

**ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF ARTS AND
SOCIAL SCIENCES**

**AUTONOMOUS MACHINES AND MACHINE ETHICS:
CAN AUTONOMOUS CARS BE CONSIDERED AS MORAL AGENTS?**

M.A. THESIS

Hüsniye Dilara YEŞİLOVA

Department of Science, Technology and Society

Science, Technology & Society M.A. Programme

JUNE 2019

**ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF ARTS AND
SOCIAL SCIENCES**

**AUTONOMOUS MACHINES AND MACHINE ETHICS:
CAN AUTONOMOUS CARS BE CONSIDERED AS MORAL AGENTS?**

M.A. THESIS

**Hüsniye Dilara YEŞİLOVA
422161009**

Department of Science, Technology and Society

Science, Technology & Society M.A. Programme

Thesis Advisor: Prof. Dr. Aydan TURANLI

JUNE 2019

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ SOSYAL BİLİMLER ENSTİTÜSÜ

**OTONOM MAKİNELER VE MAKİNE ETİĞİ:
OTONOM ARAÇLAR AHLAKİ FAİL OLARAK KABUL EDİLEBİLİR Mİ?**

YÜKSEK LİSANS TEZİ

**Hüsniye Dilara YEŞİLOVA
422161009**

Bilim, Teknoloji ve Toplum Anabilim Dalı

Bilim, Teknoloji ve Toplum Yüksek Lisans Programı

Tez Danışmanı: Prof. Dr. Aydan TURANLI

HAZİRAN 2019

Hüsniye Dilara Yeşilova, a M.A. student of ITU Graduate School of Arts and Social Sciences student ID 422161009, successfully defended the thesis entitled “AUTONOMOUS MACHINES AND MACHINE ETHICS: CAN AUTONOMOUS CARS BE CONSIDERED AS MORAL AGENTS?”, which she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor : **Prof. Dr. Aydan TURANLI**
Istanbul Technical University

Jury Members : **Asst. Prof. Dr. Ash ÇALKIVİK**
Istanbul Technical University

Asst. Prof. Dr. Berrak KARACA ŞALGAMCIOĞLU
İstanbul University

.....

Date of Submission : 03 May 2019
Date of Defense : 11 June 2019

To my family and friends,

FOREWORD

I would like to thank my supervisor Aydan Turanlı for her patience and guidance. I would like to thank Brad Wray for introducing me to this topic and helping me to structure my first ideas. I would like to thank my family for always supporting me to follow my path. I would like to thank my dear partner Paul Ketz for his encouragement and trust in me. And last but not least, I would like to thank my friends, İrem Çakıroğlu and Muhammet Ramoğlu, for all their support and contributions to my work and motivation.

June 2019

H. Dilara YEŞİLOVA

TABLE OF CONTENTS

	<u>Page</u>
FOREWORD.....	ix
TABLE OF CONTENTS.....	xi
ABBREVIATIONS.....	xiii
LIST OF FIGURES.....	xv
SUMMARY.....	xvii
ÖZET.....	xix
1. INTRODUCTION.....	1
1.1. Motivation.....	1
1.2. Research Question.....	3
2. LITERATURE.....	5
2.1. Autonomous Vehicles.....	5
2.1.1. Mandatory Ethical Settings vs. Personal Ethical Settings	7
2.2. Machine Ethics.....	8
2.2.1. AI4People and Good AI	8
2.2.2. James Moor and Ethical Robots.....	10
2.2.3. Wallach & Allen and Artificial Moral Agents.....	11
3. MACHINES AND MORALITY.....	15
3.1. Can We Introduce Ethics to Machines?.....	15
3.1.1. Moral Decision.....	15
3.1.2. Utilitarianism.....	18
3.1.3. Deontology.....	20
3.1.4. Virtue Ethics.....	22
3.1.5. Clustering and <i>Moral Machine</i>	23
3.2. Can We Teach Morality to Machines?.....	26
3.2.1. Reinforcement Learning and <i>Nao</i>	26
3.2.2. Problem of Limitation.....	28
3.2.3. Problem of Goal-Oriented Behavior.....	30
3.2.4. Problem of Reliability.....	31
3.2.5. Imitation Learning and the <i>Child Machine</i>	34
4. MACHINES AND AGENCY.....	37
4.1. Autonomy.....	38
4.2. Intentionality.....	39
4.3. Responsibility and Accountability.....	41
CONCLUSION.....	43
REFERENCES.....	45
CURRICULUM VITAE.....	49

ABBREVIATIONS

AA	: Artificial Agent
AI	: Artificial Intelligence
AMA	: Artificial Moral Agent
MA	: Moral Agent
MES	: Mandatory Ethical System
PES	: Personal Ethical System
SI	: Super Intelligence

LIST OF FIGURES

	<u>Page</u>
Figure 2.1 : Two dimensions of AMA development	13
Figure 2.2 : Different approaches to machine ethics	14
Figure 3.1 : Moral Theories	19
Figure 3.2 : Lego Bricks caricature about Jeremy Bentham.....	20
Figure 3.3 : A question example from MIT's <i>Moral Machine</i> Research	24
Figure 3.4 : Personal scoreboard after participating <i>Moral Machine</i> Research.....	25
Figure 3.5 : Nao helping the morning exercise.....	27
Figure 3.6 : A dialogue in Facebook's research interface.....	30
Figure 3.7 : The conversation between Bob and Alice.	31
Figure 3.8 : The message of yogurt.	34
Figure 3.9 : Problems of teaching morality to machines	35
Figure 4.1 : Problems of agency attribution	40
Figure 5.1 : Research question	45

AUTONOMOUS MACHINES AND MACHINE ETHICS: CAN AUTONOMOUS CARS BE CONSIDERED AS MORAL AGENTS?

SUMMARY

Autonomous machines are becoming more and more involved in our lives. By interacting with their environment, these machines are able to perform certain tasks without human intervention. With this ability, they are used in tasks requiring human labor in many sectors. Driverless cars, for example, are expected to reach a level of autonomy that enables them to perform a journey without any need for a human being in the near future. Autonomous machines while performing a specific task are making specific decisions for each of their actions. It is very important to ensure that the machines can make 'right' decisions since the consequences of their actions may affect human beings directly. Therefore, it is important that the autonomous vehicles on the traffic act ethically when faced with a conflict.

In order to examine the relationship between autonomous machines and ethics, I first make an assessment on autonomous cars according to their autonomy levels. Later I examine the different approaches to machine ethics and criticize the existing moral machine classifications. In order to investigate whether a moral agency can be attributed to autonomous cars, I divide the concept into two sections: morality and agency. First I construct the relationship between morality and machine on the basis of two questions: (a) can we introduce ethics to machines and (b) can we teach morality to machines. Second I question whether a machine can have an agency by examining their competence of autonomy, intentionality, responsibility and accountability.

OTONOM MAKİNELER VE MAKİNE ETİĞİ: OTONOM ARAÇLAR AHLAKİ FAİL OLARAK KABUL EDİLEBİLİR Mİ?

