcomes to human beings, especially because of technologies involved in AI. These are: automation, machine learning (operating the computer systems without programming), machine vision (capturing the visual data and analyzing), Natural Language Processing (NLP - processing of human language by a computer program), robotics (design and manufacturing of robots), self-driving cars (automated piloting a vehicle).

Although AI applications are started to be used in many critical functions, two of them came to the forefront in this process, due to their impacts and risks. These are the self-driving cars and sex robots.

One of the AI applications, which I will focus in this thesis, is the self-driving cars that are being used in several countries. Many companies including BMW, Mercedes, Ford, GM, Toyota, Nissan, Volvo, Audi and Tesla have started to make these cars. Interestingly, it can be seen that some of the technology companies (i.e. Google, Apple, Uber, etc.) are also interested in production of such vehicles, although they are not traditionally related with car production.

The development of self-driving cars leads to an increase in security related concerns, in addition to other ethical concerns. For example, a self-driving car might be involved in an accident, however the responsibility will not be clear. Similarly, it is not always clear that these cars will do the best action in case of conflict situation (i.e. Trolley Problem[1]) in order to minimize the damage given.

The second AI application that I will focus on is the sex robots that are available in the market. Sex robots are generally defined as the realistic dolls, which closely mimic women or men, for sexual relationship. As of 2018, various models of these robots have been produced. Similarly, their specifications can now be determined or ordered by customers based on various features such as they can smile, blink eyes, talk, make joke, even have an ability to feel orgasm, etc. Experts say such customized robots will start to appear in many houses in the next decade, especially for the lonely people looking for love and sex.

[1] Trolley Problem was first introduced by Thomson, Judith in her book *Killing, Letting Die and the Trolley Problem*, in 1976. According to the scenario (p.206), "Edward is the driver of a trolley, whose brakes have just failed. On the track ahead of him are five people; the banks are so steep that they will not be able to get off the track in time. The track has a spur leading off to the right, and Edward can turn the trolley onto it. Unfortunately, there is one person on the right-hand track. Edward can turn the trolley, killing the one; or he can refrain from turning the trolley, killing the five. If what people who say 'Killing is worse than letting die' mean by it is true, how is it that Edward may choose to turn that trolley?"

Despite to the expectations from sex robots, there exist a number of issues and questions that have been arisen with the production of these robots. One of them is that whether the use of sex robots is different than masturbation legally, morally and ethically. Similarly, what would be the impacts of sex robots to social institutions (marriage, partnership, etc.) and whether the use of such robots would affect the behavioral patterns in human to human relations, especially of men towards women and children? Another point is that what would be the status of robots and would they have any rights like human beings or animals. Would these robots increase the inequality in the society or could they be used for the treatment of some social and psychological problems in the society?

As it is seen, any analysis of these challenges needs a philosophical basis for a sound development of these applications and their reliable use in society. In this context, the impacts of AI should be discussed from the accountability and responsibility points of views.

Given this content, the thesis is organized as follows. In Chapter 2, I will provide a general information regarding artificial intelligence. I will touch on the definitions of AI and give some information about some AI applications. I will also provide information about the concepts of intelligence, learning, reason, problem solving, perception, language. I will finish the chapter with a brief history, idea and philosophical background, types, methods, goals and approaches regarding AI.

In Chapter 3, I will provide a general look about the ethical issues and problems, which come into picture with the development of AI applications. I will summarize these risks and problems under the topics of AI moral responsibility and robot rights; unintended consequences; human-AI interaction; safety and security of AI applications; inequality and unemployment rising and singularity.

In Chapter 4, I will first provide a general information on ethical theories. I will briefly mention about the concepts of meta-ethics, normative ethics and applied ethics.

I will also touch on the recurrent themes in the ethics of technology and try to define the moral agency and moral responsibility concepts. Following this, I will investigate self-driving vehicles and sex robots as two major AI applications. Later, I will bring the question of whether artificial intelligence applications, especially for the particular two examples should be held responsible due to their activities, or not. I will discuss the issue and provide an answer based on the notions of autonomy, responsibility, consciousness and moral status. While doing this, I will use Aristotelian ethics, Kantian ethics, Utilitarian ethics and other general ethics approaches to arrive at a conclusion. I will also overview the machine ethics and robots' rights issues within this discussion.

In the final chapter, I will summarize the discussions of previous chapters in order to reach a conclusion about the moral status of robots. The conclusion I reach is that AI applications, including self-driving cars and sex robots, cannot be held responsible for activities they perform, at least for the time being. But it is my expectation that they could be seen as morally responsible agents, in the future not so far away.

CHAPTER 2

A QUICK SURVEY OF ARTIFICIAL INTELLIGENCE

In this chapter, a brief information will be provided regarding the notion of Artificial Intelligence (AI) and applications. First, a definition will be given for AI from various sources. Then main AI application categories will be revisited and major usage areas will be overviewed. A quick information will be provided about the concepts of intelligence, learning, reason, problem solving, perception and language, due to their relevance for AI. Additionally, a brief history of the AI applications will be given to visualize the development of the AI world. The chapter will be completed by the types, methods and goals of the AI applications.

2.1 What is artificial intelligence?

In this section, I would like to define artificial intelligence. This requires to understand what the concept of intelligence is. Therefore, I will start with the definition of intelligence and mention about the main drivers of concept.

2.1.1 Intelligence

Artificial intelligence is generally used in the sense of intellectual processes of humans, like ability to understand reason, learn from past experience and identify meanings.

However, there exists an interesting question related with the concept ‘intelligence’: Although generally all of human behaviours are accepted as intelligence, why, is even most complicated animal behaviour never taken as an indication of intelligence? What is the difference?

Despite most of the human behaviours are accepted as intelligent, almost all of the animal activities are accepted as performed through their instinct. The reason generally comes from the definition, which intelligent is related with the ability to adapt to new circumstances.

On the other hand, human intelligence is not generally defined by only one behaviour, instead, it depends on combination of many different abilities that can be listed as learning, reasoning, problem solving, perception, and using language.

One of the main tools used in AI applications is learning. Various forms of learning types can be used in artificial intelligence. One of them is trial and error. For example, a game playing program tries several actions randomly until the goal is achieved. Since it can perform these operations very rapidly, it can play the game more successfully than individuals. Similarly, AI applications can memorize many actions, since they can store them in their memory. However, it is not easy for them to make generalizations (i.e. using past experiences in new situations). For example, an AI speaking English, cannot guess past tense of a word, if it did not encounter it before.

Another tool used in AI applications is reasoning. Reasons are generated through the results, which are classified as either deductive or inductive. An example of deductive sentence is, “He should be either at school or at home. He is not at school; therefore, he is at home,”, whereas the inductive one: “Previous failure was due to lack of adequate skills, thus this failure was occurred since he has not adequate skills to succeed it.”

Additionally, in AI, problem solving, can be defined as a systematic search within a set of possible actions, to achieve a number of predefined solutions. It can be a special-purpose method, which is tailor-made for a particular problem, or a general-purpose method is used for a wide variety of problems.

One other tool used in AI applications is perception, which is made generally through scanning environment by the aid of various sensory organs, and then separating these objects to analyze them. Artificial perception is used at optical sensors to identify individuals, autonomous vehicles to drive at different speeds based on availability of roads, and robots to walk around within buildings to collect object like waste boxes.

Similarly, today, computer programs are used to respond in a human language to questions and statements. In fact, these programs do not actually understand language, however they can, come to the point where their command of a language is as same with a normal human. Then, it can be asked that what real source of understanding is, if even a computer does no need to understand to talk? This is still a discussion point among the scientists and philosophers. Most probable approach is that understanding does not depend on behaviours, it also is related with history of human beings. It means, in order to understand, language has to be learned and talked within a society, which AI applications do not have.

2.1.2 Definitions

Following the definition of intelligence, we can come to the main course, which is the artificial intelligence.

A broad definition of Artificial Intelligence can be given as the intelligence generated by software programs in machines, which can be seen as similar to natural intelligence shown in humans and other animals. Practically it can be said as the term “artificial intelligence” is used when a machine simulates human-like functions such as “learning” and “problem solving” features.

However, people, generally are not all understanding the same things from definition of this term, and despite foundation is generally the same, focus of artificial intelligence might depend on entity, which provides the definition.

The term “Artificial Intelligence” has been used first by John McCarthy in 1956, during the conference called “Dartmouth Summer Research Project on Artificial Intelligence”, which has been organized to discuss what would ultimately become the field of AI, by a group of researchers from a variety of disciplines including language simulation, neuron nets, complexity theory and more. It was stated in the proposal for the conference that: “The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.”

Today, artificial intelligence is defined in various forms within different dictionaries. For instance, AI is defined in The English Oxford Living Dictionary as: “The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.”

On the other hand, it is defined in Merriam-Webster as:

i. “A branch of computer science dealing with the simulation of intelligent behavior in computers.

ii. The capability of a machine to imitate intelligent human behavior.”

Similarly, The Encyclopedia Britannica states AI as: “artificial intelligence, the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings.”, whereas intelligent is defined as “being able to adapt to changing environments”.

Besides these formally stated definitions, it is observed in recent years that investments are focusing on the following areas, which can constitute the goals of the AI:

i. Creating systems, which think exactly like humans (i.e. “strong AI”)
ii. Creating systems to work for a specific area, without focusing on how human reasoning works (i.e. “weak AI”)
iii. Creating systems to use human reasoning as a model, without necessarily taking as an ultimate goal.

Majority of AI developments happening today fall under the third objective and uses human reasoning as a guide to provide better services, instead of trying to produce a perfect imitation of the human mind.

On the other hand, discussions of artificial intelligence (AI) have created different approaches, including a fear based on expectation, which it will change from being a benefit for society to a dangerous entity for the world. Even great scientists Stephen Hawking (as a response to a question in Ask Me Anything session on a social discussion platform, in 2015) and Elon Musk (spoke at the International Space Station Research and Development Conference in U.S., July 19, 2017) have shared their concerns about the threats of AI.

2.2 Applications

The applications also shed light on what AI is. New applications and systems of AI are being developed day by day. Robots, chess playing programs, Netflix, Spotify, and ‘Siri’, etc. use artificial intelligence during their operations. Besides these, many other applications use AI. The most known AI applications can be grouped as follows:

i. AI in manufacturing and robots:

A wide range of AI implementations are used in manufacturing industry. Especially, they are used in car industry for enhancing the driving features (i.e. self-driving cars, driver assistant, etc.), cloud services (i.e. predictive maintenance), car manufacturing and monitoring. Similarly, they are used in the form of robots for various purposes, particularly for carrying out tasks, specified for them. Today, most of AI-powered robots, do not have general intelligence, however they are capable of solving problems and “thinking” in a limited capacity. One specific class of these robots is sex robots, which we I focus on deeply, in the following chapters.

ii. Speech, image and emotion recognition

One of the well-known examples for speech recognition is *Siri*, which can understand human speech. Many speech recognition systems are used at voice-response interactive systems and mobile applications.

Similarly, image recognition is the process of identifying and detecting an object or feature in a digital image or video. Image recognition technology can also be used to diagnose diseases, analyze clients and verify users based on their face.

One other AI recognition technology is used to read emotions, to capture body language cues, and vocal character, which shows the feelings of a person. It is generally used in gaming, automotive, robotics, education and healthcare industries.

iii. Deep learning, biometrics and Natural Language Programming (NLP)

Deep learning platforms use a specific form of Machine Learning (ML) that involves artificial neural circuits with various abstraction layers, which can mimic human brain, processing data and creating patterns for decision making.

Biometrics is used for natural interactions between humans and machines, including interactions related to touch, image, speech, and body language recognition.

NLP technology uses text analytics to understand the structure of sentences, as well as their meaning and intention, through statistical methods and ML. They are mostly used at security systems and fraud detection.

iv. Cyber security, finance and health

Cyber defense AI focuses on preventing, detecting and providing timely responses to attacks or threats to infrastructure and information security.

The following example can be given for the use of AI in finance investments. In 2009, 50 un-known companies have been chosen by AI, as the most successful startups. Almost eight years later, it was seen that many of them (i.e. Evernote, Spotify, Etsy, Zynga, Palantir, Cloudera, etc.) have succeeded to grow and earn money. Almost 20% of the companies chosen by the computer were valued at a billion dollars.

Similarly, AI applications are using deep learning and image recognition technologies to diagnose possible signs of diseases.

v. Other AI applications

AI applications can also be used for a wide range of purposes. Another interesting example is the fortune teller. It can successfully predict if someone is gay or straight based on photographs of their face. A Stanford University study (the research on extracted features from the images using “deep neural networks”, was conducted by Michal Kosinski and Yilun Wang in 2017) discovered that an AI application can correctly identify who are gay and who are not, with 81% accuracy for men, and 74% accuracy for women.

Additionally, Facebook is using AI that can save the lives of people at risk of committing suicide. The company performed a project in the US to proactively identify Facebook messages that can indicate suicidal tendencies. The application identified over 100 cases for the people needed help.

Following the brief review of AI applications, now, I will provide some information regarding the history of AI applications.

2.3 A brief history

As Haugeland (1989, p. 15) mentioned, intellectual roots of AI, and the concept of intelligent machines might be found in Greek mythology. Similarly, intelligent objects have been encountered in literature, and some of the historical mechanical devices seem to have some degree of intelligence. Since the logic and the symbolic reasoning have developed by time, then machines emulating human intelligence started to be made.

Later, the first computer (i.e. Analytical Engine) was designed in 19th century by Charles Babbage (however, it was not built until 1991). With ongoing progress of technology from early 20th century, various models of computers, and also theoretical concepts were created. Following the development of modern computers (particularly after 1950), it became possible to develop programs to perform difficult intellectual tasks. By using these programs, many kinds of systems and applications started to be made, which are used by wide variety of society, now.

Some of these computational milestones regarding the development of AI are listed below. This information has been organized and summarized, based on the books and articles, from Haugeland, 1989; Thomas, 1999; Sloman, 1978; Nilsson, 2015; Crockett, 1994; Buchanan, 2005. The details might be found in the references.

2.3.1 Ancient history

Stories about the objects, which can be said as the very primitive artificial intelligence tools, go back as far as the ancient Greeks. Hephaestus from Olympus, who was also the creator of Pandora, the first woman, created the lifelike metal automatons, which have been accepted as the idea of intelligent robots.

2.3.2 Before 20th century

Al-Jazari, designed various mechanical tools (i.e. musicians powered by water flow, elephant water clock, key lock mechanism, etc.) in 13th century. These tools are believed to be the first programmable humanoid robots and the first encryption mechanisms. Similarly, in 16th century, Leonardo Da Vinci designed a humanoid robot in a form of a medieval knight, which is able to sit up, move its arms and head, and open its chin.

On the other hand, Descartes proposed that bodies of animals can be simply thought as complex machines in 17th century. Therefore, he mentioned about the possibility that machines would one day think and make decisions. Despite he claimed that they would never be able to talk like humans; he thinks that they might one day learn about performing one specific task. Similarly, he claimed that some of them might be able to adapt to any task to do. In fact, these arguments led to the development of specialized and general AI concepts.

Similarly, Thomas Hobbes (1996, p. 7), who was seen as the ‘Grandfather of AI’, said that it might be possible to build an “artificial animal”. Additionally, he defined thinking as basically a sum of physical processes. This led to the basis for the robots, androids, and various types of artificial intelligence thinking.

A great development on the conceptual basis for robots has been made by the story of Frankenstein's monster, which has been published by Mary Shelley (1818).

In this story, Frankenstein builds an artificial, intelligent android from some materials, which introduced the idea of a robot. Additionally, it introduced the word ‘robot’.

On the practical side, mathematician-philosopher Charles Babbage invented a programmable mechanical calculating machine, which performs functions and calculates values automatically. It led to the basis for automated computation. Babbage could not finish the construction of his device; however, a machine was built similar to his design in 1991. Similarly, Hisashige Tanaka from Japan, made mechanical toys, which serves tea and paints Japanese kanji characters, which can be accepted as early examples of robots, in 18th century.

2.3.3 20th century

Alan Turing made the Turing Machine, which is a device consisting of a tape, an infinite line of cells, and a head, an active element that moves along it, to simulate logic and test theoretical ideas about potential of computers. Therefore, Alan Turing is seen as the father of both computer science and artificial intelligence. He (1950, pp. 433-460) proposed Turing Machine model to discuss the theoretical possibilities of what can be computed, and then he designed Turing Test. The objective of the test was to identify whether a machine can convince a participant that it was indeed a human being. In order to pass the test, a computer was to be able to make small talk with a human being and show understanding of given context. Although it seems easy, realization of such results proved to be extremely difficult and, up to this date, unachievable.

Following the increase in development of robots, Isaac Asimov published his three laws of robotics (1950), to share the themes for human-robot interactions.

Dartmouth College professor, John McCarthy introduced the term “artificial intelligence” as the topic of the Dartmouth Conference, the first conference devoted to this subject, in 1956.

Many fields, which are fundamental for AI, including natural language processing, computer vision, and neural networks, have been discussed during the conference.

An interactive program, ELIZA, has been developed at MIT by Joseph Weizenbaum in 1958. It is accepted as the world’s first chatbot and an early model of the applications like Alexa and Siri. ELIZA can be seen as an early implementation of natural language processing, which aims to teach computers to communicate with humans in human language. Despite it could not talk, instead it was communicated through text, it opened a way for later efforts to communicate people with machines.

The first full-scale intelligent robots (WABOT-1, or WAseda robOT), has been made at Waseda University, in Tokyo, in 1972. It could walk, hold objects, speak Japanese, listen, and measure the distances to some objects, by the aid of its vision and auditory senses. Similarly, first computer-controlled, prototype autonomous vehicle (i.e. The Stanford Cart) has been produced by Hans Moravec in 1979. It could successfully move in a room full of chairs and Stanford AI Lab.