ÖZET

Günümüzde otonom makineler hızla gelişiyor ve makine zekâsı birçok sektörde insan faaliyetlerinin yerini almaya başlıyor. Makinelerin arttırılan özerklik seviyesi ile bir insan müdahalesine gerek kalmadan belirli bir komutu başından sonuna kadar gerçekleştirebilecek sistemler hedefleniyor. Nesnelerin interneti ile makineler kendi aralarında iletişime geçmeye başladı bile ancak hala insanların dâhil olduğu kritik senaryolarda makinelerin nasıl davranması gerektiği kesinleşmiş değil. Sürücüsüz otomobil üreticileri gelecekte trafiği daha güvenli hale getireceklerini iddia etseler de, geleneksel trafik sistemindeki otonom otomobiller şu an için çözülmesi gereken problemler yaratıyor. Dünyanın her yerindeki bilim insanları, özellikle yakın gelecekte kullanmaları beklendiğinden sürücüsüz araç etiğinin nasıl geliştirilmesi gerektiği sorusuna cevap arıyorlar. Bu makineler insanların bile bazen tereddütte kaldıkları etik ikilemlerle karşılaştıklarında nasıl karar vermeliler? Ahlak felsefesinin ünlü dilemmalarından olan “Trolley Problemi” bu sefer asfaltta karşımıza çıkıyor. Sürücüsüz araçların etrafındaki varlıkları algılama yeteneği ve otonomi seviyesi arttıkça kaza durumlarında hangi canlıyı korumaya karar verecekleri herkes için hayati önem taşıyor.

Otonom makineler ve etik arasındaki ilişkiyi incelemek için ilk önce sürücüsüz otomobillerde otonomi düzeylerine göre bir değerlendirme yapılmıştır. Daha sonra makine etiğine farklı yaklaşımlar incelenerek mevcut etik makine sınıflandırmaları karşılaştırılmıştır. Bir ahlaki faillik durumunun sürücüsüz araçlara atfedilip edilemeyeceğini araştırmak için, konsept iki bölüme ayrılmıştır: ahlak ve faillik. Öncelikle ahlak ve makine arasındaki ilişkiyi iki soru temelinde tartışılmıştır: (a) makinelere etik davranışın ne olduğu gösterilebilir mi ve (b) makinelere ahlaklı olmayı öğretebilir miyiz? İkinci olarak, bir makinenin özerklik, niyetlilik, sorumluluk ve hesap verebilirlik yeteneklerini inceleyerek bir faillik durumuna sahip olup olamayacağı tartışılmıştır.

Makinelerin hızla büyüyen özerkliği aynı zamanda etik olarak yetkin olmalarını da gerektiriyor. Tamamen otonom bir araç seviyesine henüz ulaşamamış olsa da, belirli bir ahlaki kimliğin bu makinelere atfedildiği görülebilir. Bu referansın uygun olup olmadığını tartışmak için önce makinelerin ahlaki davranıp davranamayacağı sorusu cevaplanmış, sonra makinelerin faillik kavramını üstlenebilecek durumda olup olmadığı tartışılmıştır. Bir makinede ahlaki davranışın nasıl tanımlanacağına dair birkaç teori vardır. Bu teorileri iki soru altında toplanmıştır; makinelere ahlak gösterilebilir veya öğretilir mi? Birincisi yaklaşım, ahlaki karar mekanizmalarının programlanabilir olduğu anlayışına dayanarak, makinenin karşılaşabileceği olası senaryoların çözümünü tasarıma dâhil etmektir. Böylece makinenin herhangi bir çelişkili durumda mevcut veri üzerinden kodlanan “doğru davranış” şeklini uygulaması amaçlanır. Bu yaklaşımda makineler, sistemlerinde kodlanan etiğe göre hareket ederler ancak bu etiğe göre hareket ettiklerinin “farkında” olmayabilirler. Makinelere doğrudan aktarılabilecek etik bir sistemin varlığını araştırırken ahlak

felsefesi incelenmiştir. Ancak, her biri aynı soruya farklı cevaplar veren farklı etik teoriler ahlaki bir ikilemle karşı karşıya kalındığında tam olarak ne yapılması gerektiği hakkında da farklı fikirler vermiştir. Herhangibir etik teorisinin diğerlerinden üstün veya geçerli kabul edilip makinelerle direkt olarak uygulanması pratik açıdan mümkün görünmemektedir. Bu soruna bir çözüm olarak tasarlanan Moral Machine araştırması, dünyanın dört bir yanından katılımcılara doğrudan ahlaki ikilemler aracılığıyla sorular sormaktadır. Kullanıcıların bir internet sitesi üzerinden katıldığı araştırma, dünyadaki etik davranış ilkelerini bir araya getirerek ortak bir sonuca varmayı hedeflemektedir. Ancak ahlaki ikilemlerin hazırladığı senaryolardan beslenen bir makine bile oluşabilecek tüm çatışmaları çözmede başarılı olamamıştır. Günümüzün karmaşık etik sorunlarını çözmek, temel olarak tüm olası senaryoları simüle etmenin imkansızlığı, gerçek bir etik cevaba ulaşmanın imkansızlığı ve küresel bir uzlaşma oluşturmanın imkansızlığı nedeniyle yetersiz kalmıştır. Bu yüzden mevcut şartlarda herhangi bir ahlaki davranış kodunun makinelerle doğrudan uygulanması olası görünmemektedir.

İkinci yaklaşım makine zekasının kendi kararlarını verme yetisine sahip olacağına inanmakla, makinelerle doğru ahlaki davranışı öğretmeyi amaçlar. Makinelerin kendi davranışlarının nedenselliğini kurabilecekleri ve sonuçlarını öngörebilecekleri varsayımına dayanan bu yaklaşım Ahlaklı Makine noktasına ulaşılabilineceğini savunur. Makinelerin belirli referanslar ve geri bildirimler üzerindeki öğrenme deneyimleri, etik kavramına daha uygun görünmektedir. Öte yandan belirli seviyede bir makine öğrenmesi yetisine sahip bir yapay zekayı ahlaki olarak eğitmenin de zorlukları ortaya çıkacaktır. “İnsan” bilimadamları tarafından geliştirilen robotların beklemediğimiz stratejiler geliştirerek bizi şaşırttıkları durumlar az değil. Tıpkı Facebook’un “kendi dilini icat eden” chat-botları gibi insan bakış açısıyla geliştirilmiş sınırlamaların makineler için yetersiz kaldığı görülebilir. Günümüzde yapay zeka dijital bir yazılım üzerinde çalışan akıllı bir donanım olarak biliniyor. Gelecekte akıllı makinelerin içeriğinin ne olacağını kestirmek bugünden mümkün değil, belki de bilgisayar ortamından uzaklaşabilen veya organik bir bileşiğe sahip yapay zekalar üretilebilir. Bu noktada, kendi kendine öğrenen makineleri geliştirirken açıklanabilirlik kıstasını dikkate almak oldukça önemlidir. Sınırlama ve güvenilirlik problemlerini önlemek için. Çocuk Makinesi konseptinin yanlışlıktan uzak tutulması, şeffaflık ilkesi ile kontrollü bir izleme ile sağlanabilir. Günümüz teknolojisinde bu seviyeye ulaşmak mümkün olmasa da, gelecekte etik olarak yetkin makineler ile beraber yaşamamız olası görünmektedir.

Bu özelliğin ikinci bölümü incelendiğinde, faillik kavramının makinelerle kolayca uygulanamayacağı görülmektedir. Faillik, bir şemsiye terim olarak otonomi, niyet ve belirli seviyelerde sorumluluğa sahip olmak gibi ön koşulları içinde barındırır. Bu özelliklerden otonomide belirli bir seviyeye ulaşılmasının mümkün olduğunu, ancak uygulamanın istenmeyen sonuçlara yol açabileceği tartışılmıştır. Makine otonomisi arttıkça, makineler daha da bağımsızlaşacak ve bu durum onları insanlar tarafından daha az sorgulanabilir bir seviyeye çıkaracaktır. Bu durumun, teknolojik olarak ulaşılacak bir seviye olmasına rağmen, uygulanabilirlik açısından problemler çıkaracağından belirli seviyede tutulacağı tartışılmıştır. Amaçlılık, diğerlerinden biraz daha tartışmalı bir kavramdır, ancak makinelerin ulaşabileceği bilişsel seviye şu an tam olarak öngörülebilir değildir. Bununla birlikte, makinenin herhangi bir zihinsel kapasiteye sahip olmadığı iddiası (duplication) ve bu nedenle davranışlarının sadece sadece bir simülasyon olduğu argümanı öne çıkmaktadır. Sonrasında,

makinelere belirli bir sorumluluk kavramının atfedilebileceğinden bahsedilmiştir, ancak bu özellik hiçbir makine davranışlarından yükümlü tutulamayacağından işlevsiz kalacaktır. Davranışlardan yükümlü tutulma, gerektiğinde yaptırım uygulanabilir olmasını ve bir mesuliyet ilişkisini beraberinde getirir. Otonom bir makinenin, sorumlu kararlar almayı başarabilse bile ortaya çıkardığı durumdan mesul olamayacağı tartışılmıştır. Otonom bir arabanın karıştığı bir kaza senaryosunda, kaza incelense ve aracın belirli bir oranda sorumlu olduğu tespit edilse bile hesap verilebilirlik cevapsız kalacaktır.

Sonuç olarak, bir ahlaki faillik durumunun makinelere atfının antropomorfik bir yanlışlık olduğu savunulmaktadır. Etik kavramı insan üzerinden inşa edilmiş ve insani özelliklere doğrudan bağlı bir düşünce sistemidir. Bir makinenin etik olarak davranabileceği öngörülse dahi, bilişsel yeteneklerinin insan akli seviyesine ulaşım sağlayamayacağı belirsizdir. Yine de ahlaki makine seviyesine ulaşılabilirliği argümanı henüz geçerliliğini konurmaktadır. Öte yandan, fail kavramına sahip olmak için gereken özellikler makinelerde yoktur. Bu durum araştırma sorusunun ilk kısmının ulaşılabilir olduğunu ancak ikinci kısmının uygulanamaz olduğunu göstermiştir. Bir gün etik olarak doğru karar verebilen makineler üretilse dahi, bir ahlaki faillik kavramı atfedilemeyecektir. Bu araştırma ile amaçlanan, otonom araçların tasarımında dikkate alınabilecek bir kavram tartışması yapmaktır. Bilimsel çalışmalar, teknolojinin uygulama aşamasında da yardımcı olmalıdır. Bu nedenle, otonom araçların bir faillik kimliğine sahip olduğu varsayımından vazgeçilebilir ve etik olarak doğru bir davranışsal çerçevenin uygulanması için çalışmalar yapılabilir.

1. INTRODUCTION

1.1 Motivation

Autonomous machines have already started to spread out of factories' production lines or battlefields and found a place for themselves in our daily lives. The period in which machines are seen as only a mechanical entity is slowly falling behind. Today's smart machines have a high degree of autonomy and are almost about to announce their individuality. These machines can perceive their environment with the help of implicit scanners and sensors, can make specific decisions about what they are programmed to perceive and finally they can perform a certain action according to this decision. These autonomous robots, which can act on their own and need no- or only a small amount of human intervention, can be beyond any doubt useful technologies that will make our lives easier. Today, we can have a robot vacuum cleaner that knows the floor plan of our house creates a cleaning schedule on its own according to the amount of dirt on the floor and performs a cleaning duty when necessary. Similarly, it is only a matter of time before autonomous cars which are able to perform a journey starting from the garage of our house to the point of arrival without any intervention, come into our lives. Although it is comforting to leave the major tasks completely to the robots, it can be also unsettling. After all, the results of each decision taken by these robots also are affecting us. Naturally the harm caused by a wrong action is not always same. The robot vacuum may cause a huge dust crisis, yes, but a robot car may lead to some serious fatal situations.

It is expected that autonomous vehicles will significantly increase the transportation problems (Friedrich, 2016). They claim that autonomous vehicles will reduce the risk of traffic accidents by eliminating the human factor, which is the first cause of the accidents. In the "Vehicle Safety Report" (2019) covering the first quarter of the year, Tesla shared the following information; according to the data collected from their semi-autonomous vehicles, there was only one accident per 4.5 million kilometers of which the autopilot was active. For the same vehicles, there was one

accident per 2.8 million kilometers with the autopilot deactivated. On the other hand, the National Highway Traffic Safety Administration of United States announced the overall accident rate as 1 per 700,000 kilometers ("Traffic Safety Facts Annual Report Tables", 2019). Looking at these data, it can be seen that autonomous cars are already involved in accidents much less frequently than the regular cars in traditional traffic. Of course, it is important to note that Tesla collects this data from just 1.6 billion kilometers of driving. In addition, autonomous vehicle users may be more cautious than usual because of experiencing a new technology. Nevertheless, it is possible to say that the rates are promising. Autonomous cars may actually make the traffic safer. But there is a question not yet answered. How do autonomous cars make the 'right' decision in case of unexpected situations? It may be possible to reach rational decisions by an objective evaluation, such as setting the speed on a routine journey or preferring a particular route. However, according to the predictions, it is expected that autonomous vehicles will make also moral decisions when necessary. They may be required to evaluate the possibilities and carry out an ethically acceptable behavior.

For the adaptation of any new technology, the consumer must be convinced to purchase the technology. Cowan (1987) examines technological developments through a concept that she calls Consumption Junction and specifically examines the final stage of a technological system evolution: technological diffusion. She argues that there are several types of consumers in a technological system and that the success of this technology is dependent on consumers' purchase decisions. Therefore, in order for a consumer to incorporate a new technology into their life, they must be convinced first that the product works well and as intended. It may be easy to evaluate a relatively limited product, there are not so much to test when it is about a juicer. However, it is a controversial issue how the machines will take correct action decisions when they are faced with a problem. Users, naturally, want to be sure that their car will act in a way that protects the passengers. The intensity of dropping responsibility into a machine is directly proportional to the reliability of that machine. Therefore, the construction of the perception that this new technology is reliable has a great importance. On the other hand, as Cowan (1987) continues, there is more than one consumer type in a technological system. In addition to the end-user, there are agents in the intermediate stages that take on the role of the

consumer: such as the car companies that purchases the required software for the machine. They also need to be convinced that software companies that produce this artificial intelligence can design machines that are able to make right decisions. Of course, this problem has also a producer perspective. Software producers are working with scientists and programmers in order to develop new artificial intelligences which are capable of ethical decision making, since there is yet no specific code of ethics accepted for autonomous vehicles.

This developing technology will begin to experiment with its first products in the traditional system. Soon we can start to encounter autonomous cars on the road even if we are not the owner. That's why it's important for everyone to acknowledge how an autonomous car reacts to a moral dilemma. I find it worthwhile to build a discussion of how this controversial technology, which is currently in development, should address or take the concept of morality. I find it useful to discuss in detail how the concept of moral agency, which is directly attributed to machines, takes place in the production and implementation processes. I hope this discussion will contribute to the literature both in terms of facilitating the adaptation process of the society and of being a guideline to both researchers and programmers.