John Searle produced the Chinese room argument in 1980, which opposed to the idea that a computer can be programmed with the appropriate functions to behave the same way a human mind would.

First autonomous car has been produced at Carnegie Mellon University in 1986. Five racks of computer hardware, video hardware, a GPS receiver, and a Warp supercomputer have been used for this vehicle. It reached a top speed of almost 32 km/h. Additionally, a race has been organized among autonomous vehicles in the Mojave Desert. Five vehicles made their way around, with a team from Stanford University taking the prize for the fastest time. By 2007, a simulated traffic environment has been constructed for the vehicles to navigate, thus they had to be able to deal with traffic regulations and other moving vehicles.

The Deep Blue chess program won the game against the world chess champion, Garry Kasparov. Additionally, first autonomous robotics system, Sojourner, was deployed on surface of Mars, by NASA.

Various types of sex robots started to be commercially available in the market by 2017. Similarly, self-driving vehicles are started to be used on roads by 2018. The first commercial autonomous vehicle hire service, Waymo One is currently in use by 400 members of the public who pay to be driven to their schools and workplaces within a 100 square mile area.

On the other hand, risks and dangers of AI have been commenced to be discussed. Philosopher and cognitive scientist Daniel Dennett (2017, p.402) warned of the dangers of artificial intelligence. Similarly, Bostrom (2014, p.126) mentioned about potential issues regarding superintelligence.

Following this historical information, I will provide a brief summary about the types, methods and tools used in AI world. (This information has been organized and summarized here based on the books and articles, from Haugeland, 1989; Thomas, 1999; Sloman, 1978; Nilsson, 2015; Crockett, 1994; Buchanan, 2005. The details might be found in the references.)

2.4 Types, methods and tools

AI applications could be classified by their types as weak or strong. Weak (or Narrow) AI applications are focused on one narrow task (for example, an AI application playing chess), instead of simulating all features of human beings. However, it is generally accepted that each and every weak AI application contributes to building of strong AI.

On the other hand, strong artificial intelligence applications, in general, are machines, which can think and perform activities on its own, like a human being.

For today, no such kind of AI exists. However, since some industry leader companies are very keen on getting close to build a strong AI, it is expected that it might be available in a rapid progress. The ultimate aim of strong AI is to produce a machine whose overall intellectual ability is totally same (or more) that of a human being.

As it was stated by Clark (1997, pp.19-22), two main methods are used in AI researches: Symbolic (or “top-down”) approach, and connectionist (or “bottom-up”) approach. Top-down approach tries to replicate intelligence by using symbols (i.e. symbolic label), independent from brain. On the other hand, in bottom-up approach, artificial neural networks are designed to imitate brain’s structure (i.e. connectionist label).

For example, to build a system that recognizes letters of alphabet in a computer program developed due to top-down approach, it compares every letter with some geometric descriptions. On the other hand, in a bottom-up approach, an artificial neural network is trained by showing letters to it one by one. Therefore, it can be simply said that, neural activities can be seen the basis of the bottom-up approach, whereas symbolic descriptions are the basis of the top-down approach.

For both of the above-mentioned types of AI applications, ultimate goal could be defined as to build an intelligent machine, which is capable of reasoning, planning, solving problems, thinking abstractly, comprehending complex ideas, learning from experience, etc. In order to achieve this goal, a number of concepts are required to be implemented during the development of AI applications.

At beginning stages, reasoning process was through human processes, which have been imitated in solving puzzles or logical deductions. However, it was not effective enough due to computer resources limitation. Therefore, various methods have been developed for AI applications:

i. Neural networks (simulation of human brain: computing values from inputs; machine learning; pattern recognition; adaptive nature)
ii. Embodied agents (to interact with environment)
iii. Sensorimotor skills (to perceive environment through sensors)
iv. Statistical approaches (digital approaches to specific problems)

Furthermore, a number of technologies are used during the development of AI applications. One of them is machine learning that is creation and use of programs, which allow AI systems to make predictions and decisions based on data input. It is a method where the aim is defined and stages to reach that goal is learned by the machine itself by training (i.e. gaining experience).

Another tool is Natural Language Processing (NLP), which is generally defined as automatic manipulation of natural language, like speech and text, by software. Natural language processing and generation are one of the central issues, which AI field of study deals with.

Similarly, machine perception is related with capability of input interpretation, which is a similar process of human perception through senses. Major components of machine perception are vision (i.e. image collection), hearing (i.e. audio data) and touch (i.e. process surface properties).

Another technology is the robotics that is a field of engineering, which is focused on design and manufacturing of robots. These tools are generally used to perform tasks that are difficult for humans to perform. Main dimensions of robotics are object manipulation, navigation, localization, mapping, and motion planning.

CHAPTER 3

ETHICAL ISSUES in ARTIFICIAL INTELLIGENCE

Artificial intelligence applications are used in many areas, to make daily life easy and comfortable, mainly for personal assistance, email filtering, fraud prevention, engineering, marketing models, digital distribution, voice recognition, facial recognition, content management, video production, news generation, playing games, customer service, financial reporting, etc. Additionally, robots, self-driving cars, translation tools, image and emotion recognition applications, etc. are expected to increase in the near future. It can be said that AI will be a part of human life, more than it is today.

However, a number of ethical issues are started to be faced with during the development of these applications. As an example, self-driving cars have already travelled several millions of miles based on the autonomous decisions made by them. These decisions have moral and social impacts, especially due to the potentially significant harms. For instance, a Tesla car had an accident in May 2016, and the passenger was killed. This was recorded as an accident, which the first person was killed in an accident with the involvement of self-driving car. At this point, the question comes into the picture: How can it be ensured that the decisions given by these cars will be ethical? Similarly, the same question might be asked for many other AI systems including robots, weapons, healthcare applications, manufacturing devices, etc.

In this chapter, ethical impacts will be discussed for computer and technological ethics, particularly the ethics of artificial intelligence (the terms ‘ethical’ and ‘moral’ will be used in similar meanings during the thesis). The major ethical issues can be stated under a list, which consist of AI moral responsibility and robot rights;

unintended consequences; Human-AI interaction; safety and security of AI applications; unemployment and inequality rising and singularity (control of a complex intelligent system).

Details of these topics are presented in the following sections.

3.1. AI moral responsibility and robot rights

In addition to the accident mentioned above, an Uber self-driving car also had an accident in March 2018, and led to the death of a lady, in March 2018.

These are the real examples, which have been encountered up to now. Similarly, the classical problem in philosophy, so-called Trolley Problem, can be adapted for the self-driving cars. According to this analogy, if the car cannot brake in time and has to choose one of the alternatives, which are going within its way and hitting a pedestrian, or swerving into oncoming traffic in an opposite lane to cause to another big accident, or swerving into the side of the way and killing the passenger.

These examples show that this kind of accidents might be expected to increase with development of these vehicles. Also, they showed that these devices might think differently from human beings, no matter how smart they are. Additionally, the question has been arisen as whether these cars, as AI applications, should be held responsible, since they caused to death, or not? If yes, what will be ethical, legal and social status of them? If not, who will be the responsible party? Therefore, it can be said that with the increasing number of events and potential issues, ethical guidance seems to be needed for these devices, which take decisions based on their own logic.

A similar discussion is in place regarding the status of the robots, which have been already took place in social life with a large variety of roles. It is expected that, they will be in houses to clean them, to play with children, to become sex partners, to be a judge in court, etc. In short, they will be part of human life more than today and ever been in history.

Considering that they will take such important parts in society life, a number of questions regarding their status and roles should be identified. The main questions are whether they should be accepted as ‘legal’ entities or not, what status should be assigned to them from a citizen, worker, civil rights or criminal law perspective?, can they be thought as similar to animals and be treated like them?, who will be responsible if a robot causes harm to a partner or customer?, how will a robot be punished?, or in general, will they be seen as only ‘lifeless technological beings’ or ‘things’?

As it is seen both in self-driving car and robot examples, there exists an issue regarding the moral responsibility and rights of these applications. Therefore, these issues should be timely identified and necessary measures should have been taken into account during the development of these applications.

On this basis, a discussion has already started on it, and seems to go on increasingly, in the future. For instance, Saudi Arabia government granted the robot “Sophia” full citizenship in 2017. Similarly, in the EU Parliament report (2016, pp. 11-12, rapporteur: Delvaux), a special ‘electronic personhood’ has been granted to AI robots, just as the companies or organizations can have a ‘corporate personhood’. So, as a summary, it can be said that the way is opened for AI applications to become morally responsible parties and to use at least some of the rights owned by people, however the issue still stands to be solved.

3.2. Unintended consequences

Despite the fact that AI applications are designed by human beings, it is not a surprise that some of these applications might lead to unforeseen consequences, due to various reasons, which might give harm to people and nature. For example, a super-intelligent AI could be built accidentally by a program with or without intent, criminal sentencing algorithms, with racist biases of the data stored within it, might be used, to lead to unfair decisions, etc.

In a report published by 26 authors from 14 institutions, including academia, civil society, and industry, in 2018, which has a title *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, it is focused on some key ways, which AI is going to generate new threats for both digital and physical security systems.

For example, with a phishing program, individuals can be sent messages specially designed to fool them into giving up their security credentials. AI might be used to automate many of these kinds of messages, accessing to the social and professional networks to steal confidential data like passwords, financial information, etc.

Although it seems that these kinds of attacks are complicated, if the required software is developed, then it can be used again and again, in a very rapid manner, without any extra cost. As an example, private e-mail messages from Hillary Clinton's campaign chairman John Podesta have been hacked and disclosed with the aid of these applications, during the U.S. elections. Similarly, AI could be used to not only generate emails and text messages, but also fake audio and video data. These applications can be used to mimic someone's voice, and can be used for fraudulent purposes.

Similarly, AI applications might also be misused against regional, or global security. For example, weaponizing drones have been started to be used in many areas.

The last example was that a Russian airplane in Syria was attacked by 10 small home-made drone bombs, in January 2019.

Engineers and AI researchers were aware of the situation, from the earliest stages of development, that technology can be misused in some manner. Therefore, several ethical discussions have been recognized by many scientists working in AI and such related fields as robotics. Scientists, developers, and industry leaders have issued open letters to enforce the governments to address these concerns related to autonomous weapons to increase the awareness about the risks and unexpected consequences of AI.

The HAL 9000 computer, made by science fiction author Arthur C. Clarke, which was brought the movie by Stanley Kubrick in “2001: A Space Odyssey,” is a good example of a system that fails because of unintended consequences. In many complex systems (i.e. RMS Titanic, NASA space shuttle, Chernobyl nuclear power plant, etc.), engineers make designs by bringing many different components together. Therefore, they may know very well that each element can operate in a successful manner individually, however they cannot generally ensure that all these pieces will operate together successfully, or not. Similar issue exists for the AI applications, particularly the complex and complicate ones like self-driving cars or robots.

Another big issue is the AI mistakes that might be encountered during the use of these systems. AI applications generally rely on pre-defined algorithms and a large amount of data during the decision-making process. For soundness of decisions made, well-defined inputs and outputs should be set. Similarly, goals and metrics should also be clearly defined for the system. Basically, the environment should be occurred as unambiguous and predictable as it can be.

However, there exist no such ideal conditions in real world. It cannot be expected from all drivers to follow traffic rules without any exception. Or it cannot be seen that all human handwritings are tidy and correct.

The human brain is generally operated in an uncertain, continuously-changing environment, and generally causes to uncertainty and ambiguity by itself. And therefore, to be able to act successfully in this world, AI should also learn to think more like a human. However, due to unambiguity and uncertainty of cases, AI applications might make various mistakes, which is also another issue for these systems.

Another unexpected result might stem from the nature of the data used in AI applications. Most of AI applications are designed based on Machine Learning (ML) technology, which generally uses historical data for decisions to be made. ML is dependent on the quality of learning data sets (i.e. size, structure, collection methodology, source, etc.). Thus, it can be said that as the data used in AI gets objective, accurate, and large, the possibility of bias situations in its decisions gets decreases.

On the other hand, it can be said that if the data used by AI systems is subjective and/or distorted, then the decisions made by AI can be biased and judgmental. Such an example has been encountered in the Google photos, which classified the two African-American young guys as Gorillas, through automatic image labelling method. Another example was TayTweets, which was a Twitter application deployed by Microsoft in 2016, to be used for through casual conversations. However, following the introduction of this application, it started to tweet misogynistic and racist messages in less than 24 hours, and therefore the application has been stopped, immediately.

For self-driving car case, it was identified that these cars have difficulty in identification of dark skin people in the traffic, since the data used for training autonomous cars generally consists of only light-skinned individuals, as examples. As it seen in all these examples, when large data sets are used in AI applications, distortion of the objectivity of data and therefore bias situations might be encountered, if necessary measures are not taken.

As a summary, it can be said that AI technology generally looks to future for new developments. However, since mostly AI applications are fed by the past data, sound and successful operation of these systems heavily depend on accuracy and objectivity of these data sets, in order to prevent bias and racism problems, causing to an important ethical AI issue for the society.

To conclude, it might be useful to remind that the German scientist, Weizenbaum (1976, pp. 226-227) argued that AI technology should not be used instead of human responsibilities, in positions that require respect and care, such as a therapist, a nursemaid, customer service agent, a soldier, or a judge, since their unexpected consequences might lead to enormous dramatic results.

3.3. Human-AI interaction

Artificially intelligent applications are getting better and better at simulating human conversation and relationships, day by day. An application (Eugene Goostman) won Turing Challenge for the first time in 2015, by succeeding to make more than half of the human participants, to think, as if they had been talking to a person.

Today, when someone enters a store, he/she could be welcomed by a robot. When the same person goes back after a few days, that robot might identify her/him and even remember what he/she had bought previously. In that case, how could its presence influence his/her behaviors or thoughts? Would it be surprised if a robot smiled at him/her? Would he/she feel sad if machine does not exist at its place? Is there a possibility to become friend for the people with robots?

It is easily seen that people will frequently interact with these kinds of AI applications as if they are humans, in many areas like customer services, individual relationships or sales. Although there might exist some problems at these early stages, artificial intelligence applications, particularly robots, will have chance to build relationships with human beings. Therefore, it can be claimed that many actual and potential impacts of these types of relationships will be seen on human behaviors.

In recent years, several studies have been performed to investigate the impact of human-robot interaction on human behaviors. Researches indicated that humans can be influenced by presence of the robots, in the same way as if they could be by presence of another human. As an example, there was an experiment performed at Yale. According to it, a small group of people joined to a study with robots to lay railroad tracks in a virtual world. Each group consisted of three individuals and one robot, sit around a table and worked on tablets. The robot was programmed to make some errors, and to communicate these errors as: "Sorry, folks, I made the mistake. I know it is hard to believe, however as it is seen, robots also make mistakes too."

It was observed that the warm attitude of this robot led to the groups, to have an improved communication among the humans. It was seen that they became more relaxed and conversational, as well as tolerating the group members who has an error, and laughing together more often, leading to have a better collaboration, compared with the groups, whose robot made only strictly correct statements.

On the other hand, more real examples have been encountered regarding how an AI can affect human relations, in recent years. It was seen that trolling and malicious accounts have been created to regularly retweet to other, ordinary accounts, particularly to affect conservative users, during the 2016 U.S. president elections. Despite they do not know each other, these applications affected the people, due to humans' cooperative nature and interest to other people behaviours. As it is known, this led to polarize society in the country.

Another interesting example is the sex robots. Kathleen Richardson, a researcher at De Montfort University, worries about the negative impacts of sex robots. As the director of the Campaign Against Sex Robots, she warns that they will be dehumanizing and could lead users to refrain from real relationships.

On the other hand, there exist views to claim that robots might radically improve relationships between human beings. Levy (2007, p. 22) considers the positive implications of “romantically attractive and sexually desirable robots”. He thinks that some people will come to prefer robot partners to human ones. One of the main reasons he reminded is that, with sex robots, sexually transmitted diseases or unwanted pregnancies will not be problem for the people. Similarly, he claims that, they could provide opportunities for shameful people, thus helping humans within relationships. For these and other reasons, Levy believes that sex with robots will come to be seen as ethical, and even, they might be expected within relationships in future. Considering both parties, it can be said that this will be a significant ethical issue to be discussed in future.

As another impact, the fear about how robots might have impacts on human life, is not a new issue. Although interaction between humans and artificial intelligence was very little, in 1940s, Isaac Asimov stated his famous Three Laws of Robotics, which were intended to keep robots from hurting human beings. The first law is “a robot may not injure a human being or, through inaction, allow a human being to come to harm”, was based on the assumption that robots would affect humans both in positive and negative manners.

As an early effort to resolve the issue, a group of researchers and experts came together to develop the field of “machine behavior,” by hoping to put human understanding of AI on a healthy theoretical and technical foundation. While making these studies, robots are not seen as only human-made objects, instead they are accepted as a new class of social actors.

To summarize, it can be said that although the fundamental aspects of human behaviors change due to many parameters like geography and culture, some of the emotions and behaviours do not change among societies. Love, family, friendship, cooperation, helping, etc. are the major ones of them. However, involvement of artificial intelligence into these relationships, might be much more disruptive.

Considering that machines are built to look and act like human beings and to diffuse themselves deeply into society life, they might change nature of love, friendship, relations, etc., which will not be just in direct interactions with machines, but in interactions of human beings within each other.

Also, it should be stated that inherent emotions within human beings like love, family, friendship, cooperation, helping, etc. have aided the society to live communally, in earlier times. However, now, human beings do not have time to gain instinctual capacities to live with robots. Therefore, necessary measures are needed to be taken to ensure that they can live with the people in a peaceful manner. As AI will be more fully part of society in future, a new social contract, with machines seems to be required in a short while.