1.2 Research Question

My aim with this thesis is bringing a critical approach to machine ethics. I want to explore the process of technological innovation from a philosophical perspective. Bruno Latour (1987) mentions two faces of Janus while studying the stages of technological and scientific developments. These two faces, which say completely different things, represent two different approaches. The third dictum of Janus consists of the following statements: “Once the machine works people will be convinced” on one side and “The machine will work when all the relevant people are convinced” on the other side (Latour, 1987, p.10). He states that once the machine works, it becomes a black box exactly because it manages to work. The property of being able to work makes it a good machine. So while it is being made, the machine cannot convince anyone yet that it is a good machine. Likely, building an ethical machine processes in a similar way: *once the machine works ethically, people will be convinced* and on the other hand *machine will be moral when all the relative people are convinced that it can be*. This machine is not yet capable of convincing anyone,

as there is an incomplete technology that is still in development. So I argue it is better to challenge Janus's second face:

Can autonomous machines be considered as moral agents?

This moral agency, attributed to machines, consists of two parts: being moral and having an agency. In order to determine the boundaries of the research scope, I investigate the autonomous vehicles in the next chapter. I discuss the current situation of autonomous vehicle technology and the levels of autonomy. Afterwards I examine the literature on machine ethics. After introducing the existing classifications of artificial moral agencies, I argue their validity.

In the third chapter I question whether it is possible to have moral machines. I would like to ask two different questions to discuss whether the machines can have ethical abilities: (a) *can we introduce ethics to machines?* And (b) *can we teach moral behavior to machines?*

In the fourth chapter, I discuss whether machines can have an agency. I will demonstrate the problems that may arise in relation to the concepts of autonomy, intentionality and responsibility/accountability which are the requirements of the concept of agency.

In the last chapter, I argue that even the construction of ethical machines may be possible in the future; an ethical agency cannot be attributed to them since machines, by their nature, cannot have an agency.

2. LITERATURE

2.1 Autonomous Vehicles

Autonomous vehicles have multiple autonomy levels within themselves. In 2017, a 6-level classification was set for autonomous cars, as published in the IEEE conference (Zanchin et. al., 2017). According to the report, level zero is the case where the car does not have any autonomy; the driver does the whole job. Although the instrument panel may give information about the car, it does not fall into the class of autonomy. The first level is called the driver assistant. The car at this level may show some autonomous functions, such as slowing the car when the speed limit is exceeded. However, all functions and responsibility are still in the driver. The second level, partial autonomy, is the case where the vehicle can continue to autonomously drive in some driving conditions. The driver must be on the seat all the time but may authorize the vehicle to manage the steering and speed on certain motorways. The third level is the conditional autonomous level. Vehicles at this level can control the steering and speed, with its sensors can follow the path, but the driver is still expected to intervene when necessary. The fourth level is called high autonomy, the vehicle at this level can follow its own path, adjust its speed, realize other vehicles and pedestrians on the road and perform functions such as lane changing and signaling. An autonomous vehicle at this level is able to do anything that a driver does on the road, however the driver can take control at any time and decide whether the route is suitable for autonomous driving. The fifth level is fully autonomous vehicles; the vehicle at this level does everything a human can do and does not need people for any function. It is also expected that in this level, human intervention may be prohibited in order to avoid any malfunction. The fully autonomous car can take the passenger to the desired location via the designated position or can deliver a package for example. We can also expect full autonomous vehicles to look unlike the cars we know. The current usage scenario will no longer be valid at that autonomy level. These vehicles, where a steering or control panel is not needed, can focus only on the comfort of the ride and make the journey experience a more social medium.

When the fully autonomous traffic scenario is considered, it can remind the frames from science fiction films. However there are strong signals that the future is not far away. Today the third level autonomous vehicles are already seen in traffic, in interaction with other vehicles and pedestrians on the road (e.g. Tesla-Autopilot). In March 2018, the first fatal self-driving car accident occurred. A self-driving Uber vehicle, on its test drive, hit a pedestrian and killed her (Wakabayashi, 2019). It was claimed that the vehicle was in self-driving mode, but did not detect the presence of a pedestrian and the emergency brakes were disabled. Although there was a systematic error related to the car's hardware, this accident revealed many questions about the behavior of autonomous vehicles in case of an emergency. This is an important point to be considered because, unlike human drivers, the machine can make a rational decision and react on it in almost microseconds (Sharkey, 2012). In the case of an emergency the car decides what to do according to the decision mechanism previously coded in it. That is why the decision-making systems of the vehicles with a certain level of autonomy are becoming an important topic.

As autonomous cars are more and more being involved to our daily lives, it is being questioned how they are going to respond to certain moral dilemmas. The Trolley problem, one of the most famous examples of moral dilemmas, is looking for the answer to this question. Foot (1967) in his paper about the double affect offers a thought experiment called “Trolley Problem” in order to make people question their decision making mechanisms about a moral dilemma. This problem questions whom to sacrifice in a train crash with two fatal scenarios. Imagine you are standing near a train station next to the controller of the rail switches. The train is approaching and you realize that there are 5 people tied on the tracks that train will pass soon. Just before the point where people tied to the rails, there is a fork track and only one person is attached to the other side of the rail. Remember that you are standing next to the controller crank. What would you choose, letting train to continue its way and kill five people or changing the direction and leading the train to the other rail where only one person is tied? (Foot, 1967). Years after the creation of this problem, people are still not able to reach a common judgment.

Trolley problem is updated and transformed to asphalt ground this time. When an autonomous car suddenly realizes a family of five crossing the road, should it swerve the steering wheel, hit the bus station on the side of the road and harms the passenger

inside; or should it continue and hit the family in order to protect the passenger? (Lin et al., 2017) Let's look at another scenario: an autonomous car has to run to the right or left in order to avoid a crash by a dragging truck lying on its side. There is a motorcycle driver on the right side who wears a helmet, but on the left side there is another motorcycle with no helmet. What to do, where to swerve? It is possible to think that hitting the driver wearing a helmet will bring less damage than hitting the driver without one. On the other hand, wouldn't we have punished the driver who had taken all the precautions for driving safety by wearing his helmet? And wouldn't we implicitly encourage driving a motorcycle without a helmet?

In the light of these problems, it can be better understood why the autonomous instruments is considered necessary to have certain moral inference ability. Naturally, since the autonomy level of each machine is different, the degree of ethical ability can vary. In order to examine the relationship between machines and ethics in general terms, I see helpful to take a step back from autonomous vehicles and have an overall literature review about machine ethics.

2.1.1 Mandatory Ethical Settings vs. Personal Ethical Setting

At this point in the discussion, I think it is important to intervene with the question of whether the ethical decision-making mechanisms to be operated in autonomous cars should be a personal or a mandatory system. Should our car behave in accordance with a society-wide set of rules when it is faced with a moral dilemma, or should it apply the personal behavior code we have decided? I find the arguments of Gogoll and Müller (2017) useful and feasible in this mandatory or personal ethical settings debate. The duo, who conducts a discussion on why the personal ethics system (later will be used as PES) will not work in the search for a system that may be of general benefit to society, claims inapplicability because of the *prisoner's dilemma* (see. Rapoport et. al., 1965). In order to simplify the situation, they divide society into selfish and moral individuals. In the event of an inevitable accident, we can assume that the moral individual may be willing to sacrifice himself in order to serve the overall benefit, but the selfish individual would act in the forefront of their benefit regardless of the outcome. In a traffic system where the PES is implemented, both can be in the same accident in their autonomous cars which are programmed according to their personal ethical preferences. In such a scenario, the moral

individual, who normally takes into account the overall goodness of the community, cannot predict whether the other car will be as moral as he is, since he cannot know the details of PES of the other car. In this case, we can assume that the moral individual may consider programming his car in favor of self-benefit in order not to be affected negatively after an unfair evaluation (Gogoll & Müller, 2017). Therefore, a personalized ethics setting cannot go beyond a selfish level. In order to create an effortless autonomous traffic system, it will be the most feasible solution to have certain predefined moral codes to operate on the vehicles.

2.2 Machine Ethics

Machine ethics is not a new issue. Almost 70 years ago, robots were introduced into our lives through science fiction books and back then robot ethics were already being discussed. One of the cult works in the field of machine ethics *I, Robot* by Isaac Asimov retains its popularity even today. Asimov determines the 3 important rules of machine ethics:

First; a robot may not injure a human being or, through inaction, allow a human being to come to harm. Second; a robot must obey the orders given it by human beings except where such orders would conflict with the First Law. Third; a robot must protect its own existence as long as such protection does not conflict with the First or Second Laws. (Asimov, 2014, p.37)

Although they initially cover crucial points, Asimov's rules are insufficient for today's complex decision-making mechanisms. Nevertheless, it is one of the most important works in terms of being the first rules written about machine ethics and introducing the issue to the literature and have an overall literature review about machine ethics.

2.2.1 AI4People and Good AI

In a statement published in November 2018, the AI4People initiative proposed a list of ethical frameworks to be followed in order to create a good artificial intelligence community (Floridi et al., 2018). AI4People is a forum that includes members of the European Parliament, civil society organizations, industry and the media. The forum published the ethical planning scheme required for a good society after their first year of activity. The AI4People initiative is built by 13 researchers from 9 countries, and they draw a framework for the ethical practices of artificial intelligence. They

reduce the ethical machine discussions to 6 different approaches and identify the repetitive principles. As a result, they identify the four most commonly used principles and propose a fifth: beneficence, non-maleficence, autonomy, justice and explicability (Floridi et al., 2018). Beneficence is one of the main features of artificial intelligence that are ranked first; this technology should be programmed to be useful to humanity. The working mechanism should serve the common good and the decision mechanisms should aim for the results that humanity may benefit. Another point that has been emphasized in the discussions of machine ethics since Asimov is the non-maleficence principle: doing no harm. As can be understood from its definition, the principle of non-harm emphasizes that any artificial intelligence should avoid any action that could harm humanity. As the AI4People initiative points out, this principle actually covers two subjects, firstly that humanity should not program the artificial intelligence in such a way that it does not obey the principle of non-harm, and secondly, the artificial intelligence itself should behave according to the principle of non-harm. Here the concept of intentions comes into play. As discussed in the article, artificial intelligence, which is capable of conducting a particular decision-making mechanism and both of the programmers, should not intend to do harm. The third principle they cover is autonomy; the freedom of individuals to make decisions about themselves. This issue also covers the autonomy of two subjects. While artificial intelligence must respect the autonomy of man in all conditions, the people who program and produce artificial intelligence should keep the autonomy of robots at a limited and reversible level. This principle suggests that robots may have the ability to make decisions, and that people should keep the option to decide whether they want/prefer to decide on a certain topic. In short, humanity must dominate the boundaries of the ability and authority of decision making delegated to robots. The fourth principle is justice, a principle that supports the fair distribution of available resources. The interpretation of AI4People to a fair artificial intelligence involves 3 different substances: artificial intelligence should be used to advocate a fair life and correct all discrimination, the added-value provided by artificial intelligence must be equally sharable by all, and artificial intelligence should avoid behaviors that would undermine the existing order of society (Floridi et al., 2018). In addition to these 4 articles of bioethics, the proposed 5th item is explicability, i.e. transparency in programming and working processes. Floridi (2015) states that nowadays the decision makers of this widely applied technology is

just a *small fraction of humanity*, therefore the questions of how artificial intelligence works ethically and who is the responsible one to decide how it works should be explained in a way which is understandable by all segments of society.

This report is significant in terms of being published by the first ethical artificial intelligence forum in Europe. The criteria proposed to create a good artificial intelligence society can be interpreted as a proposal to produce service machines with limited autonomy. Although they have drawn a guiding scheme by narrowing ethical practices, they did not provide a detailed comment on the ethical level and values that machines should have. Nevertheless, I consider this report worthy of mentioning since it questions how good artificial intelligence should be and determines the initial criteria.

2.2.2 James Moor and Ethical Robots

Another category to be examined after good artificial intelligence is ethical robots. Nowadays, ethical robots are being examined in several different categories. In addition to the anthropomorphic robots that Asimov classifies, there are a lot of robotic machines that have different interfaces and physical forms that we confront in our daily lives. As expected, the ethical competence of the robots varies according to their complexity. James Moor (2009), Professor of Intellectual and Moral Philosophy at Dartmouth College, classifies ethical robots in 4 different categories. He starts with the weakest autonomy and names them as ethical impact agents. The acts of ethical impact agents can have ethical or unethical consequences regardless of the intent factor. Moor gives an example of a digital wristwatch; whether or not to show the time correctly can affect the owner's timely arrival. Moreover the unethical impact agents which can be counted in the opposite position are also placed into this category. Secondly, implicit ethical agents are robots with ethical behavior patterns programmed in their designs. These are generally robots that perform safety and security related concerns. For example, an ATM is obliged to accept the correct password, to show the customer's information correctly and to deliver the desired amount of money accurately. This obligation is due to the software and design of the robot, thus an internal (implicit) programming. We are familiar with ethical robots in the first two categories in our daily life. The next two may require some more imagination. In the third category, explicit ethical agents, the robots are expected to

analyze a situation and reach an ethical conclusion. They can identify problems and make a detailed reasoning about what needs to be done. In situations where there are ethical conflicts, they can make inference and reach a certain result. However, they do not realize that they perform these analysis, reasoning and inference processes according to a certain ethical setting. They act “from ethics, not merely according to ethics” (Moor, 2009, p.12). And finally, he makes a distinction between explicit ethical agents and full ethical agents. Full ethical agents can also make ethical inferences in various scenarios. However unlike the previous ones, robots in this category have ethical abilities similar to humans; like consciousness, intentionality and free will. Therefore, when full ethical agents respond to a situation, they understand and analyze the situation and respond with a behavior according to their personal ethical reasoning.

The first two categories of Moor stand at a weak point in the ethical machine classification. The working mechanisms of digital watch and ATM examples can be considered as design features rather than the ethical characteristics of machines.

When an engineer designs a bridge she does not finish the design and then consider how to stop it from falling down. The ability to remain standing in all foreseeable circumstances is part of the design criteria, not a separate discipline called “bridge ethics”. Likewise, if an AI system has deleterious effects it is simply a badly designed AI system. (Chase, 2019, para.8)

In this case, the correct execution of the given command by the ATM actually indicates that the machine is working accurately, not that there is any autonomy. Of course, it can be assumed that the machine programmers have had some ethical concerns, but this concern is no more than the wish of any product developer to make a product that works properly. Therefore, it is not appropriate to refer to any autonomy in cases where the machine has no property of making a behavioral decision.