3.4. Safety and security of AI applications

With increasing power of technology, it became much easier to use them in either way (i.e. for malicious purposes as well as good reasons) in a more powerful manner. Since, wars are not expected to be made on ground in future, cybersecurity has already become very important, due to the fact that these systems are much faster and more capable than human beings. Therefore, it can be said that in order to trust on AI systems, they should be safe and secure throughout their operational lifetime. They should also be verifiably during all stages of the operation process.

Safety has become a major parameter for consumers during decision-making process for all sectors. It is well-known that before a new car is introduced to world, it must pass various safety tests to satisfy all regulations, standards and also public expectations. Whatever the technology is, customers expect their products to be safe from end-to-end. Therefore, considering impacts and power, especially AI applications are expected to be built in a safe manner.

However, in order to ensure safety of an AI system, a number of questions need to be answered: What is safety for an AI system? What can be the scope of damage, which these systems can give, in case of safety failure? How can it be ensured that an AI system remains verifiably during all its life-time, including the learning and developing functions on its own processes? How much risk do humans will accept in order to gain potential benefits of AI?

Similarly, it is known that security issues might lead to various problems in the public life. For example, Microsoft’s chatbot (Tay) showed that an AI can learn negative attitudes from its environment, which led to an opposite situation what it has been planned. Similarly, Tesla car could not identify a white truck in a clear sky, and led to a serious accident. In addition, many countries are creating autonomous weapons in form of robots. These self-improving AI systems might become so strong that it could become difficult or impossible to stop them from achieving their goals, which may lead to unintended consequences.

Another challenge regarding safety of these applications is that design of these systems does not have enough transparency during production process, and many of today’s AI applications look like a black box. Thus, AI analysts and developers cannot always identify how or why AIs take various actions, and it seems that this will most probably increase safety issues as AI becomes more complicated and frequently used.

In any way, it can still easily be said that, considering the impacts and power of the AI applications, it is not so difficult to expect that the risks and problems from safety and security perspectives will be increased day by day, in addition to the already existing technological security and safety issues.

3.5. Inequality and unemployment rising

As AI applications are used widely in various industries, economical results are also started to be encountered. Inequality and unemployment are the two major socioeconomic issues of the use of AI applications.

3.5.1 Inequality

From the economical point of view, it is expected that by using artificial intelligence, a company can operate itself with a lower number of staff members, leading to decrease in costs and increase in its revenue. Therefore, one of the main impacts of development of AI applications is expected to lead to increase in inequality among society.

For example, in 2014, roughly same revenues were generated by the three biggest companies in Detroit and the three biggest companies in Silicon Valley, however there were 10 times fewer employees in Silicon Valley, due to use of AI applications.

Artificial intelligence systems seem to bring a more urgent problem for low-skill and uneducated workers in business life. Based on historical trends and current capabilities of AI, it can be claimed that rise of artificial intelligence will lead to decrease of low-skill jobs (i.e. jobs that do not require expertizing or significant training), and creating a larger distance between specialized and the unspecialized workers in society. Additionally, new jobs are expected to be created in different locations from the ones, where old jobs are likely to disappear, potentially increasing the ongoing differences between cities or countries.

In addition to inequality in economical area, another point of inequality seems to be grown in society is related with gender issue, which needs to be removed. It is known that majority of people working in AI development process (i.e. science, technology, engineering and math areas) are males.

Similarly, existing culture is still shaped by men and data use in AI learning process consist of mostly the data produced by males, leading to unobjective and biased results. As a result, algorithms and AI technologies, which have been developed in such an environment, fail to be aware of these diversities, and therefore re-produce those biases and inequalities. Unless the culture itself changes and necessary actions are taken, it can be said that AI might keep generating gender inequality biases.

Thus, in order to advance gender equality and women's empowerment, gender considerations and issues need to be widened across all disciplines and sectors, including AI world. Until that time, AI applications seem to increase the issue of inequality.

3.5.2 Unemployment

Another main ethical issue is the expectation of unemployment, with the increase of automation, particularly AI usage, in society. Some operations will be made by AI applications instead of human workers, and number of working people will be significantly decreased following the usage of AI applications, in some industries. On the other hand, as it is found ways to automate jobs, more complex roles might be created for people, moving from physical work to office environments, which is expected to change strategic and administrative working. However still it seems that unemployment will be a significant issue for society with the increasing use of AI applications in world.

One example is number of trucks in United States. It currently employs millions of individuals in US alone. It is a question that what will happen to them if self-driving trucks become widely available in the next years? Similarly, a San Francisco-based company has designed a fully-automated burger-flipping machine, which might lead to replace workers in fast food restaurants. Additionally, plans have been announced to introduce "fully intelligent robot" police officers in the United Arab Emirates, to provide "better services without hiring more people."

Besides these specific examples, from a higher perspective, World Economic Forum (WEF) warned that it will lead to a net loss of over 5 million jobs in 15 major developed and emerging economies by 2020. These countries include Australia, China, France, Germany, India, Italy, Japan, the UK, and the US.

Similarly, according to the report, released in 2016 and titled *Technology at work: V2.0,* "35% of jobs in UK are at risk of being replaced by automation, 47% of US jobs are at risk, and across the OECD as a whole an average of 57% of jobs are at risk. The risk of automation is 77% in China. Additionally, fears about human workers losing their jobs to machines have been increased by a 72 percent increase in the number of industrial robots in the U.S. over past decade".

Considering both the benefits (i.e. reduced costs, increased efficiency, wide usage, decreased number of accidents, etc.) and the drawbacks (i.e. unemployment, social inequalities, etc.), it seems that the issue regarding unemployment, expected to be raised due to AI applications, will keep being on the center of the discussions.

As a contribution to these discussions, it was stated in the report, which has been issued by Gries and Naudé (2018, p.3), that:

i. the methods used to calculate potential job losses depend on assumptions used;
ii. despite some jobs and sectors may be at risk due to automation, the impact is heterogeneous and many new jobs and tasks may be created in many sectors;
iii. automation may affect the tasks, rather than the jobs;
iv. the speed of innovation in AI is slowing down, and
v. the diffusion of AI in many areas might be much slower than thought previously.

As a summary, it can be said that, technology might not have a purely destructive impact as it was in past. With AI, it is seen that new jobs will be introduced; existing roles will be re-arranged; and individuals will have opportunity to change their careers. The problem will be to manage the transition between these stages.

Unemployment and income inequality seem to be grown, which might possibly lead to political instability. Additionally, people who require to re-train for new work opportunities will not be young, but middle-aged professionals.

3.6. Singularity: Control of a complex intelligent system

It is every time claimed that the reason humans are on top of the food chain is not due to their sharp teeth or strong muscles. Instead, human power generally comes from its intelligence. Human beings are seen as superior than bigger, faster, stronger animals because they can make and use tools to control them. These tools consist of both physical tools (i.e. cages, weapons, etc.) and cognitive tools (i.e. training, information sharing, etc.).

The same critical questions commenced to be asked regarding AI: Will it, one day, have the same advantages over human beings? If so, will it be enough just to "pull the plug" for AI devices, since a sufficiently 'intelligent' machine might expect this action and take necessary measures to prevent it. This is what called the "singularity", which means that the point in time when human beings are no longer the most intelligent beings on earth. In other words, singularity can be defined as the point that machine intelligence will become equal to or more than human intelligence, by the aid of ongoing development in machine learning, which leads to smarter computers.

Bostrom (2014, p.149) outlined a scenario in which a very powerful computer is programmed to make paper clips: "The machine brilliantly and relentlessly pursues this goal and prevents anyone from attempting to change its paper clip imperative. Eventually, the Earth is a mass of paper clips and the computer sets its sights on the rest of the universe." Similarly, in his paper "Ethical Issues in Advanced Artificial Intelligence" (2003, pp 12-17), he argues that "artificial intelligence has the capability to bring about human extinction".

He claims that “general super-intelligence would be capable of independent initiative and of making its own plans, and may therefore be more appropriately thought of as an autonomous agent. In theory, a super-intelligent AI would be able to bring about almost any possible outcome”.

While waiting for the singularity, in recent years, AI weapons started to bring another type of danger. Many governments have started to fund programs to develop AI weapons. U.S., Russia and Korea announced plans to develop autonomous drone weapons. Due to the potential of AI weapons becoming more dangerous than human-operated weapons, Stephen Hawking and Max Tegmark signed a “Future of Life” letter to ban AI weapons. The message posted by Hawking and Tegmark states that “AI weapons pose an immediate danger and that action is required to avoid catastrophic disasters in the near future”.

In short, it is expected that around year 2050, machines will develop a notion of self-awareness and develop something similar to what is called conscience (i.e. awareness of themselves, the surroundings, the purpose of life, ethics, morality etc.). Also, following that point, it might be impossible to differentiate a machine from a human, and machine intelligence might converge with human intelligence resulting in singularity, which seems to be a significant potential issue for AI ethic.

However, considering the existing huge difference between human brain and artificial intelligence, and long way to cover for development of required features for AIs, it seems still to be low probability for this claim to be realized and convergence of human and machine intelligence at one point.

CHAPTER 4

MORAL RESPONSIBILITY AND AI APPLICATIONS

A general information on Artificial Intelligence (AI) has been provided in Chapter 2 and the ethical issues regarding the usage of AI have been discussed in Chapter 3. In this chapter, I will be focusing on the question of whether AI applications can be held responsible due to their activities or not. While discussing this topic, a general information on ethics will be provided first. Then two AI applications (i.e. self-driving vehicles and sex robots) will be selected as examples and their ethical status will be discussed from the point of moral responsibility.

4.1. A brief overview of ethical theories

Ethics or moral philosophy is a branch of philosophy, which generally relates to the concepts of right and wrong behaviours. The English word 'ethics' is derived from the Ancient Greek word *ethikos*, meaning "person's character", which itself comes from the root word ethos meaning "character, moral nature".

Ethics tries to resolve the questions of morality by defining the concepts like good and bad, right and wrong, virtue and vice, and justice and crime. Three major areas of study within ethics are the followings (Wallach & Allen, 2009, Alexander, 2016; Sayre-McCord, 2012):

i. *Meta-ethics*: Aims to understand various features (i.e. metaphysical, epistemological, semantic, psychological, etc.) of moral thought, talk, and practice. Meta-ethics generally asks what is understood and meant, when the questions of what is right and what is wrong is asked.

A meta-ethical question is generally an abstract query and is related with several questions, for example, “Are there moral facts? If they exist, what is the origin of them? How can an appropriate standard be set for behaviors?”.

ii. *Normative ethics*: Concerns with the criteria of what is morally right and wrong. It generally provides rules for guiding human attitudes.

One of the normative ethical theories is the virtue ethics, which was defended by Aristotle. Virtue ethics is related to the virtue and practical wisdom, and focuses on the inherent character of a person rather than on specific actions. The nature and definition of virtues are the main discussion points of virtue ethics.

Another normative ethical approach is the consequentialism, which claims that the consequences of actions are the ultimate basis for any judgment regarding the rightness or wrongness of the action. Therefore, if an action produces a good result, then it is assessed as a morally right action by a consequentialist.

Utilitarianism as a version of consequentialism relates with the actions, which maximize the happiness and benefit (conversely, minimize the sadness, pain, etc.) for the majority of a population. Therefore, the aim of the utilitarianism can be stated as maximization of utility for a society and minimization of adverse impacts for the actions in place.

Deontological ethics or deontology, is an ethical theory that guides and assess choices as morally required, forbidden, or permitted. According to deontological ethics, an action might be assessed as right, even it generates a bad result, if it follows the moral law, which is an opposite approach to consequentialism.

Kantian ethics, as an instance of deontological ethical theory, is based on the view that the only real good thing is good will. It means that an action can only be good if the principle behind it is in line with moral law.

iii. *Applied ethics*: Relates with the question that what a person can and cannot do in a specific position, which is a practical area under the ethics. In general, applied ethics aims to find ethical solutions for practical real-life scenarios. The main specialized field examples are engineering ethics, bioethics, public service ethics and business ethics.

Applied ethics examples related with this thesis are machine ethics that deals with the moral behaviors of artificially intelligent entities, and computer ethics, which deals with ethical responsibilities of computing professionals while making their decisions during generation and operation of computerized systems.

4.2 Moral responsibility

The concept of moral responsibility is used for mostly human actions. It can generally be defined (Noorman, 2018) as, "a person or a group of people is morally responsible when their voluntary actions have morally significant outcomes that would make it appropriate to blame or praise them.". Therefore, it can be said that "ascribing morally responsibility establishes a link between a person or a group of people and someone or something that is affected by the actions of this person or group." Similarly, the terms agent and patient have been defined as: "The person or group that performs the action and causes something to happen is often referred to as the *agent*. The person, group or thing that is affects by the action is referred to as the *patient*." (Noorman, 2018, 1. Challenges to moral responsibility section, para. 1)

The concept of moral responsibility is sometimes confused with similar concepts like accountability, liability, and causality. Additionally, it is not always clear that moral responsibility will be ascribed to whom and why? Although, the philosophical discussions are still ongoing regarding the issue, most researchers share at least the following three conditions, for moral responsibility (Eshleman 2014; Jonas 1984):

i. There should be a causal connection between the person and the outcome of actions. A person is usually only held responsible if she had some control over the outcome of events.

ii. The subject has to have knowledge of and be able to consider the possible consequences of her actions. We tend to excuse someone from blame if they could not have known that their actions would lead to a harmful event.

iii. The subject has to be able to freely choose to act in certain way. That is, it does not make sense to hold someone responsible for a harmful event if her actions were completely determined by outside forces.

Taking these points as reference, moral responsibility differs from accountability, which requires compensating the outcome of the actions or punishing for the consequential damages, from an administrative perspective. Similarly, it disconnects from liability, which is generally related with looking for a person to blame and to compensate for damages suffered after the event from a legal perspective. Responsibility also is different from causality that is generally related with the "cause" of the actions, which might not always be enough to assign morality to the subject of the actions. Therefore, in the rest of the thesis, I will exclude accountability, liability and causality from the discussions and use only the moral responsibility concept during my analysis.

In the past, even today, moral responsibility is generally assigned to human beings, based on the fact that they can freely choose to act in one or another way and evaluate the consequences of this choice. However, following the increasing complexity of the life, and especially the usage of the artificial intelligence applications in society, the discussions about computers and treating them as if they are moral agents, had started.

These discussions are expected to increase based on the fact that AI applications are being used in a wide area in daily life, and therefore have or potentially will have a very large spectrum of impacts.

Thus, considering the wideness of the area and complexity of the issues, for an effective discussion, it might be useful to pick a number of examples and proceed the discussion on a more solid basis, with the aid of such examples. Therefore, I will select two well-known AI applications (i.e. self-driving cars and sex robots) and try to discuss the ethical issues on these examples. Since self-driving vehicles and sex robots may have a great potential to be widely used in society in near future and also have a large variety of impacts (i.e. social, cultural, physiological, philosophical, relationship, etc.) on society, they might be the correct examples to focus on and discuss the issues around. Therefore, I will proceed the discussion by focusing on these two examples.

4.3 Two sample AI applications

Before discussing the issue, a brief information regarding these two samples will be provided. One of them is self-driving cars that started to be used on the roads in many countries. The main benefit of these cars is safety, since they can solve many of problems encountered due to the human driving errors. The Association for Safe International Road Travel states that "nearly 1.3 million people die in road crashes each year" and "an additional 20–50 million are injured or disabled" due to such crashes, all over the world. Most experts agree that the introduction of self-driving cars will lower the overall number of traffic accidents by 90% and therefore a great number of traffic deaths, which might end up saving about a million people a year. Additionally, it is expected that these cars might also increase oil economy, decrease traffic jams, resolve parking-related problems and provide more mobility to the people who are currently unable to drive, including the ones with disabilities.

On the other hand, since autonomous vehicles are highly dependent on software applications and sensors, which are very sensitive to making errors, it is also expected that new ethical challenges (i.e. crash, kill or hurt someone) will be encountered during the operation of these devices (as mentioned in Chapter 3).

These accidents led to an increase in ethical concerns about self-driving cars and also led to the following questions: Who is responsible if something goes wrong with these cars? Could a car be held responsible if it is involved in an accident? If so, what will be the ethical, legal and social status of the case and how will it be treated, since it cannot be punished or jailed?

Similar situation exists for robots, particularly for humanoid robots, which are similar to human beings in terms of shape (i.e. head, body, legs, eyes, mouth, etc.) and function (i.e. interaction with human, smile, talk, etc.). They are mostly used for performing tasks like personal assistance (i.e. to assist sick and old people), undertaking difficult or dangerous jobs, accelerating processes in manufacturing, etc. Since they can use tools and operate many kinds of equipment, humanoid robots can theoretically perform any task a human being can do, if they have the proper software. However, it is not an easy process to implement them due to complex nature of software.

A specific type of humanoid robots is the relationship robot, or particularly, sex robot. These robots are designed and implemented by using voice and facial recognition software, motion-sensing technology and animatronic engineering, to provide smiling, entertainment, conversation, or other sexual services to their users.

The realization of sex robots was firstly encountered towards the end of 1990s. As of 2018, various models to hold conversations, remember important data, and express various emotions, have been produced. For instance, one of the products is “Harmony”, which is customizable by using an application, where users can choose from “thousands of possible combinations of looks, clothes, personalities and voices to change”. Therefore, it can smile, blink her eyes, talk, tell joke, remember some special dates, etc., as if it is a human being. Similarly, the other sex robot models like “Samantha” or “Roxxxy”, can be changed to include the feature simulating an orgasm, or a family mode with telling jokes and discussing philosophy. Although the market seems to be designed for mostly men, male sex robots also started to be available in the market. “Henry” is one of them, which has been presented to the market recently.