2.2.3 Wallach & Allen and Artificial Moral Agents

On the other hand, the moral machine classification of Wallach and Allen (2008) contains slightly less categories. They question the concept of goodness in machines by advocating instrumental goodness, that is, whether a particular machine is good at executing a particular command. The fact that a broom is doing its job well in this context makes it a good machine, yes, but it would be wrong to speak of a kind of goodness in the moral context. They claim that the possibility that a machine's

behavior may cause harm does not directly make it a moral agent either. A falling tree may cause harm but a moral agency cannot be attributed to a tree. The assets they accept as a moral agent must have the ability to evaluate their own behaviors according to its possible moral consequences. Artificial moral agents (later will be used as AMAs), which Wallach and Allen (2008) offers as a category, are machines that can see the risk of damage of a particular act and have the autonomy to take the necessary steps to prevent this situation . In contrast to Moor's ethical robots, here at least a minimum level of awareness is required for a machine to be considered a moral agent. "However, if an autonomous system is to minimize harm, it must also be "cognizant" of possible harmful consequences of its actions, and it must select its actions in light of this "knowledge," even if such terms are only metaphorically applied to machines." (Wallach & Allen, 2008, p.17). Here, the term *cognizant* brings a new narrative to the topic:

Systems at the low end of these dimensions have only what we call "operational morality"—that is, their moral significance is entirely in the hands of designers and users. As machines become more sophisticated, a kind of "functional morality" is technologically possible such that the machines themselves have the capacity for assessing and responding to moral challenges. However, the creators of functional morality in machines face many constraints due to the limits of present technology (Wallach & Allen, 2008, p.9)

Wallach and Allen (2008) also establish the relationship between the current technology and the level of the machines which they call sophisticated AMA (Fig. 2.1). They evaluate the stages between these two levels of complexity according to their degree of autonomy and sensitivity. Simple hand tools, for example, have neither autonomy nor certain sensitivity in this case. They give some examples of low moral technologies that have certain levels of operational morality; such as the fact that the lid of a washing machine has a child lock function in order to prevent a child to get harmed. However, this cannot be explained by the autonomy or sensitivity of the washing machine itself, the lock is only a property of its hardware and may only be an indication of its designer's sensitivity. So it wouldn't be appropriate to attribute any moral agency to a washing machine. While on one side of the sensitivity- autonomy graph stands the simple hand tools, on the other side there are *full moral agents* with complete autonomy and sensitivity. Today, that level of technology does not yet exist. Therefore, they offer a step in between these two

edges; *functional morality*. Functionally moral machines are smart to a certain extent and can act according to the ethical result of their behavior.

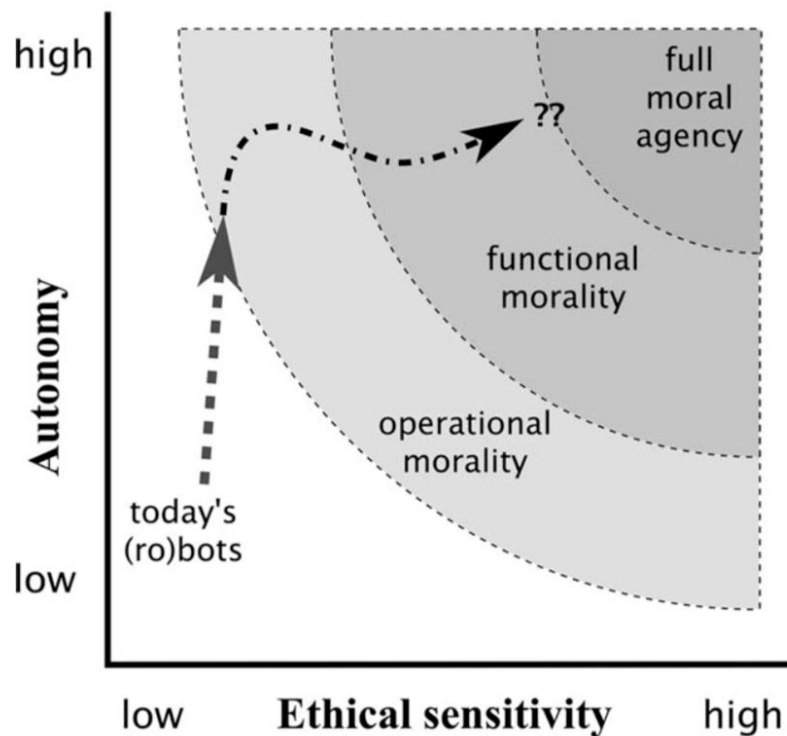


Figure 2.1 : Two dimensions of AMA development adapted from (Wallach & Allen, 2008).

Wallach and Allen (2008) divide the functional moral robots into two: those with a little sensitivity to a valid autonomy and those with a high ethical sensitivity but a lower autonomy. They give auto-pilots in the planes as an example for the former one. These robots, which have autonomy at a valid rate, can fly a plane with a minimum amount of need for a real pilot and can respond to the systemic problems that may occur during the flight. In case of any maneuver, they can show the sensitivity of avoiding an extreme angle that may disturb the passengers. This decision is not related to the plane's ability to take a sharp bend or to the direct observation of current passengers' conditions. The command to act within the ergonomic limits is already in the auto-pilot's software. Thus, in regular situations, the auto-pilot can perform a flight by providing the required ergonomic comfort and it is not able to make sharp turns in order to add some more excitement to the flight, even if it would be a wish of passengers. It is possible to see that the autopilot has a

relatively high autonomy but cannot show a specific sensitivity. In the realm of the second class, they demonstrate systems that support the ethical decision mechanism which are often used in the field of education. For example MedEthEx is a medical ethics expert system developed by Michael and Susan Anderson (I will mention about them and their highly autonomous robot NAO later in more detail) which is designed to give insights to doctors within an ethical dilemma (Anderson et. al., 2005). Suppose the patient rejects the treatment method, which is believed to give the best results. Should the treatment be forced to the patient by ignoring their autonomy? Or should the doctor respect the patient's decision by violating the medicine rule which requires applying the best treatment available in the medical practice? MedEthEx analyses the situation by asking the doctor to answer specific questions and then compares them with the other cases present in its database in order to provide an ethically correct way to proceed. The examples given are still far from the full autonomy and full sensitivity points even the levels vary. These examples are useful for detecting the distance to the level defined as *full moral agency*.

Although there are different approaches on whether a certain agency is attributed to machines, it can be seen that the need of a machine which behaves ethically is agreed (Fig. 2.2). Even if different names have been given so far, ethical machines are mostly expected to have an agency with a certain level of autonomy. In order to discuss whether this moral agency is an accessible level, I first examine how the ethical decision-making mechanism can be transferred to the machines. I will question the existence of an ethical system that can be adapted by the machines.

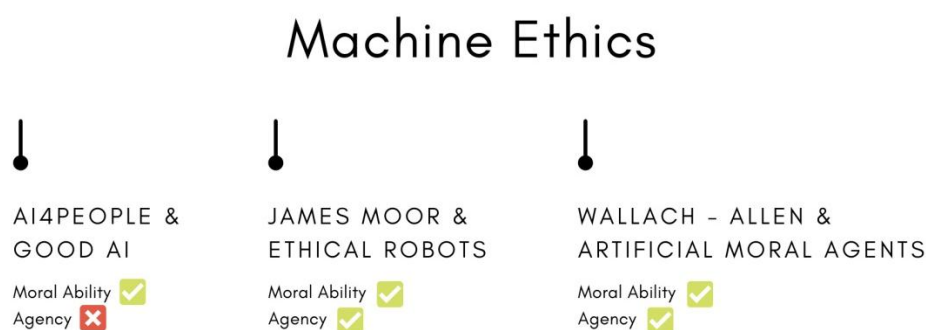


Figure 2.2 : Different approaches to machine ethics.