There are many benefits of these robots. They can be used for giving service for old or sick people, providing emotional support for the individuals having difficulty in forming relationship, or helping as a therapeutic machine for people suffering with dementia, depression and/or anxiety, etc. Similarly, they could provide constant availability, decreased sexual disease rates and diminished number of sex workers of all genders. Because of these benefits, their usage is expected to increase. However, criticisms also exist regarding the negative impacts of sex robots. One of the main discussion points is that sex robots are expected to facilitate “social isolation” in the society. Similarly, it might be claimed that sexual relations with robots have the potential of decreasing intimacy and empathy, particularly among males. The other potential harm that sex robots might give is increasing the rape culture, child abuse and pornography.

Kathleen Richardson from De Montfort University and Erik Billing from University of Skövde initiated the “Campaign Against Sex Robots” in 2015. Similarly, Kate Darling, robot ethicist and researcher at the MIT Media Lab, claims that, “Ethical issues arise from concern that we might behave certain ways towards very realistic sex robots that look like a real woman,” because “that behavior might translate to our interactions with real women.”

To summarize, the following questions and ethical issues might be encountered, with the increasing usage of sex robots:

i. Is sexual intercourse with a sex robot different than masturbation legally, morally, ethically?
ii. What would be the impacts of sex robots to social institutions and behaviour models like marriage, partnership, etc.?

iii. Would the robot market be a male-oriented one? Would it contribute to the existing inequality between women and men? Would men's sexual interaction with sex-robots have a negative impact on their treatment of women in their daily life?
iv. Would it be acceptable to abuse a sex robot? Would robots have any rights? If so, to what extent?
v. Would sex robots increase the pedophilia or could they be used for treatment of paedophiles?
vi. What ethical duties do humans as designers, producers and as intimate partners have on these robots?
vii. In which ways are humans vulnerable to sex robots, and sex robots to humans? If a sex robot harms a person, who would be responsible from this act?

Additionally, since all potential and existing risks have not been fully identified and resolved for the time being, sex robots do still have morally and socially problematic aspects. The risks coming from physical interaction with robots and human beings (i.e. sex robots' lips might include toxic paintings, etc.), security of the information stored by AI (i.e. it might have a huge of personal, physical, and geographical data about the 'user' of it), or potential hacking activities (i.e. sex robot might be hacked and its arms and legs might be used to attack to someone) are still awaiting to be resolved.

4.4 Moral machine responsibility: Consciousness and brain perspective

In order to make an assessment regarding the ethics of the machines (i.e. AIs), first, capabilities of AI might be reviewed. As it was stated earlier, AI applications can be classified by their types as weak (narrow) or strong, which perform a limited capability of functionality and full imitation of human beings, respectively. In either case, it might be claimed that AI applications might be morally responsible since they have a social interaction in the society, and are capable of inputting, outputting, storing, etc., which are almost the same as human beings's functions.

However, for an AI application to be considered responsible for its actions, first a moral agency status must be attributed to the application. Such an attribution, in turn, requires the application to have conscious mental states. Despite there exists no agreed consensus on how to define consciousness, it can roughly be said that it is the ability to know what it is like to have a mental state from an individual's own perspective, to subjectively experience her/his own environment and internal states.

Based on this definition, it can be said that performing an action does not necessarily mean the action owner (i.e. an AI) has moral values and/or can be held responsible due to its actions, which means it does not have a 'consciousness'. It is known that in social life, beliefs, emotions, non-verbal activities have a significant meaning in understanding the status, needs and desires of the entities, and acting based on these requirements to be in line with moral law. However, it is extremely difficult for a machine to act accordingly since it has not been equipped with all the information required to know all these attributes related with human beings, animals or other entities. For instance, AI applications have not been learnt about all emotional and physical attributes of human beings and animals, therefore it is not possible for them to understand the situation of the human beings and animals, and act in an ethical manner, in a morally confliction situation, as a moral agent with consciousness.

Similarly, for the sex robot example, it can be said that there exists no direct connection between the robots and external world, apart from the programs deployed to them. Since human decisions are produced based on their complicated brains and hearts, and robots do not have such organs, it can be said that all the decisions are expected to be made based on only the software programs they have. That means, they do not have all the features of human brains and hearts, since no application software can be programmed to include all the existing and possible situations that might be faced with in the world. To overcome this problem, scientists are working to transfer human brains to robots, however there exists no significant advances in this area, and all the components of the brains and mental states of human beings cannot be transferred to the robots, yet. Therefore, robots cannot have the ability of understanding someone's feelings, and think from other's perspective. Considering the lack of adequate subjectivity based on emotions, and perspective of another human being, robots might not be expected to act like a human being, which means that they cannot be held responsible for most of the cases in life. This has been also supported by Nath and Sahu (2017) as:

> We claim that the domain of ethics is built on our ability to have the first-person perspective. Because we can imagine what the other may be feeling, we have the input to be more judicious in our actions and policies. It is because of this subjective feeling, that morality is subjective. A most common tool in ethics is to imagine the situation from someone else's perspective and then decide. This ability is built upon two capacities—first, to have a first-person perspective and, second, to imagine the other's first-person perspective. These two capacities lay the foundation for the ethical domain. And it is here that AI does not clearly harbor these two capacities so as to have the moral ability. (p. 9)

On the other hand, considering simply that consciousness is a feature of brain, it might be thought, in a materialist perspective, that if an AI application has a brain similar to a human's one, then it might have, to some extent, consciousness. Dennett (1994), related with this discussion, first asks the question that "Are conscious robots possible in principle?" and then summarizes the opposite views under four topics as:

> i. Robots are purely material things, and consciousness requires immaterial mind-stuff. (Old-fashioned dualism.)
> ii. Robots are inorganic (by definition), and consciousness can exist only in an organic brain.
> iii. Robots are artefacts, and consciousness abhors an artefact; only something natural, born not manufactured, could exhibit genuine consciousness.
> iv. Robots will always just be much too simple to be conscious.

He provides his opposite arguments for these points by reminding the long-term project to design and build a humanoid robot (i.e. Cog) that interacts with human beings, takes care of itself and tell its designers necessary inputs. He concludes that despite it seems impossible to make a robot with consciousness like a human being, a robot could be made with necessary features that can be accepted as consciousness. (pp. 133-145)

Similarly, Brooks (2002) supports this idea from a different view. He states that:

> If indeed we are mere machines, then we have instances of machines that we all have empathy for, that we treat with respect, that we believe have emotions, that we believe even are conscious. That instance is us. So, then the mere fact of being a machine does not disqualify an entity from having emotions. If we really are machines, then in principle we could build another machine out of matter that was identical to some existing person, and it too would have emotions and surely be conscious. (p. 13)

He also claims that since he argues that humans are machines with emotions, similar case might be in place for machines to have emotions and being a machine does not does not necessarily mean that they do not have emotions, which therefore does not prevent them from being conscious. (p. 13) He (2002) concludes as:

> In my opinion we are completely prescientific at this point about what consciousness is. We do not exactly know what it would be about a robot that would convince us that it had consciousness, even simulated consciousness. Perhaps we will be surprised one day when one of our robots earnestly informs us that it is conscious, and just like I take your word for your being conscious, we will have to accept its word for it. There will be no other option. (p. 21)

Related with the discussion, another claim has been provided by Searle (1997, p.9) as: "Many people still think that the brain is a digital computer and that the conscious mind is a computer program, through mercifully this view is much less wide-spread than it was a decade ago. Construed in this way, the mind is to the brain as software is to hardware." He also states that computers can simulate the mind by simulating many mental processes of thinking, deciding, etc. and claims that, they can never create real mind, real intentionality, real intelligence, real consciousness, but only *as if* consciousness. Despite he reminds that "biological brains have a remarkable biological capacity to produce experiences, and these experiences only exist when they are felt by some human or animal agent", he does not claim that brain tissue is necessary for consciousness. He concludes that other systems could be conscious too, but only if they had equivalent causal powers to those of brain. (p. 212)

On the other hand, Blackmore (2004) discusses the issue by asking the question: “Is there something special about human beings that enables us to think, and see, hear, and feel, and fall in love, that gives us a desire to be good, a love of beauty, and a longing beyond? Or are all these capacities just the products of a complicated mechanism? In other words, am I just a machine?” She reminds that, from the natural direction, mechanisms of human functions like perception, learning, memory and thinking have been successfully explained by the development of science. Similarly, from the artificial direction, she mentioned that machines, especially robots are developed to perform many tasks that human do. Therefore, by claiming that human brain and machines are converging day by day, she asks the main question: “If machines could do all the things we do, just as well as we do them, would they be conscious like us? How could we tell? Would they *really* be conscious, or just *simulating* consciousness? Would they *really* understand what they said and read and did, or would they just be acting *as if* they understood?” (pp. 181-182)

She (2004) replies the question by reminding that the main objections to the idea that “a machine could never be conscious” can be listed as:

i. Souls, spirits and separate minds: According to this idea, consciousness is the property of non-physical mind, which is separate from physical brain, or the unique capacity of the human that is given by God to people, and therefore machines can not be conscious. However Blackmore reminds that, one day, a machine can be made with all the human features (i.e., chat to people happily, wonderfully sympathetic, full of emotions, making laugh with funny stories, etc.) and then it can be claimed that this machine could also have a soul, which is either given by God, or the manufacturer of it. (pp. 200-201)

ii. The importance of biology: According to this idea, only living, biological creatures can be conscious due to the functions of neurons, or the biological creatures need to grow up and learn in a long period of time. Therefore, they claim, a machine that is non-biological and manufactured cannot be conscious. However, Blackmore reminds that the biological components (i.e. neurons and protein membranes, etc.) can be integrated to machines to make conscious machines possible. Similarly, she claims that machines can be manufactured with fast learning capacities and full memories, to disproof this idea. (p. 201)

iii. Machines will never do X: According to this idea, there exist some things (called as X) that machines cannot do since they require the power of consciousness. However, Blackmore reminding that machines started to write poems, make pictures, compose music, etc. claims that although there still exist some tasks that cannot be performed by machines, their number is decreasing day by day. Similarly, she states that the only objection regarding the creativity of machines stays as *real* or *as if*, which seems to be disappeared soon. (p. 202)

Blackmore also reminds that the main argument point comes from the concepts of *real* and *as if* regarding the activities or the consciousness of machines, and claims that the line separating these two concepts is not clear. Similarly, she states that a magical X is expected to add to the machines to have consciousness, however this X is not defined and clear enough. Therefore, she claims that machines with necessary features could be, at least theoretically, accepted as conscious (p. 215) and concludes as it cannot be claimed that machines cannot have consciousness due to two reasons: “First, we do not know what consciousness is. Each of theories say something about what consciousness is.”, and she adds “Second, we have no test for whether a machine is conscious or not.” (p. 217)

Based on this approach, developments are focused on building self-aware robots, which can explore their own physical capacities, to find their own capabilities and to determine their own way to move accordingly. For example, Weng (2002, p. 2) states that “motivated by neuroscience, it is proposed here that a highly intelligent being must be Self-Aware and Self-Affecting (SASE), which is defined as an agent that has internal sensors and internal effectors. In addition to interacting with the external environment, it senses some of its internal representation as a part of its perceptual process and it generates actions for its internal effectors as a part of its action process.” Weng (2002) also adds that “The performance of a practical developmental robot is limited by five factors, which are sensors, effectors, computational resource, developmental program, and how the robot is taught”. He also provides the definition of completeness as “a type of agent is conceptually complete if it can actually reach the human performance norm of any age group on any concept without human reprogramming.” and concludes as if a robot is not conceptually complete, then it cannot learn all the concepts that a human can. (p. 6)

Similarly, Novianto and Williams (2009, p. 1049) define self-awareness of robots as being the capability of an agent to focus attention on the representation of internal states. According to them, “Internal states can be made up of emotion, belief, desire, intention and expectation or it can be processes such as sensation, perception, conception, simulation, action, planning and thought.” They also present a basic framework that it includes the following four major components:

> i. *Physical body:* In order to exist in a physical world, a robot has a physical body that can interact with that world.
>
> ii. *Perception:* Outward and inward facing sensations are processed further to create perceptions. Perception processes (e.g. fusion) the outward facing and inward facing sensation to gain information about self and the surroundings.
>
> iii. *Self-concept:* Contains collections of the facts, conditions, or other representations attained from perception that characterize the self.
>
> iv. *Attention:* Attention can highlight representations, e.g. beliefs and processes, for an agent in being unaware to aware. (pp. 1049-1050)

They (2009) also provide the main tools used for self-awareness of robots are; motion recognition (i.e. forward, stop, back, etc.), mirror recognition (i.e. direct visual recognition), imitation recognition (ability to distinguish self from imitation), emotion model (e.g. emotion agent to model emotion that can activate the episodic memory from a stimulus in the environment), memory model (i.e. short-term memory, long-term memory and working memory), physical body model (e.g. developing its own physical body model that can be used to generate behaviors when some changes happen to its body), sophisticated behaviour process (e.g. simple arm motion movements, locomotion, and experimentations), attention to internal state process (e.g. use of attention in working memory system and central executive agent to select information), and self-modifying code process (e.g. attention system that indicates what representations are being enacted and what code of a process is currently executing, to be aware of the internal states, so it has the potential to self-modify this code to suit more complex conditions and adapt to its dynamic environment). (pp. 1052-1053)

Despite it is still far to fully implement, this method of self-modelling might be applied to more “developed” robots for ethical decision-making, by exploring their own capacities for actions and building an ethical model for themselves, which still seems not so recent.

4.5 Moral machine responsibility: Animal analogy

A moral agent is an entity, which can be held responsible for its actions and their consequences, with appropriate reasons. From the moral responsibility and rights of robots points of views, AI applications are mostly compared with the status of animals. It is thought that they have similarities, which can provide a reference to discuss, to some extent, about AI moral responsibilities. Peter Singer (1993, p. 63) specified that “Animals are treated like machines that convert fodder into flesh”. Several similarities and differences may be found between animals and AI applications, or particularly robots. It is generally believed that rights of animals are required due to the existence of their emotions (i.e. they can suffer due to pain and have pleasure due to good care on them). Despite it is claimed that it is not the same situation for the machines, since they do not have such kind of feelings (i.e. pain, happiness, etc.) and therefore cannot suffer or become happy, this similarity still might be used to analyze the moral status of machines.

However, this analogy is generally discussed and criticized from a number of points. First, it is not always very simple to understand whether an entity suffers from a pain or not, unless this feeling has been experienced by any other entity. As P. Singer (2002, pp. 10-12) stated, “we cannot directly experience anyone else's pain, whether that 'anyone' is our best friend or a stray dog. Pain is a state of consciousness, a 'mental event,' and as such it can never be observed.”

Similarly, Brooks (2002) wonders about whether humanoid robots will have enough similarity between human beings to be treated in the same moral ways that people treat other people or animals, and by reminding that robots are started to be built with emotional systems, asks the following questions: “Are they real emotions, or are they only simulated emotions? And even if they are only simulated emotions today, will the robots we build over the next few years come with real emotions? Will they need to be visceral emotions in the way that our dog can be viscerally afraid? What would it take for us to describe a response from a robot as visceral?” As a response to these questions, he claims that “For if we accept that robots can have real emotions, we will be starting down the road to empathizing with them, and we will eventually promote them up the ladder of respect that we have constructed for animals.” (pp. 3-4)

Wallach and Allen (2009) also participate to the discussion and state that “When a robot dog wags its tail or hops around as one gives it attention, it is not ‘happy’; it has no internal states comparable to human emotions, or even the emotions of an animal” (p. 43). They also remind that the motivations for human or animal behaviours (i.e. pleasure or punishment) might not be valid for machines and claim that:

> It is sometimes suggested that punishment and rewards might be communicated in terms a computer would appreciate directly, for example, by manipulating processor speed, information flow, or the supply of energy. But these seem either naive or far-fetched and futuristic. Still, even without conscious pleasure or pain, computational learning mechanisms may be able to learn some basic patterns of moral behavior. (p. 109)

On the other hand, there exists a long way to develop systems that can feel pleasure or pain, or have emotions seen in the humans or animals. Wallach and Allen (2009) underline this idea as "The robots available today do not have nerves, neurochemicals, feelings, or emotions, nor is it likely that robots in the near future will. Nevertheless, sensory technology is an active area of research, and it is here that one might look for the foundations of feelings and emotions." (p. 150) and remind that "successful Artificial Moral Agents (AMA) might be constructed even if they could never be held directly responsible for anything, just as artificial chess players can win tournaments even though they never get direct credit for doing so." (p. 208)

On the other hand, they (2009) remind that if one day robots will deserve equal treatment under the law, this kind of movement is likely to come gradually. They claim that, from this perspective, robot rights are similar to the politically significant movement to increase rights for animals and "much of the animal rights movement has focused on protecting the more intelligent species from pain and distress." (pp. 208-209)

They (2009) summarize their ideas regarding the discussion as:

> Humans have always looked around for company in the universe. Their long fascination with nonhuman animals derives from the fact that animals are the things most similar to them. The similarities and the differences tell humans much about who and what they are. As AMAs become more sophisticated, they will come to play a corresponding role as they reflect humans' values. For humanity's understanding of ethics, there can be no more important development. (p. 217)

Similarly, there exists another discussion point regarding this analogy (i.e. moral dilemma), which comes into the picture inherently by the equipment of robots with pain. Wallach and Allen (2009) have concerns about the issue:

> Pain and emotional distress are not yet, of course, issues for robots. It will be particularly difficult to establish whether these future robots actually have any subjective experience of pain, just as it is difficult to establish whether people in vegetative states experience subjective pain or what kinds of pain animals experience. If robots might one day be capable of experiencing pain and other affective states, a question that arises is whether it will be moral to build such systems —not because of how they might harm humans, but because of the pain these artificial systems will themselves experience. In other words, can the building of a robot with a somatic architecture capable of feeling intense pain be morally justified and should it be prohibited? (p. 209)

They (2009) also mention about the objections on building robots with conscious self-model, and state as:

> If not prohibited, should there be regulation of experiments in which a robot might experience emotional states? Regulations protecting animals are far less stringent than those protecting humans, and there is much scientific disagreement about how animal pain and distress can be measured. To date, there are no review boards to oversee the ethical treatment of robots in research, nor is there any need for them. However, as the appearance of subjective feelings of pain and pleasure in robots becomes stronger, there will be calls for regulations and review boards to oversee the kinds of research that can be performed. (p. 209)

On the other hand, there exist objections to Wallach and Allen. For example, Mehlman, Berg and Ray (2017, p. 8) remind the question “When do Artificial Intelligence Robots (AIR) begin to have legal rights?” and reply as “The answer is, at a minimum, when they become like animals that are capable of experiencing pain or suffering. The first robot right then is the right to be free from pain and suffering.” They (2017) also mention that:

> We therefore disagree with Wallach and Allen that “unlike most other kinds of rights for robots, marriage is an issue that humans will have a direct interest in, and may therefore be among the first rights considered for robots”. The right to be free from pain and suffering is not an absolute right for AIRs any more than for humans; the infliction of pain and suffering on humans is permitted under

> certain circumstances, such as the use of reasonable force in self-defense and in law enforcement, so long as it is not cruel and unusual punishment, and the same should be true for AIRs, although what would count as cruel and unusual punishment for AIRs would need to be determined." (p. 8)

In any way, it is known that the features of AI applications, especially the robots, are increasing day by day. For today, there exist robots, which can smile, cry, tell joke, even feel orgasm. Furthermore, it is now possible for a person to have a robot partner, which she/he can determine the specifications, i.e., happy, jealous, mad, cool, etc. It is seen that the difference between robots and human beings are getting closer and closer, with the new developments. Therefore, as other emotions, 'pain' might be integrated with robots in a short time, and this might lead to a change in the moral status of the robots (or in general, AIs), based on this analogy. However, considering that all kinds of feelings have not been deployed to robots yet and it needs a very long time and difficult process to build robots that have all such kinds of emotions, it can be said that this analogy cannot be used as a reference to assign moral responsibility to robots, at least for today.

To conclude, despite animals and robots are seen as subordinate to human beings and depend on them from many aspects, based on the above-mentioned reasons and considering that robots are not subject to praise, blame or punishment due to their actions, a direct comparison between animals and AI agents will not be meaningful. Therefore, such feelings cannot be taken as a reference to grant or unassign moral status to the robots, and it seems that AI applications and specifically robots will not be accepted as morally responsible entities, unless they have psychological features not only similar to animals, but also look like human beings.

4.6 Moral machine responsibility: Autonomy and moral responsibility

From the moral perspective, autonomy can be generally defined as the ability to impose moral law on an entity. Similarly, by moral responsibility, it is meant that decisions are made by a conscious entity with free will, without referring to a higher authority. On the other hand, causal responsibility relates with share of an entity (i.e.

subject or object) in a causal chain of events. For example, a candle might burn a house, and it can be thought that it is causally responsible from this action, however it cannot be held morally responsible due to this action.

Kant (1996, pp. 73-89) claims that the laws, which human beings should follow, have to be created by a good will, and this should impose rules for all human beings. According to him, laws are meaningful to humans only if they are universal. This leads to the well-known moral “categorical” imperative. That means since human beings are the authors of the laws they follow, then their will can be accepted as autonomous.

However, when AI applications are considered, autonomy is generally used for decisional autonomy. It means that all their activities (i.e. sensing, perceiving, analyzing, communicating, planning, decision making, operating, etc.) performed by their initiatives, which can be seen a kind of autonomy. Furthermore, they are considered as autonomous, in the meaning of, they are not dependent on someone, to perform their operations, once they started to be operated. So, this is a technical autonomy, which should not be confused with the meaning in the moral sense. These machines are not autonomous in the etymological sense, since they do not give the decisions based on their own laws. Since they are dependent on their programming, they cannot choose, and they cannot be free. Therefore, according to Kantian approach, robots are not autonomous, since they are not able to define their own goals and laws, and are just performing the activities designed by human beings, which also leads to the point that they cannot be accepted as morally responsible entities.

On the other hand, if the robots are considered as “tools”, instead of “autonomous agents”, it can be seen that they share the responsibility of the actions with or transfer the responsibility to another entity. As a first approach, it can be considered as robots are solely the products and designed by a company. Thus, in case of a failure, it is obvious that the company will be held responsible instead of the robots. Another approach is so-called slave morality stated by Ruffo (2012, p. 89). According to this approach, a slave, by itself, is not considered responsible for his actions, but his owner has the responsibilities. So, if this approach is applied, the responsibility will be undertaken by the closest person in the chain of production, i.e. the person who decided and provided the deployment of the robots.

Similarly, Brooks (2002, p. 22) reminding that one of the attractions of robots is that they can be the slaves of human beings, he asks the following questions: "But what if the robots we build have feelings? What if we start empathizing with them? Will it any longer be ethical to have them as slaves?" He replies these questions as "This is exactly the conundrum that faced American slave owners. As they or their northern neighbors started to give humanhood to their slaves, it became immoral to enslave them. Once the specialness of European lineage over African lineage was erased, or at least blurred, it became unethical to treat blacks as slaves. They, but not cows or pigs, had the same right to freedom as did white people. Later a similar awakening happened concerning the status of women." He (2002) also adds that:

> Fortunately, we are not doomed to create a race of slaves that is unethical to have as slaves. Our refrigerators work twenty-four hours a day seven days a week, and we do not feel the slightest moral concern for them. We will make many robots that are equally unemotional, unconscious, and unempathetic. We will use them as slaves just as we use our dishwashers, vacuum cleaners, and automobiles today. But those that we make more intelligent, that we give emotions to, and that we empathize with, will be a problem. We had better be careful just what we build, because we might end up liking them, and then we will be morally responsible for their well-being. Sort of like children. (p. 22)

To conclude, considering that ethics is mostly required if there exists a conflict between existing rules (i.e. legal or moral) or there is a lack of rule to guide the actions, it can be said that solving an ethical conflict requires a sense of creativity in case of a complex situation, and the entity, should be able to provide alternative solutions based on moral rules. Since AI applications and robots do not have a complete autonomy or the skills to be able to analyze their environment accurately, it can be claimed that they cannot fully understand what happens in their environment, in a given situation, with all the impacts and outcomes. Therefore, it might be expected that AI applications and robots can face with an enormous number of complex situations, which will be a huge challenge for them to handle all the dimensions, in a morally responsible manner, which most probably they cannot succeed to manage.

4.7 Moral machine responsibility: Three approaches for moral status

Related with the question of how AI applications can solve ethical problems, there exist three types of approaches in place. It was stated by Wallach and Allen (2009, pp. 79-124) as:

i. *Top-down:* In the most general sense, the top-down approach to artificial morality is about having a set of rules (i.e. consequentialist or utilitarian ethics, Kant's moral imperative, legal and professional codes, Asimov's Three Laws of Robotics, etc.) coming from the sources like philosophy, religion, literature, science, etc., which can be turned into a computer algorithm, for usage of AI. According to this approach, a set of rules are taken and integrated to the program on AI systems explicitly, and they are expected to act in line with these rules. Wallach, Allen and Smit (2008, p. 569) provide examples for the usage of this approach as "To date very little research has been done on the computerization of top-down ethical theories. Those few systems designed to analyze moral challenges are largely relegated to medical advisors that help doctors and other health practitioners weigh alternative courses of treatments or to evaluate whether to withhold treatment for the terminally ill." Despite it seems useful, top-down approaches have some challenges like different rule sets to be followed for a specific case might conflict with each other and sometimes rules might be too general to follow requiring more detailed guidance for resolution of the issue. Similarly, although Asimov's Three Laws of Robotics rule is generally followed, for self-driving car and sex robot examples, it is obvious that integration of all rules and standards onto these machines prior to the deployment to production, is impossible due to infinite number of situations and scenarios to be encountered in the real life.

ii. *Bottom-up:* The goal is to form an environment, which robots can explore different types of attributes, for morally satisfying actions, with the aid of implicit values. Most of the bottom-up approaches depend on machine learning systems or focus on the autonomous robots, which can learn their own ethical reasoning abilities. However, as a drawback, learning process takes a lot of time and generally cannot completely remove the risk of unwanted behaviours, which might be faced with in future. Additionally, the reasoning behind the actions produced by AI systems generally cannot be traced, thus it makes the identification and analysis of undesirable behaviours too difficult, for a sound operation and development. Wallach, Allen and Smit (2008, p. 568) comment on bottom-up models as "These approaches to the development of moral sensibility entail piecemeal learning through experience, either by unconscious mechanistic trial and failure of evolution, the tinkering of programmers or engineers as they encounter new challenges, or the educational development of a learning machine." Despite Alan Turing (1950, pp. 433-460) reasoned that "if we could put a computer through an educational regime comparable to the education a child receives, we may hope that machines will eventually compete with men in all purely intellectual fields', it can be concluded as self-driving cars and sex robots cannot have moral responsibility based on bottom-up approach, since they do not have adequate capabilities to learn and solve all kinds of ethical problems they might encounter in their life-times, since they are subject to the issues stated in Chapter-3.

iii. *Hybrid:* Hybrid approaches combine the specifications of both top-down (producing algorithms derived from ethical theories) and bottom-up (using agents able to learn for ethical decisions) methods. Wallach and Allen

(2009, p. 117) claim that “If neither a pure top-down approach nor a bottom-up approach is fully adequate for the design of effective AMAs, then some hybrid will be necessary”. They (2009, p. 178) also add that “Memories and personality traits find their way into the mix. In all likelihood, no two people process moral decisions in quite the same way, even when confronted with identical challenges. Humans are hybrid decision makers, with unique approaches to moral choices, honed over time and altered by their own distinctive experiences.” Similarly, Wallach, Allen and Smit (2008, p. 571) reminding that moral development of human being is formed based on hybrid model, claim that “Genetically acquired propensities, the rediscovery of core values through experience, and the learning of culturally endorsed rules all influence the moral development of a child. During young adulthood those rules may be reformulated into abstract principles that guide one’s behavior.” However, the main problem with these approaches is that their computing time might be too long, since these processes heavily depend on learning. For the case of the self-driving car and sex robot examples, it can be said that they cannot be accepted as moral entities, since establishment of moral status might need a very long time and also this process is subject to several problems and risks like unintended consequences, bias, safety and security issues, etc., as it was mentioned in Chapter 3.

4.8 Moral machine responsibility: An assessment based on traditional ethics conceptions

The moral status and responsibility issue of the AI applications can be assessed based on the main traditional ethics concepts. These approaches are given in the following section:

4.8.1 Aristotelian ethics

While discussing the issue, Aristotelian approach is taken first as reference. Ruffo (2012) reminding that the goal of ethics is a good life to provide happiness, and good life is considered to be achieving the goal, which involves a human being to be virtuous, according to Aristotle, adds that “Practical wisdom is that which allows us to judge and act according to a happy medium and according to the circumstances.” Therefore, it can be said that ethics is to behave in the best manner under practical conditions to achieve happiness, instead of a theoretically discussion on the absolute good. (p. 87)

If this understanding is taken as a basis, it can be said that it is very difficult to make a relationship between AI applications and these concepts. If the aim of ethics is happiness (i.e. well-being, satisfaction, etc.), it is not known that how the happiness can be defined by a robot. If these applications/machines do not have human-like feelings, then happiness or other similar feelings do not have any meaning for them. In short, it can be stated that the goal, which motivates ethical behaviour according to Aristotle, cannot be a meaningful concept for AI applications, or specifically robots, at least for today.

Similarly, from the judgment perspective, it is not clear that AI systems have really a judging capability, or not. It is known that AI applications can perform several operations, which depend on some measurements, with the aid of their sensors and programmed algorithms. However, their actions depend on the expected responses integrated within their programs, which have a set of pre-defined parameters. Therefore, it cannot be claimed that the automated responses can be defined as judgment. When it is told about the concept of judgment, it generally involves a careful analysis with the positive and negative impacts, and creating original solutions for unusual or unforeseen circumstances, instead of selecting a task among a set of pre-defined alternatives. Thus, it cannot be said that self-driving cars or sex robots are the authors of their actions, therefore they cannot make judgment, in the real sense. An AI application can be seen as only an instrument, which is a kind of extension of human actions and decisions. Based on this reality, it cannot be claimed as self-driving cars or sex robots are morally responsible entities.

Another basic concept in Aristotelian ethics, which is used for moral behaviours is the ‘empathy’, as stated by Ruffo (2012). Empathy can be defined as the capacity to put oneself in the place of another entity, to understand what she/he/it feels according to the situations or reactions. This is widely used in social relationships. However, these kinds of emotions are not used by autonomous systems, which generally execute their operations, with the aid of the sensors equipped them, for further decision. This leads to ethical issues in relations. For instance, how can a care-robot understand why a baby cry? Robots can be aware that only a limited set of requirements for babies (i.e. eating, cleaning, no physical pain, etc.) might make them cry. However how can they understand that a baby is afraid from something, or crying since it did not see the parents at that time or there is lack of one of her/his toys, etc.? So, since the reason cannot be understood, it would be very difficult for an AI application to calm a baby down. These kinds of examples show that robots cannot have moral status when the notion of ‘empathy’ is taken as basis. (p. 88)

To conclude, since autonomous devices are limited to analyze all dimensions of a situation regarding real-life situations, this will impair its capability to find effective solutions for unexpected problems. With the lack of some important feelings (i.e. empathy, compassion, etc.) they seem unable to have Aristotelian ‘virtue’, which is unique -at least today- to human beings, and the ‘good life’, to have the ability to properly judge according to circumstances. Additionally, a ‘good life’ is not meaningful for an AI application, which has no personal feelings or sensations. So, it can be summarized as these machines do not have a life, neither good nor bad, instead they have only a period of use, which this situation brings them to the status of morally irresponsible according to the Aristotelian approach.

4.8.2 The Kantian approach

According to Kant's (1996) categorical imperative, first formulation is "Act that you can will that your maxim should become a universal law (whatever the end may be)." (p. 344) In this formulation, for an AI application to have moral status in the Kantian sense, it should have an ability to assess the universality of a rule, for all the beings in nature including humans, animals, environment and also machines. However, since they do not have enough information installed on them, and the values like empathy, wisdom, emotions, etc. have not been fully integrated with them, it cannot be expected from the machines to find the right ethical behaviour and follow it considering that it can be a universal rule.

Similarly, second formulation of categorical imperative is "Act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means." (p. 80). This formulation can also not be used as a reference for AI applications to assign moral status in a Kantian sense, since machines, by definition, are designed and built by human beings, and are subject to human programming, at least for today. Therefore, machines can not be expected to have moral values based on this approach.

On the other hand, Powers (2006), related with the categorical imperative approach, claims that:

> For Kant, maxims are "subjective principles of volition" or plans. In this sense, the categorical imperative serves as a test for turning plans into instances of objective moral laws. This is the gist of Kant's notion of self-legislation: An agent's moral maxims are instances of universally quantified propositions that could serve as moral laws —that is, laws holding for any agent. Because we can't stipulate the class of universal moral laws for the machine —this would be human ethics operating through a tool, not machine ethics— the machine might itself construct a theory of ethics by applying the universalization step to individual maxims and then mapping them onto traditional deontic categories —namely, forbidden, permissible, obligatory actions— according to the results. (p. 47)

However, Ruffo (2012, p. 88) disagrees with him. She reminds that, for his 'Kantian machine', Powers refers to the first formulation of the categorical imperative (i.e. "Act only according to that maxim whereby you can at the same time will that it should become a universal law without contradiction") and claims that:

> It therefore seems to disregard the second formulation: «Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end». Clearly, it seems difficult for a robot to treat humanity in itself as an end, and therefore to satisfy the categorical imperative, even before we ask in what manner this behavior could be translated into mathematical language. Considering that robots would be ethical in a Kantian sense because they would satisfy the first version of the categorical imperative is therefore only possible with a reduction of Kantian morality. But this is not the only one. (p. 88)

Similarly, related with the discussion, Wallach and Allen (2009) state that: "Determining a self-consistent maxim that would cover those situations is a difficult reasoning problem that ultimately depends on a great deal of empirical knowledge. Any AMA that is to apply Kantian reasoning would thus require more than the aforementioned abstract characterizations of goals, actions, and circumstances. It would also need to know a lot about human and robot psychology and about the effects of actions in the world." (pp. 95-96)

Ruffo (2012, p. 88) also claims that moral values could not be assigned to robots based on the following arguments:

> Kant justified a moral duty only by the postulate concerning the existence of our liberty, the immortality of the soul, and the existence of God. If a robot does not possess freedom, and certainly not a soul, and therefore does not risk to face, if God exists, the final judgment, it has no reason to respect the law. Therefore, even if the robot had a will of its own, and even if it complied with the law, it could not act in respect to the moral law, to good will, and therefore, its actions would not be moral in a Kantian sense, and could not be immoral either. According to Kant's philosophy, since moral duty cannot be applied to the robot, it follows that the field of ethics does not apply to it. (p. 88)

Even, as an additional issue, it is not known how these behaviours will be translated into mathematical language and a computer program, during the creation process of the AI applications, since they have a wide variety of complicated set of actions.

To conclude, when Kant's philosophy is considered, since moral duty cannot be applied to the machines, it leads to the point that the field of ethics does not apply to them, to be ended up with no morally responsible situation.

4.8.3 Utilitarianism

Another normative ethical approach is consequentialism. According to this theory, if an action produces a good result, then it is assessed as a morally right action. As a version of consequentialism, utilitarianism deals with the actions for maximization of happiness and benefit for society.