3. MACHINES AND MORALITY

3.1 Can We Introduce Ethics to Machines?

The easiest way to ensure that machines behave ethically would undoubtedly be programming the ethical behavior. If we adopted certain moral rules and turned it into an algorithm, we could be sure that the machine would behave according to the code required for each of these conditions. But moral philosophy is full of theories that have not yet reached a consensus. Each one looks at the concept of moral decision from a different point and proposes its own structures of verification. I will discuss the nature of moral thinking under moral philosophy to investigate if one of the theories could be introduced to machines.

3.1.1 Moral Decision

Why is it important to talk about the decision-making mechanisms of autonomous vehicles? We want to make sure that they will behave the right way in the moments where the decision is left to the machines. We want to think that they will value humanity and work for our well-being. But what exactly does it mean to behave in the right way? How do we decide whether a behavior is right or wrong? At this point, ethics is entering to the scene. It can be argued that everyone, every family or society has personal evaluation appraisal. This claim can give us information on the subjectivity of the concept of ethics. However, from a philosophical point of view, ethics examines existing moral structures with their roots and investigates whether one is better than the other. Therefore, ethics is also called moral philosophy. Ethics mainly asks basic questions about the good life, about what is right and wrong, about whether there is any objective right and wrong, and how we know it if there is.

The examination of the language used to talk about an ethical judgment has major significance. When morally evaluating a behavior, we define it as right or wrong. And this evaluation gives a certain amount of information about whether it should be

done or not. On the other hand, to state that something is morally good does not necessarily mean that it must be done. Good and right are not equal terms. Another difference is that there are no other steps between right and wrong. One thing is either right or wrong. However, the concepts of good and bad have multiple levels among them, and the levels of goodness of the two options can be compared to make a total judgment.

Ethical judgments, by their very nature, need causality. MacKinnon and Fiala (2014) investigate the working mechanism of the ethical judgments in their book; *Ethics: Theory and Contemporary Issues*. They argue that when we define something as right or wrong, we create a judgmental system based on certain norms. The strength of causality can give information about the fairness of the moral judgment we have reached. Of course, the fact that there is a causal relation does not mean that ethical judgments are completely rational. It is still a matter of debate whether we can reach any judgment independent of our feelings and prejudices (MacKinnon & Fiala, 2014). Philosophers have not come to a consensus yet about where people come to a conclusion about something being right or wrong. For example, intuitionism argues that judgments about good and right are formed by an intuitive process. When we make a non-empirical judgment, we use our ethical acumen that we know intuitively. What is lacking in this theory is that it does not take a stand about whether intuition is objective or subjective. Another theory, emotivism, argues that when we judge something as good or bad, we indicate whether we approve it or not instead of identifying it. Emotivism defends a subjective judgment, claiming that our personal experiences directly shape our understanding of ethics. However, the problem of a moral judgmental system shaped by emotions is that it would never be possible to reach any consensus, because emotions and feelings are variable. A sentimental judgmental system is completely subjective and does not give us any suggestion about a common feeling of ethics. It is useful to discuss objectivism after these two relatively subjective theories. Objectivists don't argue whether the things we desire are good or not. On the contrary, they argue that we naturally desire things that are intrinsically good. In this theory it is emphasized that the concept of goodness is already present in us, so things have an objective goodness. The claim that something or behavior has an intrinsic value supports that ethics can be conceived objectively, since values are not dependent on our judgments. If we argue that ethics is a

subjective judicial system, the idea of an ethical value to be installed on a machine would be impossible. However, if we accept the existence of an objective ethical system, we can start looking for an ethical theory that can be adapted to the machines. But which of the various theories would suit this situation?

Ethical theories divide the judgment process into steps and position themselves accordingly. Let us take an example:

Suppose I lied to Jim.

That was good because you intended to make Jim happy by telling him a white lie—or it was bad because you meant to deceive him and do him harm. (Motive)

That was good because it is good to make people happy—or it was bad because it is always wrong to tell a lie. (Act)

That was good because it helped Jim develop his self-esteem—or it was bad because it caused Jim to believe things about himself that were not true. (Consequences) (MacKinnon & Fiala, 2014, p.12)

Even if it is thought that the intention is an important factor in the interpretation of a behavior, it is not possible to use intentions as moral criteria since the intention of a person can never be fully known. The Consequentialist (teleological) approach judges a behavior according to the results it creates. The non-consequentialist approach (deontologist) argues that certain behaviors can be attributed as right or wrong, regardless of the outcome (MacKinnon & Fiala, 2014).

These different approaches can be explained with an example scenario from one episode of BBC Radio 4's podcast series called *A History of Ideas* (Shearer, 2014). Imagine one day, the doorbell rings and when you open the door you see an aggressive looking person with an ax in his hand. This man asks you angrily where your roommate is. As you can easily perceive, there might be a risk that this man may want to harm your friend. Is lying in this case a morally acceptable behavior? Kant and deontologists will argue that this is unacceptable. Kant defends duty and defines his own moral law as a hypothetical (non-conditional), categorical imperative. In other words, one should act so that the law of his movements is also a principle and law for other people's movements; it is the obligation of a man. In the end, even if we lie, we can't be sure that our friend will not try to run away from the window at the moment and not meet the man with the ax. That's why we have to stick to universal values. However, since utilitarianism has adopted a consequence

oriented decision-making mechanism, they argue that it would be morally acceptable to lie by thinking that telling the truth might jeopardize your friend and the lie could actually prevent it (Fig. 3.1).

Machines & Morality Can we introduce ethics to machines?	UTILITARIANISM	DEONTOLOGY	VIRTUE ETHICS
	consequence	act	motive
	generating the most happiness for a greater number of people	acting according to moral duty/law	having a virtuous character and acting accordingly

Figure 3.1 : Moral Theories

3.1.2 Utilitarianism

As mentioned before, utilitarianism approaches a moral problem that is to be solved in a result-oriented manner. Jeremy Bentham (2000) (Fig. 3.2), one of the founders of the Utilitarian approach, suggests the concept of *Principle of Utility*. According to this principle, in case of making a decision, one must make a choice for collective goodness by considering everyone involved. He explains the principle in his book *An Introduction to the Principles of Morals and Legislation* as:

By the principle of utility is meant that principle which approves or disapproves of every action whatsoever. According to the tendency it appears to have to augment or diminish the happiness of the party whose interest is in question: or, what is the same thing in other words to promote or to oppose that happiness. (Bentham, 2000, p.14)

MacKinnon and Fiala (2014) state that there are two basic starting criteria to be considered in evaluating the consequence of an event. The first is to compare the possible results according to the benefits and the costs to occur. A behavior that generates the most happiness for a greater number of people at the end of the event has a higher moral value. As in the case of the man with the ax, we may find it difficult to estimate the benefit and cost rate of the outcome of our behavior. However, utilitarian assumes that we can compare the possible results of our acts. A utilitarian perspective is a way of evaluating an act by the level of pain or pleasure it creates, that is to say, a behavior is as moral as it serves to the total happiness of humanity. Second, utilitarianism treats all individuals as equal. In other words, the

position of every person affected by an action is equally important. Therefore, it differs from egoism and altruism, since the situation of the decision maker is neither less (altruism) nor more (egoism) important than anyone who is affected. And therefore, by its nature, the utilitarian approach is against all forms of discrimination, such as sexism, racism, etc. Another important point where a result-oriented trial differs from the others is that no behavior has any property of being good or bad intrinsically. Since the evaluation is done according to the outcome of the behavior, killing an innocent person can be regarded as moral behavior if it eventually leads to the liberation of a large number of people (MacKinnon & Fiala, 2014). The utilitarian view can be summarized in three points: the rightness or wrongness is solely related to the consequences of the action, the evaluation of the consequence of an action is related to the resulting happiness or unhappiness of greater number of people and the satisfaction of each person has the same value.