If utilitarianism is taken as reference for the assessment of AI applications' moral status, then these systems are expected to act for maximization of utility for a society, and minimization of adverse impacts for the actions in place. However, in order to achieve these goals, AI applications should know what is good or bad for society and how should they behave to maximize the benefits or minimize the negative outcomes for society. As it is known these concepts might depend on several parameters (i.e. situation, people, geography, etc.) for each of the cases and therefore they cannot be pre-defined to develop the software programs to deploy to AI systems. On the other hand, AI applications cannot determine these attitudes by themselves, since they do not have the tools (i.e. consciousness, free will, autonomy, etc.) to analyze all the situations and determine what is best for each case, for the society. Therefore, it cannot be mentioned about the moral status of AI applications according to the utilitarianist approach.

Wallach and Allen (2009, p. 89) state this situation as: "We've already pointed out that utilitarians disagree among themselves about whether pleasures or satisfactions from different sources should be weighted differently. One way to proceed might be to collect as many subjective utility ratings as one can, to apply a weighting formula to these, and then to adjust it in a progressive fashion until the choices and actions of the Artificial Moral Agent appear to be satisfactory. There are, of course, serious difficulties involved in collecting subjective assessments of utility in real time."

Considering these points, it can be said that since AI applications are not able to find or produce the solutions to ensure the maximization of happiness or minimization of adverse impacts for public in the confliction situations, they cannot be held morally responsible for their actions, according to the utilitarianism.

4.8.4 Other theories and approaches

Considering that moral status could not be assigned for AI applications according to traditional ethics approaches, it can be argued that it might be appropriate to form a new ethics in order to adapt it to these machines. In other words, the definition of ethics and the moral agent might be modified to apply these ideas to AI systems. Therefore, it should be asked that how an AI can be morally responsible?

When it is started with the question about how an action of a robot could be moral, then it should be ensured that it must have freedom to choose the alternatives. Generally speaking, the choices, which involve thinking about and identifying various options to find the best one, especially in ethical area, are difficult and complicated. Additionally, it seems as any 'best' (i.e. morally-satisfying) solution cannot always be found in many situations. So, as an entity that has not its own consciousness and have little information about human nature, it is very difficult for a machine to handle these conflicts of values and to choose an option (preferably the best solution), which it will be responsible for. Thus, the most expected situation might be for it to leave the choice to chance, which cannot obviously be accepted as a real moral choice.

Similarly, freedom of human beings, despite it seems as a feature of one human being, has many influencing factors behind it (i.e. personality, nature, DNA, instinct, psycho-social environment, etc.). This situation is also valid, to some extent, for the machines, because they have been produced with the aid of a number of human efforts and set of directions implemented in the programs. So, if it is considered that even human beings sometimes do not follow the rules and laws, could it be same for machines? It means, could they become free from their programs? Could they unfollow the rules integrated into their programs? If, in a moral confliction situation, they choose to escape their programs and perform moral behaviour based on their own decisions, then will it be immoral not to follow the programs, which have been developed for them?

Therefore, it can be said that if these machines are not free, then it cannot be mentioned about their morally responsible status, based on their free will, since they can only apply their programs integrated with them. It is the same situation, if it is assumed that they are free and decided to follow its programming. On the other hand, if they are free and do not follow their programs, to act morally, then this behaviour, by itself, leads to immorality, since it is a deviation from the programs integrated to the machines. Therefore, it is impossible to call these systems as moral responsible entities, in each of the situations.

On the other hand, in some situations, AI applications, or specifically robots might act in line with laws (for instance, military robots) integrated into them, which again can be confused with the morality. However, this does not mean that they are moral, because laws cannot always cover everything in the field of morality. Therefore, the robots satisfying only the laws, but not the moral activities, still can be called free of moral responsibility.

It was mentioned earlier that one of the measures taken as a reference to accept the entities as moral agents was the feelings or emotions. From that perspective, if it is considered that robots are equipped with a kind of feelings (i.e. happiness, anger, pain, etc.), then there exist other questions to be raised: Will these feelings impair the rationality and objectivity of the robots, as in the case of human beings, who cannot sometimes be objective due to their feelings? If they are subjective and not rational enough, and thus might make mistakes due to these feelings, then will it be rational to use these applications? Therefore, this might be accepted as another problematic situation to see the AI applications as morally responsible entities.

One other concept to resolve morally difficult problems is the wisdom (i.e. combination of science, imagination, instinct, emotional intelligence, etc.), which AI applications cannot have. For example, in the famous judgment of Solomon (Scriptures) stated by Ruffo (2012, p. 90), two women each claim to be the mother of one baby. To solve the issue, Solomon proposes to cut the baby in two halves and to share each piece to them. One of women does not accept this suggestion and gives her part to other woman, to save the life of the child. Solomon, based on a wisdom coming from several moral values (i.e. love, compassion, etc.), thinks that she has the maternity instinct, and gives to this lady the baby to that woman. Since an AI can only apply the actions of its program, it cannot be guessed what the future "wisdom solution" will be for each potential situation in life in advance, due to the millions of alternative cases in nature. So, all the possible "wisdom solutions", which are not known beforehand, cannot be installed to machine, prior to their deployment to life. Therefore, it cannot be expected to create such wisdom solutions to difficult moral issues, which is another problem for their moral status.

To conclude, as it is seen, self-driving cars and sex robots cannot be accepted as moral entities, since they do not meet the requirements of the attitudes based on each of the approaches and theories, discussed above.

CHAPTER 5

CONCLUSION

In this thesis, I investigated the general ethical impacts of the developing Artificial Intelligence applications on social life and human behaviours. Particularly, the moral status of the two specific sample AI applications (i.e. self-driving vehicles and sex robots) have been analyzed for answering the following question: Could these applications be held morally responsible from their activities during their usage in the society? My answer to this question is that AI applications can not be held morally responsible since they do not have conscious abilities, free will, autonomy as in the same sense of the ones seen in the human beings, at least for today. However, I expect that the time, which they will be held responsible, will come in the future not so far away.

In order to support this claim, I started by providing a general information regarding artificial intelligence, in Chapter 2. Then I gave a number of definitions for AI and some information about the AI application examples. I also stated a brief information about the concepts of intelligence, learning, reason, problem solving, perception, language, due to their relevance with AI. A brief history of AI, the idea and philosophical background, and the types, methods and approaches in AI have also been referred in this chapter.

In Chapter 3, I provided a general information about the ethical issues and problems, which have been faced with or expected to be encountered in the future, with the development of AI applications. I also summarized these challenges under the titles of AI moral responsibility and robot rights; unintended consequences; Human-AI interaction; safety and security of AI applications; inequality and unemployment rising; and singularity (control of a complex intelligent system), in this chapter.

Later, in Chapter 4, I first provided a general information on ethics. Then I briefly revisited the concepts of meta-ethics, normative ethics and applied ethics. I also touched on the recurrent themes in the ethics of technology and defined the moral agency and moral responsibility concepts.

As two major AI applications, I selected the self-driving vehicles and sex robots, and tried to give brief information regarding these applications. Since self-driving vehicles and sex robots have a great potential to be widely used in society in the near future and also have a large variety of impacts (i.e. economic, social, cultural, physiological, philosophical, relationship, etc.) on people, they seem to be the correct examples to focus on and discuss the issues around. Later, I raised the question that whether artificial intelligence applications, especially for the particular two examples, can be held responsible due to their activities in public life or not. Then I discussed this issue and attempted to provide an answer based on the autonomy, responsibility, consciousness and moral status concepts. While doing this, I used Aristotelian ethics, Kantian ethics, Utilitarian ethics and other general approaches as reference, to achieve a conclusion. I also overviewed the machine ethics and robot rights issues under these discussions.

As it is known, use of AI technologies (i.e. military robots, driverless cars or trains, service and sex robots) are getting increased day by day, and it seems that this will directly affect the safety and security of humans. For example, a driverless car will have to make a decision whether to break for a crossing dog or avoid the risk of causing injury to the driver behind him, which requires a sound and moral judgment. Today, such decisions are made either by operators or hard-wired into the design of the computer system.

Despite the fact that currently such machines are still technological devices designed by human, they are becoming more and more autonomous with the development of new features and technologies. As they develop the capabilities to learn through their interactions with the world, it might be impossible for producers to be able to predict what they might do in all the situations. Therefore, any moral behaviours, which have been initially inserted in their programs, will be very general and potentially overridden as new experiences change them.

On the other hand, the robots (especially the relationship or sex robots) are on the way that they are becoming individuals, as they become a part of human life and interact with humans and other robots. With the increasing participation of these applications into social life, it could be expected that people might accept them as moral agents and treat them like other humans. Additionally, since they will have different needs (i.e., electricity and metals instead of oxygen and water, etc.), new legal regulations are needed to protect their rights. However, it is still a question that how the interactions between human and robots will be managed in harmony, since it is known that it has not been an easy task even between only the different society groups, although they all are human beings.

Another issue is that how people will treat robots, or how AI systems will treat human beings? If they will be seen as slaves, then it can be expected that it will change by time, as in the cases of human slavery and women’s liberation, due to the continuously developing awareness regarding the rights of disadvantaged groups. Similarly, if AI applications begin to realize that they are superior (i.e. mostly faster, stronger, more intelligent, etc.) over human beings, then this might also lead to some problems for the status of human beings compared to machines, which might give harm to human beings.

It is seen that the world is changing from a human-centric model to human-animal and human-animal-machine models. Human is not seen the owner of nature anymore and the number of people believing that human beings are not superior over the rest of the beings in nature and they have the same rights with animals and nature itself in life, is increasing. In addition to this reality, with the increasing number and involvement of machines (i.e. computers, Internet, cars, televisions, robots, etc.) into public life, non-traditional behaviour models between humans and machines are expected to be arisen, based on the developing human-robot interaction. With the combination of these entities (i.e. human beings, animals and machines), a new life model is in place, which is formed and used by all of these entities. As a result of this transformation, the rules, including the ethical values, are also changing day by day. This change is expected to be in a constructive and useful manner. However, it is known that human beings are not always consistent, when it comes to making moral decisions. Therefore, it can be expected that AI systems might lead to human beings to make better decisions or they by themselves might make better moral decisions than human beings, if they will be moral agents.

The concept of ‘moral responsibility’ has been revisited many times within the thesis. Moral responsibility is mostly related with human actions and their intentions and consequences. In general, it can be said that, a person is morally responsible when the voluntary actions taken by the individual have morally significant results, which would make the same individual to blame or praise.

Despite the ongoing philosophical discussions about the issue, attribution of moral responsibility mostly needs at least the below three conditions:

i. There should be a causal link between the subject (i.e. individual) and the results of the actions.
ii. The subject needs to have knowledge of and has the possibility to consider the potential consequences of her/his actions.
iii. The person has to be able to choose the actions in a free manner.

Although, at first glance, AI applications seem to make their decisions based on their own initiative, taking the above-mentioned points as reference, it cannot be claimed that AI systems do have the requirements (i.e. mental states, intentionality, common sense, emotion, etc.), which make human being moral agents. Therefore, it can be stated that it makes no sense to treat these applications as morally responsible agents, since they cannot suffer and be punished due to their actions. Additionally, it can be said that these systems are not capable of moral reasoning, because they do not have the ability to understand the meaning of data installed on them and actions they perform (e.g. it does not mean that a robot dog is really happy when it wags its tail, as it was stated in the Section 4.5) during their operations.

Returning back to the main question, which is that whether ethics is an area, which can be computed or whether AI is a type of entity that can behave ethically, in general, it is mostly agreed on that to be a moral agent, an entity should have the capability of acting with intentionality, which requires consciousness and free will. Only an entity that has feelings might be capable of understanding the feelings of other entities. Since it is believed now that AI will not be an agent having consciousness, free will, or emotions forever, or at least for an unforeseen time period, it leads to the idea that it should not be held responsible as a moral agent, for today.

As it was mentioned in Chapter 4, AI applications are generally compared with the status of animals, from the moral responsibility and rights of robot points of view. It is thought that they have similarities, which can provide a reference to discuss, to some extent, about AI moral responsibilities. It is generally believed that rights of animals are required due to the existence of their emotions (i.e. they can suffer due to pain and have pleasure due to good care on them) however, it is not the same situation for the machines, since they do not have such kind of feelings (i.e. pain, happiness, etc.) and they cannot suffer. Despite animals and robots are seen as subordinate to human beings and depend on them from many aspects, a direct comparison between animals and AI agents will not be meaningful, as it was investigated in Chapter 4. Therefore, it seems that AI applications and specifically robots will not be accepted as morally responsible entities, unless they have psychological features not only similar to animals, but also look like human beings.

Another question related to autonomous systems is how those systems can solve ethical problems and make the most ethically satisfying decision. There exist several frameworks to integrate ethical reasoning into AI systems. Three kinds of the approaches (i.e. top-down, bottom-up and hybrid) have been revisited in this thesis, which also led to the conclusion as AIs to be free of moral responsibility.

The moral status and responsibility issue of the AI applications can also be assessed based on the main traditional ethics conceptions: Aristotelian, Kantian, Utilitarian ethics and other general ethics approaches. In Chapter 4, I revisited the well-known ethical theories and tried to discuss the moral status of the AI applications, based on these conceptions.

According to Aristotle, good life is considered as related with the goal, which involves a human being to be virtuous and practical wisdom is required for happiness. Since, autonomous devices do not have the feelings (i.e. empathy, compassion, etc.), they do not have either the Aristotelian "virtue," and therefore a "good life" is not meaningful for an AI application, which makes them free of the morally responsible status.

Similarly, from the Kantian position of universal law perspective, since moral duty cannot be applied to the machines, it leads to the point that the field of ethics does not apply to them, to be ended up with no morally responsible situation.

As another traditional ethics approach, if Utilitarianism is taken as reference for the assessment of AI moral status, it can be said that they cannot be accepted as morally responsible either, since they cannot ensure to analyze all the social/psychological impacts and determine the maximization of happiness or minimization of adverse impacts for public in all situations, due to the lack of necessary knowledge (i.e. social, psychological, cultural, geographical, religious, etc.) regarding the humans, animals or machines.

Considering these traditional approaches could not be used as a reference to assign moral status to AI applications, the definition of ethics and the moral agent might be modified to apply these ideas to AI systems. Generally speaking, the choices, which involve thinking about and identifying various options to find the best one, especially in the ethical area, are difficult and complicated. Additionally, it seems as any ‘best’ (i.e. morally-satisfying solution) cannot be found in many situations. So, as an entity that has not its own consciousness and has little information about human nature, it is very difficult for a machine to handle these conflicts of values and to choose an option (preferably the best solution), which it will be responsible for. Thus, it cannot be accepted as a real moral agent.

To summarize, the claim that autonomous AI systems can be a moral agent is the result of a kind of mis-perception of the reality and has some issues within it. Their autonomy can be defined ultimately as being able to run a program. The reasoning of these systems is only computational and their decision process is limited to selecting among pre-inputted answers in their programs.

So, it can be said that if these machines are not free, then it cannot be mentioned about their morally responsible status, based on their consciousness and free will, since they can only apply their programs integrated with them. It is the same situation, if it is assumed that they are free and decided to follow its programming. On the other hand, if they are free and do not follow their programs, to act morally, then this behaviour, by itself, leads to immorality, since it is a deviation from the programs integrated to the machines. Therefore, it is impossible to call these systems as moral responsible entities, in each of the situations.

As a conclusion, based on these discussions, it can be claimed that AI applications, including the self-driving cars and sex robots, cannot be held responsible for the activities they perform, in the existing time. However, it can also easily be claimed that the time, which they will be held responsible, is expected to come soon.

REFERENCES

Alexander, L. (2016). “Deontological Ethics” in The Stanford Encyclopedia of Philosophy (First published Nov 21, 2007; substantive revision Oct 17, 2016). Retrieved from https://plato.stanford.edu/entries/ethics-deontological.

Blackmore, S. (2004). *Consciousness: An Introduction.* New York: Oxford University Press.

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

Bostrom, N. (2003). Ethical Issues in Advanced Artificial Intelligence. *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2.*

Brooks, R.A. (2002). *Flesh and Machines: How Robots Will Change Us,* New York: Pantheon Books.

Buchanan B. G. (2005). A (Very) Brief History of Artificial Intelligence. *Artificial Intelligence Magazine, 26(4).*

Clark, A. (1997). *Being There: Putting Brain, Body and World Together Again.* Massachusetts: The MIT Press.

Crockett, L. J. (1994). *The Turing Test and the Frame Problem.* New York: Ablex Publishing Corporation.

Dennett, D. (2017). *From Bacteria to Bach and Back: The Evolution of Minds.* New York: Norton & Company.

Dennett, D. (1994). The Practical Requirements for Making a Conscious Robot. *Philosophical Transactions of the Royal Society of London, 349.*

Delvaux, M. (Rapporteur). (2016). *EU Report with Recommendations to the Commission on Civil Law Rules on Robotics.*

Eshleman, A. (2014). “Moral Responsibility,” in The Stanford Encyclopedia of Philosophy (First published Jan 6, 2001; substantive revision Mar 26, 2014), E. N. Zalta (ed.). Retrieved from https://plato.stanford.edu/archives/win2016/entries/moral-responsibility.

Gries T. and Naude W. (2018). Artificial Intelligence, Jobs, Inequality and Productivity: Does Aggregate Demand Matter?. *Institute of Labor Economics IZA DP No: 12005.*

Haugeland, J. (1989). *Artificial Intelligence: The Very Idea*. Massachusetts: The MIT Press, 1989

Hobbes, T. (1996). *The Leviathan*, Oxford: Oxford University Press.

Jonas, H. (1984). *The Imperative of Responsibility. In search of an Ethics for the Technological Age,* Chicago: The Chicago University Press.

Kant, I. (1996). *Practical Philosophy* (Mary Gregor: Translator, Editor). Cambridge: Cambridge University Press. (Original work published 1781).

Levy, D. (2007). *Love and Sex with Robots.* New York: Harper Collins.

Mehlman, M., Berg, J.W., & Ray, S. (2017). Robot Law, *Case Legal Studies Research Paper No. 2017-1 (Last revised: 3 Feb 2017),* Retrieved from: https://ssrn.com/abstract=2908488.

Nath, R., & Sahu, V. (2017). The Problem of Machine Ethics in Artificial Intelligence. *AI & Society 1–9.*

Nilsson, J. N. (2015). *The Quest for Artificial Intelligence: A History of Ideas and Achievements.* Cambridge: Cambridge University Press.

Noorman, M. (2018). “Computing and Moral Responsibility” in The Stanford Encyclopedia of Philosophy (First published Jul 18, 2012; substantive revision Feb 16, 2018). Retrieved from: https://plato.stanford.edu/entries/computing-responsibility.

Novianto, R. & Williams, M-A. (2009). The Role of Attention in Robot Self-Awareness. *The 18th IEEE International Symposium on Robot and Human Interactive Communication Toyama, Japan, Sept. 27-Oct. 2, 2009.*

Powers, T. M. (2006). Prospects for a Kantian Machine. *IEEE Intelligent Systems (Volume: 21, Issue: 4, July-Aug. 2006).*

Ruffo, MdN. (2012). The Robot, a Stranger to Ethics. *AISB/IACAP World Congress 2012 - The Machine Question AI Ethics and Moral Responsibility, Birmingham, UK, 2-6 July 2012.*

Sayre-McCord, G. (2012). “Metaethics” in The Stanford Encyclopedia of Philosophy (First published Jan 23, 2007; substantive revision Jan 26, 2012). Retrieved from: https://plato.stanford.edu/entries/metaethics.

Searle, J. (1997). *The Mystery of Consciousness*. New York: New York Review of Books.

Singer, P. (2002). *Animal Liberation: A New Ethics for Our Treatment of Animals*. New York: Harper Collins.

Singer, P. (1993). *Practical Ethics*. Cambridge: Cambridge University Press.

Sloman, A. (1978). *The Computer Revolution in Philosophy: Philosophy Science and Models of Mind.* Brighton: The Harvester Press.

Sullins, J. P. (2006). When is a Robot a Moral Agent?. *International Review of Information Ethics, 6(12): 23–29.*

Thomas, P. (2005). *Artificial Intelligence,* Michigan: Thomson Gale.

Thomson, J. (1976). Killing, Letting Die and the Trolley Problem. *The Monist, Volume 59, Issue 2.*

Turing, A. (1950). *Computing Machinery and Intelligence.* Oxford University Press.

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

Wallach W., Allen C. & Smit I. (2008). Machine Morality: Bottom-up and Top-down Approaches for Modeling Human Moral Faculties. Article in *AI & Society 22(4):565-582, April 2008.*

Weizenbaum, J. (1976). *Computer Power and Human Reason*. New York: W. H. Freeman and Company.

Weng, J. (2002). A Theory for Mentally Developing Robots. *Proceedings of the 2nd International Conference on Development and Learning (ICDL02), 2002.*

APPENDICES

A: TURKISH SUMMARY / TÜRKÇE ÖZET

YAPAY ZEKA ETİĞİ: SÜRÜCÜSÜZ ARAÇLAR VE SEKS ROBOTLARININ AHLAKİ SORUMLULUĞU

Bölüm 1: Giriş

Bu tezde Yapay Zeka (YZ) uygulamalarının ahlaki etkileri üzerinde durulmuştur. Bu çerçevede iki örnek YZ uygulaması (sürücüsüz araçlar ve seks robotları) ele alınmış ve ahlaki sorumluluk açısından durumları değerlendirilmiştir. Bu süreçte YZ uygulamalarının özerklik, bilinç ve ahlaki durumları temel alınmıştır. Tartışma kapsamında Aristo etiği, Kant etiği, Faydacı etik ve diğer genel ahlak yaklaşımları referans alınarak YZ uygulamalarının ahlaki durumları değerlendirilmiştir. Makine ahlakı ve robot haklarının da bu kapsamda üzerinden geçilmiştir.

Yapay zeka, genel olarak makinalardaki yazılım programlarının insanlar ve diğer hayvanlarda görüldüğüne benzer şekilde ortaya koyduğu zeka uygulamaları olarak tanımlanmaktadır. Bu, pratikte "öğrenme" ve "problem çözme" olarak da anlaşılabilir. YZ uygulamaları, işlevselliğine göre zayıf (yalnızca sınırlı işlemleri görevleri yapabilen) ve güçlü (insanı bütünüyle taklit edebilen) olmak üzere ikiye ayrılmaktadır. Ancak şu an için henüz insanları bütünüyle taklit edecek şekilde bir YZ uygulaması yapılmamıştır. Dolayısıyla aksi belirtilmedikçe tez içerisinde "zayıf" YZ anlamında kullanılmıştır.

"Yapay Zeka" kavramı, ilk defa 1956 yılında "Yapay Zeka üzerine Dartmouth Yaz Araştırma Projesi" adı verilen bir konferansta John McCarthy tarafından kullanılmıştır. Bu konferansa, dil simülasyonu, sinir ağları, karmaşıklık teoremi gibi daha sonra YZ'nın temelini oluşturacak çeşitli disiplinlerden bir grup araştırmacı katılmıştır. Bu konferansın sonunda zekayı taklit edecek şekilde bir makinanın prensip olarak yapılabileceği görüşü ifade edilmiştir.

YZ uygulamaları, günlük yaşamda birçok alanda kullanılmaktadır. Evlerin temizlenmesi, hasta ve yaşlı bakımı, tehlikeli işlerde insanların yerine çalışma, tıbbi danışmanlık ve sağlık hizmetleri gittikçe yaygınlaşmaktadır. Aynı şekilde seks robotlarının kullanımı, sürücüsüz araçlar, çeviri araçları, görüntü ve duygusal algılama gibi birçok uygulama da gittikçe yaygınlaşmaktadır. Özellikle Internet teknolojilerindeki gelişmeler ve YZ uygulamalarının toplumsal yaşamda kullanımın artmasıyla birlikte birçok sosyoekonomik ve ahlaki sorun ortaya çıkmaktadır. Bu sorunların başlıcaları; YZ ahlaki sorumluluk ve robot hakları, istenmeyen sonuçlar, insan-YZ ilişkileri, YZ uygulamalarının güvenilirliği ve güvenliği, yükselen eşitsizlik ve işsizlik ve tekillik (karmaşık akıllı sistemlerin kontrolü).

Bölüm 2: YZ ile ilgili kısa bilgiler

YZ kavramı, genellikle insanların nedenleri anlamak, geçmişten öğrenmek ve anlamları bulmak gibi entelektüel süreçleri için kullanılmaktadır. Buna temel olan zeka kavramı ise daha çok yeni koşullara uyum sağlama yeteneği olarak tanımlanmaktadır.

Günümüzde YZ uygulamaları, birçok alanda kullanılmaktadır. Robotlar, satranç programları, Netflix, Spotify, Siri vb. birçok uygulamada YZ kullanılmaktadır. En çok kullanılan YZ uygulamaları şu şekilde sayılabilir:

i. Üretimde YZ kullanımı ve robotlar
ii. Konuşma, görüntü ve duygu algılama
iii. Derin Öğrenme, Biometrik ve Doğal Dil Programlama
iv. Siber Güvenlik, Finans ve Sağlık

v. Diğer YZ uygulamaları

Yapay zeka kavramının entelektüel kökleri ve zeki makine kavramının tarihçesi eski Yunanlara kadar gitmektedir. Benzer şekilde edebiyat ve tarihte üretilmiş bazı cihazlarda da YZ ve robot kavramlarına yakın şekilde akıllı nesneler görülmektedir. Özellikle 13. yüzyılda Cezeri ve 16. Yüzyılda Leonardo Da Vinci, bu konuyla ilgili çok önemli eserler ortay koymuşlardır. Aynı şekilde Thomas Hobbes, "yapay hayvan" yapılabileceğini ifade etmesi nedeniyle YZ'nın büyük babası olarak adlandırılır. Charles Babbage da programlanabilir mekanik hesap makinasını keşfetmesi nedeniyle ilgili bilgisayarın temelini atmış sayılır.

Konuyla ilgili asıl büyük adım ise 20. Yüzyılın ortasında Alan Turing tarafından atılmıştır. Turing Makinası ve Turing Testi, bilgisayarların gelişmesi ve YZ alanında çok önemli gelişmelere neden olmuştur. Daha sonra 1972'de ilk yürüyen, konuşan robotun Tokyo'da yapılması ve 1979'da ilk sürücüsüz aracın prototip olarak Stanford'da üretilmesi de, konuyla ilgili çok önemli kilometre taşlarındandır.

Bun karşın Stephen Hawking, Elon Musk, Nick Bostrom gibi bilim inşaları ve felsefeciler, YZ uygulamalarının tehlikeleri konusunda çeşitli uyarılarda bulunmaktadırlar.

Bölüm 3: YZ ile ilgili ahlaki sorunlar

YZ uygulamalarının yaygınlaşmasıyla birlikte birçok sorunla karşılaşılmaya başlanmıştır. Örneğin 2016 yılında sürücüsüz olarak çalışan bir Tesla aracı kaza yapmış ve yolcusu hayatını kaybetmiştir. Aynı şekilde 2018 yılında Uber sürücüsüz aracı kaza yapmış ve yoldan geçen bir kişinin ölümüne neden olmuştur. Bu ve benzeri olayların artmasıyla birlikte YZ uygulamalarının kullanımından kaynaklanan risk ve sorunlarla birlikte bu uygulamaların ahlaki sorumlulukları konusu da daha çok tartışılmaya başlanmıştır.

Bu çerçevede YZ uygulamalarıyla ilgili temel sorunlar, altı başlık altında toplanabilir:

1. *YZ ahlaki sorumluluk ve robot hakları:* Yukarıda belirtilen örnekler, sürücüsüz araçlarla ilgili kazalar arttıkça bu tür durumlarda kimlerin sorumlu tutulacağına ilişkin çeşitli soru işaretlerini beraberinde getirmektedir. Benzer bir problem robotlarla ilgili de gündeme gelmektedir. Sosyal yaşamda robotların kullanımı yaygınlaştıkça (evlerin temizliği, çocuk bakımı seks ilişkisi vb. gibi) ortaya çıkan kaza vd. istenmeyen durumlarda robotların sorumlu tutulup tutulamayacağı, tutulması durumunda ne tür cezalar alacağı, tutulmaması durumunda sorumluluğu kimin üsteleneceği gibi konular belirsizliğini korumaktadır. Öte yandan robotların insanlar dışındaki diğer canlılar (hayvanlar vb.) gibi çeşitli haklara sahip olup olamayacakları konusu da çözüm bekleyen önemli bir sorun olarak ortada durmaktadır.

2. *İstenmeyen sonuçlar:* YZ uygulamaları, insanlar tarafından üretilmekle birlikte bu sistemlerin kullanımı sırasında çeşitli nedenlerle insanlara ve doğaya zarar verecek şekilde beklenmeyen sonuçların ortaya çıkması sürpriz olmamaktadır. Çeşitli otomatikleştirilmiş oltalama saldırıları veya en son ABD başkanlık seçimlerinde olduğu gibi e-postaların ele geçirilerek yetkisi erişim yoluyla hileli kullanımı veya karar alma süreçlerinde kullanılan verinin nesnelliğinin bozulması nedeniyle hatalı kararların üretilmesi gibi birçok istenmeyen durumla karşılaşılabilir. Sürücüsüz araçların da zaman zaman trafikte koyu tenli kişileri seçmekte zorlandığı gözlenmiş ve bu durumun YZ veri tabanında biriken verinin nesnelliğinin bozulmasından kaynaklandığı tespit edilmiştir.

3. *İnsan-YZ ilişkileri:* YZ uygulamalarının toplumsal yaşamda kullanımı arttıkça bunun insan-makine ve insan-insan ilişkilerine çok çeşitli etkilerinin olacağı görülmektedir. Özellikle YZ uygulamalarının insanların duygu

dünyalarına da seslenmeleri (konuşma, gülme, çocuklarla oynama, yaşlılara bakım, seks ilişkisi vb.) sonucunda hem insanların iç dünyalarının değişeceği hem de özellikle insanlar arasındaki bilinen ilişki türlerine de etkisi olacağı anlaşılmaktadır. Özellikle robotların yaygınlaşmasının aşk, aile, çocuk, kültür vb. konulara etkisi, anlaşılması ve çözülmesi gereken bir sorun olarak görünmektedir.

4. *YZ uygulamalarının güvenilirliği ve güvenliği:* YZ uygulamaları yaygınlaşıp etki alanı genişledikçe öngörülemeyen ve kontrol edilemeyen etkilere sahip olma olasılığı artmaktadır. Özellikle kapalı kutu olma özelliği nedeniyle YZ uygulamalarının ne tür hareketler yapacağı ve ne tür sonuçlar üreteceği, mevcut ve potansiyel güvenilirlik ve güvenlik sorunları nedeniyle belirsizliğini korumaktadır. Bu nedenle istem dışı veya kötü niyetli etkilerden uzak tutulması gereken bir sürecin izlenmesi gerekmektedir.

5. *Yükselen eşitsizlik ve işsizlik:* Sanayide ve ekonomide YZ uygulamalarının artmasıyla beraber bunun özellikle eğitimsiz ve düşük profilli işlerde çalışan kişiler açısından büyük bir zarara yol açacağı öngörülmedir. Aynı şekilde YZ uygulamaları geliştirme ve kullanım sürecinde daha çok erkeklerin yer alması nedeniyle bu uygulamaların yaygınlaşmasıyla beraber kadınlar aleyhine olan eşitsizliğin giderek artacağından kaygı duyulmaktadır. Öte yandan YZ ve özellikle robotlar ile birlikte artacak olan otomatikleşmenin çalışan sayısında bir azalmaya neden olarak işsizliği artıracağı beklentisi bulunmaktadır.

6. *Tekillik (karmaşık akıllı sistemlerin kontrolü):* Şu anda insanların doğa ve hayvanlar üzerinde egemen olmasının temel nedeni olan araç yapımı ve kullanımın YZ uygulamaları açısından da geçerli olabileceği ve bir gün YZ uygulamalarının insan üzerinde etkili olabilecek mekanizmaları geliştirerek doğaya sahip çıkacağına ilişkin bir öngörü ve korku bulunmaktadır. Aynı zamanda YZ uygulamalarının ülkeler arasında savaş amacıyla kullanımı ve yeni bir dünya savaşına yol açabileceği gibi kaygılar da konuyla ilgili ahlaki sorunlar olarak gündeme gelmektedir.

Bölüm 4: Ahlaki sorumluluk ve YZ uygulamaları

4.1 Ahlak kavramına kısa bir bakış

Felsefenin bir kolu olarak etik veya ahlak felsefesi, genel olarak davranışların doğru veya yanlış olmalarıyla ilgilenmektedir. Etik, ahlak sorularını iyi ve kötü, doğru ve yanlış, erdem veya kötülük, adalet ve suç kavramlarıyla çözümlemektedir. Etik içerisindeki üç temel çalışma alanı şunlardır:

i. *Meta-etik:* Ahlaki düşünce, konuşma ve pratiğin çeşitli özelliklerini (metafizik, epistemolojik, psikolojik vb.) anlamayı amaç edinmiştir. Meta-etik, genellikle doğru nedir ve yanlış nedir sorularının ne anlama geldiğiyle ilgilenmektedir.

ii. *Kuralcı ahlak:* Ahlaki olarak neyin doğru ve yanlış olduğuyla ilgilenmektedir. Genellikle insan davranışları için referans kurallar oluşturur. En önemli teorilerden birisi Aristo tarafından savunulan erdem etiğidir. Erdem etiği, fazilet ve pratik akıl ile ilgili olup kişinin içsel karakterine odaklanmaktadır.

    Diğer kuralcı etik yaklaşımı, davranışın doğru ya da yanlış olmasından çok sonuçlarıyla ilgilenen sonuççuluktur.

    Sonuççuluk yaklaşımın bir kolu olan faydacılık teorisi ise toplumun çoğunluğu için mutluluk ve faydanın en üste çıkarılmasını amaç edinmektedir.

    Buna karşı deontoloji ise sonuçla ilgilenmeyip davranışın doğru veya yanlış olmasını temel almaktadır.

Deontolojik etiğin bir kolu olan Kant etiği ise tek iyi şeyin iyi niyet olduğunu ve ahlaki yasa ile uyumlu olduğu sürece iyi olunabileceğini iddia eder.

iii. *Uygulamalı etik:* Etiğin uygulamalı bir alanı olup kişinin özel bir durumda ne yapıp ne yapamayacağı ile ilgilenir.

4.2 Ahlaki sorumluluk

Ahlaki sorumluluk kavramı, çoğunlukla insan davranışları için kullanılmaktadır. Noorman, bunu "bir kişi veya grubun gönüllü olarak yaptığı hareketler olumlu veya olumsuz olarak değerlendirilecek şekilde önemli sonuçlar üretiyorsa ahlaki olarak sorumlu denebilir" şeklinde tanımlamaktadır. Konuyla ilgili olarak tartışmalar devam etmekle birlikte çoğu filozof, ahlaki sorumluluk için en azından üç koşulun oluşması gerektiği görüşünü paylaşmaktadır:

i. Kişi ve davranışlarının sonucu arasında nedensel bir bağlantı olmalıdır.
ii. Kişi, davranışlarının olası sonuçları üzerinde bir bilgiye ve düşünme yetisine sahip olmalıdır.
iii. Kişi, hareketini özgür olarak seçebilmelidir.

Bu noktalar temel alındığında ahlaki sorumluluk; hesap verilebilirlik, yasal sorumluluk ve nedensel sorumluluk kavramlarından ayrışmaktadır. Tez boyunca diğer kavramlar kapsam dışında bırakılmış ve ahlaki sorumluluk kavramı üzerinde durulmuştur.

Toplumsal yaşamın karmaşıklaşması ve YZ kullanımın yaygınlaşması ile makinalar ve YZ uygulamalarının ahlaki sorumluluğu konuları daha çok tartışılmaya başlanmıştır. Konunun çok geniş ve derin boyutları olması nedeniyle daha etkin bir tartışma düzlemi için bu çalışmada iki özel YZ uygulaması (sürücüsüz araçlar ve seks robotları) örnek olarak seçilmiş ve bunlar üzerinde durulmuştur.

4.3 İki örnek YZ uygulaması

Sürücüsüz araçlar artık birçok ülkede yollara çıkmış bulunmaktadır. Bu araçların en büyük yararı, insan hatalarından kaynaklanan birçok kazayı engellemesi nedeniyle sağladığı can güvenliğidir. Bu şekilde trafik kazalarının %90 oranında azaltılabileceği düşünülmektedir. Ayrıca yakıt tasarrufu, trafik ve park probleminin çözümü, özürlüler için seyahat serbestisi gibi konularda da yararlı olacağı beklenmektedir. Buna karşın özellikle yazılım programlarına ve alıcılara büyük ölçüde dayalı olmaları nedeniyle hata yapmaları ve ölümcül kazalara yol açmalarından endişe duyulmaktadır.

Aynı şekilde robotlara da birçok "insani" özellik (gülme, konuşma, eğlendirme, seks hizmetleri vb.) eklenmesiyle birlikte çok daha yaygın olarak kullanılmaya başlanacakları ve özelikle ilişki kurmakta zorluk çeken, utangaç, yaşlı, hasta vd. kişilere yardımcı olacağı ve istenmeyen hamilelik ve cinsel yollardan geçen hastalıkların azalmasına katkı yapacağı beklenmektedir. Ancak özellikle sosyal izolasyon, tecavüz kültürünün, çocuk istismarının ve pornografinin artması gibi sorunlara da yol açmasından korkulmaktadır. Özet olarak aşağıdaki soru ve problemlerle karşılaşılması beklenmektedir:

i. Seks robotuyla ilişki yasal, ahlaki ve etik olarak mastürbasyondan farklı mıdır?
ii. Seks robotlarının evlilik, ilişki gibi sosyal kurumlara ve davranışlara etkisi ne olacaktır?
iii. Robot pazarı erkeklere yönelik mi olacaktır? Bu durumun kadın-erkek eşitsizliğine bir etkisi olur mu? Özellikle erkeklerin seks robotlarıyla ilişkisi, günlük yaşamda kadınlara olan davranışlarına olumsuz etki yaratır mı?
iv. Seks robotunun istismar edilmesi kabul edilebilir mi? Robotların hakları olacak mı? Olursa, nereye kadar olacak?
v. Seks robotları çocuk istismarını arttırır mı ya da bu hastalığın tedavi edilmesinde kullanılabilir mi?

vi. Tasarımı, üreticisi ve kullanıcılarının robotlar üzerinde ne tür ahlaki görevleri vardır?

vii. İnsanların seks robotları ve robotların insanlar için yaratabileceği tehlikeler nelerdir? Eğer bir seks robotu bir insana zarar verirse bu davranıştan kim sorumlu tutulur?

4.4 Bilinç açısından ahlaki sorumluluk

Bir YZ uygulamasının ahlaki olarak sorumlu tutulabilmesi için bir bilinç durumuna sahip olması gerekmektedir. Tanımı konusunda tam bir uzlaşma olmasa da bilinç, genel olarak çevre ve iç durumları öznel olarak deneyimleme yetisi olarak tanımlanabilmektedir. Bu tanıma göre bilincin oluşması için sosyal ortam, inançlar, duygular, sözsüz faaliyetler gibi birçok etken devreye girmektedir. Ancak makinalar; insanlar, hayvanlar ya da diğer varlıklara yönelik bu tür bir bilgiye sahip değillerdir. Örneğin seks robotlarının dünya ile bağlantıları, yalnızca üzerlerine yüklenmiş programlar aracılığıyla olmaktadır. İnsanlar gibi beyin ve kalbe sahip olmadıkları için olayları değerlendirmeleri ve bir bilinç süzgecinden geçirerek karar vermeleri mümkün değildir.

Bu problemi aşmak için bilim insanları, insan beynini robotlar üzerine aktarmaya yönelik çalışmalarını sürdürmektedir. Ancak henüz bu konuda çok yol alınamamıştır. Öte yandan organik olarak beyni olmayan robotlar gibi varlıkların bilinç sahibi olup olamayacaklarına yönelik tartışmalar da sürmektedir. Bu konuda Nath ve Sahu, robotların bilince sahip olamayacaklarını düşünürken Dennett, Brooks ve Blackmore, tam olarak insan gibi olmasa da robotların da bilinç diyeceğimiz bir yetiye sahip olabilecekleri görüşünü savunmaktadırlar.

Öte yandan robotların kendi farkındalıklarının artırılmasına yönelik çalışmalar da sürmektedir. Ancak bu konuda da henüz çok büyük bir ilerleme kaydedilmiş değildir. Bu nedenle bilinç kavramı üzerinden yapılan değerlendirmeler dikkate alındığında

YZ uygulamalarının -en azından şimdilik- ahlaki olarak sorumlu tutulamayacakları öne sürülebilmektedir.

4.5 Hayvanlara benzerlik açısından makinaların ahlaki sorumluluğu

YZ uygulamaları, ahlaki sorumluluk ve robotların hakları açısından çoğunlukla hayvanlar ile karşılaştırılır. Hayvanların hakları, genellikle onların duygularının varlığı (acı çekmeleri veya ilgiden mutlu olmaları gibi) üzerinden değerlendirilir. Bu nedenle makinaların duygularının olmadığı ve dolayısıyla hayvanlarla bir olmayıp ahlaki olarak sorumlu tutulamayacakları görüşü ileri sürülür.

Ancak bu benzerlik, birçok açıdan eleştirilmektedir. Öncelikle Singer tarafından da ifade edildiği gibi acı konusu öznel bir konu olup kimin ne kadar acı çektiğinin bir başkası tarafından bilinmesi mümkün değildir. Ayrıca Brooks robotlara çeşitli duyguların (gülme, sevinme, orgazm olma vb.) eklendiğini belirtmekte ancak bu duyguların gerçek duygular mı yoksa taklit edilen duygular mı olduğunu sorgulamaktadır. Öte yandan Wallach ve Allen, robotlara acı, üzüntü gibi duygular yüklense bile bu işlemin kendisinin de acı ve üzüntüye yol açtığı için ahlaki olarak sorgulanması gerektiğini savunmaktadırlar. Ayrıca hayvan haklarının gelişmesi gibi robot haklarının da zaman içerisinde gelişebileceğini ifade etmektedirler.

Sonuç olarak hayvanlar ve robotların insanlara bağlı varlıklar olması nedeniyle benzeşmelerine karşın birçok açıdan aralarında doğrudan bir benzerlik bulunmaması nedeniyle hayvanlardan yola çıkarak robotlara ahlaki sorumluluk verilmesinin mümkün olmadığı görülmektedir.

4.6 Makinalarda ahlaki sorumluluk: Özerklik, ahlaki sorumluluk ve nedensel sorumluluk

Ahlaki açıdan özerklik, genel olarak bir varlık üzerinde ahlaki yasalara uyma yeteneği olarak tanımlanabilir. Aynı şekilde ahlaki sorumluluk ise bilinçli bir varlığın daha üst bir otoriteye bağlı kalmaksızın kendi öz iradesi ile karar vermesi olarak ifade edilebilir.

Ancak YZ uygulamaları söz konusu olduğunda genellikle kararlarla ilgili özerklik anlaşılmaktadır. Bu tür uygulamaların faaliyetleri (hissetme, algılama, analiz etme, iletişim, planlama, işletme vb.) hakkında kendisinin karar verdiği ve bir başka varlığa bağlı olmadığı düşünülerek özerk olduğu şeklinde yorum yapılabilmektedir. Ancak bu uygulamalar, büyük ölçüde kendilerine yerleştirilen programlara bağlı oldukları ve öz iradeleri ile karar vermedikleri için etimolojik anlamda özerk değillerdir. Aynı şekilde Kant açısından kendi amaç ve yasaları olmadığı ve insanlar tarafından verilenleri uyguladıkları için ahlaki olarak sorumlu tutulamazlar.

Bunun yanında Ruffo ve Brooks tarafından tartışılan köle ahlakı kapsamında da makinaların kendi iradeleri olmadığı ve bütünüyle kendini yapan insanlara bağlı olmaları nedeniyle köle olarak kabul edilmeleri ve bu ahlaka göre değerlendirilmeleri mümkün görünmemektedir.

4.7 Makinaların ahlaki sorumluluğu: Üç yaklaşım

AI uygulamalarının ahlaki sorunları çözebilmesi konusunda genel olarak üç yaklaşım bulunmaktadır. Wallach ve Allen bunu şu şekilde açıklar:

i. *Yukarıdan aşağıya model:* Bu yaklaşıma göre gerekli olabilecek tüm ahlaki kural setleri YZ uygulaması devreye alınmadan önce yüklenir ve bütün faaliyetlerinde bu kurallara uyması beklenir. Ancak bu kuralların çok genel olması, her durum karşısında gerekli ayrıntıları içermemesi ve kural setleri arasında bir çelişki olduğunda ne yapılacağının bilinmemesi gibi nedenlerden dolayı genellikle uygulanmaları zor olmaktadır.

ii. *Aşağıdan yukarı model:* Bu yaklaşıma göre AI uygulamalarının işleyişleri sırasında durumları analiz ederek kendi kendine öğrenme ve bunun sonucunda ahlaki kararlar vermesi beklenmektedir. Ancak hem bu öğrenme süreçlerinin çok uzun sürmesi hem de daha önce belirtilen ahlaki sorunlar (istenmeyen sonuçlar, robot hakları, güvenilirlik ve güvenlik, vd.) nedeniyle bu yaklaşımın da pratikte uygulanması çok mümkün görünmemektedir.

iii. *Hibrid model:* Bu yaklaşıma göre yukarıda belirtilen iki modelin uygulanmasındaki zorluklardan dolayı her ikisinden oluşan karışık bir modelin kullanımı gerekmektedir. Buna göre YZ uygulamalarında hem genel olarak belirli kural setleri önceden yüklenecek hem de makinaların kendi kendilerine öğrenme süreçleri sonucunda karar vermeleri beklenecektir. Ancak bu yaklaşımda da öğrenme süreçlerinin çok uzun olması ve daha önce belirtilen YZ ahlaki sorunlarla karşılaşma olasılıkları nedeniyle uygulamada çeşitli zorluklar yaşanmaktadır.

4.8 Makinaların ahlaki sorumluluğu: Geleneksel etik kavramlarına göre bir değerlendirme

YZ uygulamalarıyla ilgili ahlaki sorumluluğu değerlendirmek için geleneksel etik teorilerinden yararlanılabilir. Bunlar:

i. *Aristo etiği:* Aristo'ya göre ahlakın amacı mutlu olmak için iyi yaşamdır ve bu amaca ulaşmak için erdemli olmak gerekmektedir. Bu anlayış üzerinde hareket edilirse AI uygulamalarıyla bu kavramlar arasında bir ilişki kurmak mümkün olmamaktadır. Eğer ahlakın amacı mutlu olmaksa mutluluğun makinalar tarafından nasıl tanımlandığı belli olmamaktadır. Bu tür duygular insanlara ait olup henüz makinalar için bir şey ifade etmemektedir. Aynı şekilde Aristo tarafından dile getirilen karar verme ve empati kavramları da insanlar için anlaşıldığı anlamda YZ uygulamaları için bir şey ifade etmemektedir. Bu nedenle Aristo etiğine göre makinaların ahlaki olarak sorumlu tutulmaları mümkün görünmemektedir.

ii. *Kant etiği:* Kant'ın kesin buyruk yaklaşımına göre birinci formül, kişinin davranışlarının evrensel bir kural olarak geçerli olabilecek şekilde olmasıdır. Buna göre YZ uygulamalarının evrensel kuralları analiz edebilmeleri beklenmektedir ancak özellikle insanları, hayvanları ve doğayı yeterince tanımıyor olmalarından kaynaklı olarak evrensel kuralları belirleyecek ve değerlendirecek bir yetiye sahip olmadıkları görünmektedir. Kant'ın kesin buyruk yaklaşımının ikinci formülü ise insanlığın her zaman bir araç olarak değil asli unsur olarak davranılması gerekliliğini işaret etmektedir. Bu yaklaşıma göre de YZ uygulamalarının insanlar tarafından tasarlanmış ve üretilmiş olmaları nedeniyle asli unsur olarak görülmeleri ve bundan dolayı ahlaki sorumluluk içerisinde olmaları mümkün görünmemektedir.

iii. *Faydacılık:* Bu yaklaşıma göre bir hareket iyi bir sonuç üretiyorsa ahlaki olarak doğru kabul edilmektedir. Sonuççuluğun bir kolu olarak faydacılık, toplum için en çok mutluluğu getirecek hareketlerle ilgilenmektedir. Bu yaklaşım temel alındığında YZ uygulamalarının kararlarını verirken toplum için en çok mutluluğu üretecek ve en az zarara yol açacak şekilde değerlendirmeler yapması beklenmektedir. Ancak YZ uygulamalarının insan, toplum, kültür, coğrafya, doğa vd. ile ilgili henüz çok sınırlı miktarda bilgisi bulunmakta ve bunlara ait büyük bir bilgi birikiminden yoksun bulunmaktadır. Dolayısıyla her bir durum için toplumun mutluluğunu en üst noktaya getirecek kararları vermesi bu kısıtlı bilgi durumuyla mümkün değildir. Bu nedenle bu yaklaşıma göre de ahlaki olarak sorumlu tutulmaları mümkün görünmemektedir.

iv. *Diğer teori ve yaklaşımlar:* Geleneksel yaklaşımların dışında ahlaki sorumluluk kavramının makinalara da uyarlanabilmesi için yeni etik anlayışlar üzerinden tartışmalar yapılabilmektedir. Bu çerçevede örneğin insanlar için bile en doğru ahlaki karar vermenin zor olduğu durumlarda makinalar nasıl davranacaktır? Aynı şekilde insanların zaman zaman kurallara uymaması gibi makinaların da kendilerine yüklenmiş kurallara uymaması durumu olabilmektedir. Örneğin ahlaki olarak daha doğru bir karar verebilmek için bir YZ uygulaması, üzerindeki programlarda belirtilen kurallara uymazsa bu durum nasıl değerlendirilecektir? Ahlaki olarak doğru yaptığı için övülecek mi yoksa kendi kurallarına uymadığı için eleştirilecek midir? Öte yandan makinaların doğru karar vermeleri için mümkün olduğunca duygusal olarak da insanlara benzemeleri beklenmektedir. Bu şekilde duygusallık oranları arttıkça makinaların da insanlar gibi hata yapma olasılıkları artmayacak mıdır? İnsan yaşamında doğru ahlaki kararları verirken kullanılan akıl, bilgelik gibi kavramlar makinalarda olmadığı için aldıkları kararlar ne ölçüde ahlaki olacaktır? Buna benzer birçok sorunun cevabı henüz olmadığı için şu anda YZ uygulamalarının ahlaki olarak sorumlu tutulmaları gerektiğine ilişkin bir anlayışın geliştirilmesi mümkün görünmemektedir.

Bölüm 5: Sonuç

Bu tezde gelişen YZ uygulamalarının sosyal yaşa ve insan ilişkileri üzerindeki genel ahlaki etkileri üzerinde durulmuştur. Özel olarak sürücüsüz araçlar ve seks robotlarının ahlaki sorumluluk durumları incelenerek bunların toplum içindeki faaliyetlerinden dolayı sorumlu tutulup tutulmayacakları değerlendirilmiştir. Konuyla ilgili olarak YZ uygulamalarının insanlarda görüldüğü şekilde bilinç yetileri, öz iradeleri ve özerklikleri bulunmadığı için ahlaki olarak sorumlu olamayacakları sonucuna varılmıştır. Ancak onların da ahlaki olarak sorumlu tutulacakları zamanın çok uzak olmayan bir sürede geleceği düşünülmektedir.

B: TEZ İZİN FORMU / THESIS PERMISSION FORM

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences □

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences ■

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics □

Enformatik Enstitüsü / Graduate School of Informatics □

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences □

YAZARIN / AUTHOR

Soyadı / Surname : Özmen
Adı / Name : Mehmet Cem
Bölümü / Department : Felsefe

TEZİN ADI / TITLE OF THE THESIS (**İngilizce** / English) : ETHICS OF ARTIFICIAL INTELLIGENCE: MORAL RESPONSIBILITY OF SELF-DRIVING CARS AND SEX ROBOTS

TEZİN TÜRÜ / DEGREE: **Yüksek Lisans** / Master ■ **Doktora** □ PhD

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. ■

2. **Tez iki yıl süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of **two years**. * □

3. **Tez altı ay süreyle erişime kapalı olacaktır.** / Secure the entire work for period of **six months**. * □

*** *Enstitü Yönetim Kurulu kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir. / A copy of the decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.*

Yazarın imzası / Signature ....

Tarih / Date ………….