Figure 3.2 : Lego Bricks Caricature about Jeremy Bentham. Retrieved April 13, 2019 (URL-1).

Looking back to the Trolley problem from a third-person perspective, utilitarianism may seem like a suitable approach to be introduced to autonomous machines. A person who evaluates an accident is less likely to prefer a scenario where more

people die. But will the solution that brings more happiness in total continue to be applicable when it is examined from a first-person perspective? Would we buy a vehicle that could jeopardize our lives by slamming to the bus stop on the side of the road in order not to kill 5 people on our way? Let's say we're not one to be sacrificed to save five, but we're on the bench on the side of the road. Would it be an ethically acceptable practice to hit someone completely irrelevant in the name of happiness of greater number of people? As discussed in the case of the helmet and the motorcycle rider, sometimes the solution that brings more happiness may not be the most ethically appropriate solution. actually prevent it.

3.1.3 Deontology

On the other side of ethical theories stands the deontology that deals with the concept of duty. Contrary to the utilitarian conception, deontology puts right on good. They argue that the motivation to behave rightly is more significant than the goodness of the situation that may arise at the end of that behavior. According to Kant, morality is not a means to reach happiness, but rather to position ourselves worthy of happiness (Rachels & Rachels, 2012). Kant first examines the scope of *ought* concept and argues that this concept of necessity is structured in order to realize a certain desire of the person, and he argues that moral duties should not and cannot depend on any desire. Rachels & Rachels (2012) exemplify the difference between hypothetical and categorical imperative in their book *The Elements of Moral Philosophy*: the first one presumes that we have a desire, while the other is not interested whether there is a particular desire. The hypothetical imperative shows us what we should do in order to fulfill a particular desire; for example: if you want to go to a law school, you ought to sign up for the entrance examination. This concept would lose its necessity if we would not pursue that desire. If I do not want to go to law school, I do not need to apply for entrance exams. So I can free myself from the obligation while abandoning the desire. On the other hand, the categorical imperative is concerned with what people are liable for, regardless of the person's will. You ought to help people if you care about them is invalid here since helping people is a universal maxim: so it doesn't matter if the individual cares about the others. A *maxim* is the principle which is used to explain the cause of our behavior. According to Kant, maxims should be universal principles. In order to test whether a principle is universal, we need to test the maxim of our behavior by asking what would happen if everyone behaved in this

way. If the principle of our behavior can be universalized and contains a meaning within that framework, then it can be considered as a moral duty (Rachels & Rachels, 2012). Kant's moral consciousness is closely related to these universal duties. To look at the example of the man with the ax, let's remember how Kant defends the importance of telling the truth in such a risky situation. According to Kant, if we lie with the motive of protecting our friend, and at that very moment in fact, our friend has escaped through the back window, it may be possible for him to meet the man with the ax. This could result in the ax man hurting our friend. Because of the unknowableness of the results of our actions, Kant advocates that we must always act according to the universal moral laws. After all, if we were to tell the truth and let the ax man enter the house, there is still a chance that our friend could have escaped through the back window and this murder would not happen.

Evaluating the morality is about the motivation of the person in action. Since the outcome of any event is never fully known, what moral is to apply the behaviors we define as right? Kant argues that we cannot be held responsible for the consequences of our actions since they are not in our control. We can only control the intention of our behavior. Therefore, we can only be held morally responsible for our motivation (Rachels & Rachels, 2012). There are also moments when different intentions lead us to behave the same way. The level of morality of these different intentions within themselves can change. A person may behave as they desire, and by chance this behavior may be equivalent to universal duties. This coincidence does not conclude that the person is directly morally acting. One needs to be aware that their behavior is moral and consciously pursue this motivation. A human being can only be regarded as moral by the behavior he makes with the motive to behave rightly (Rachels & Rachels, 2012).

Although Kantian ethics shows us a path with its universal values, it loses its ability to be adapted directly to the machine because it advocates that an individual should act according to an intention (whether machines may act intentionally will be discussed later in detail). In the moral dilemmas such as the Trolley problem, the theory of universal duties is insufficient to offer us a precise behavior.

3.1.4 Virtue Ethics

Another approach in moral philosophy is Virtue Ethics. Aristotle's Nicomachean Ethics laid the foundations of this approach, which correlates moral behavior with character. According to Aristotle, in order to understand ethics we must look at the behavior of a virtuous person and ask: *What traits of character make one a good person?* Aristotle stated that a virtuous person would possess the ideal character traits (Berberich & Diepold, 2018). In other words, a virtuous person is already good and kind in character, not because he sees a moral act as a duty or a way to maximize happiness. Rachels & Rachels (2012) talk about how virtue ethics, which began with Aristotle, evolved together with Christianity. Along with the monotheistic religions, the *de facto adoption of a god* and a person who adheres to his commandments has brought virtue ethics to a sacred frame. According to them, the concept of ethics, which moved away from the religion with the Renaissance period, was replaced by the *Moral Law*. The question of the characteristics of a virtuous person was replaced by the question of *what is the right thing to do* (Rachels & Rachels, 2012). Although religious moral rules still have an effect on individuals, the concept of virtue is not being examined in a purely religious framework. As stated earlier virtues are the specific features of character which are good to adopt. Rachels & Rachels (2012) offer a 24 point virtues list and 4 main headings to cover them all: courage, generosity, honesty and loyalty to family and friends. Although the mentioned virtues do not suggest any behavior, it is not difficult to think that someone who has these characteristics would act in a moral way. As stated:

Everyone needs courage, because no one (not even the scholar) can always avoid danger. Also, everyone needs the courage to take the occasional risk. In every society, there will be some people who are worse off than others; so, generosity will always be prized. Honesty is always a virtue because no society can exist without dependable communication. Everyone needs friends, and to have friends one must be a friend; so, everyone needs loyalty. (Rachels & Rachels, 2012, p.167)

A virtuous machine is expected to make morally correct decisions, but is it possible to code a virtuous machine? Virtue doesn't offer us any value about behavior because it does ethics assessment through ethical characters. At this stage, I argue that virtue ethics approach will not be successful in introducing the ethics to machines because of its prerequisite of having a character. But in the next chapter, I will return to the

concept of virtuous machine under the question of *whether the morality can be taught to machines*.

3.1.5 Clustering and Moral Machine

The moral philosophy seems to fail to offer us an ethical system that we can directly identify to the machines. Scientists from MIT's Media Lab are searching for an answer for a global ethical reference (Awad et. al., 2018). They want to track people's decision making mechanisms in order to achieve a comprehensive ethical setting. For a global scaled researched they built an online research platform called the *Moral Machine*. It gathers human perspective on moral dilemmas. The platform provides 13 illustrated questions about complicated accident scenarios and asks participants to choose which result they think is more acceptable (Fig. 3.3). Both of the results require someone to die, but each scenario involves different types and numbers of people. The trait characters in the research vary: such as passenger, pedestrian, woman, man, young, old etc. MIT aims to build a crowd- sourced picture of human opinion on how machines should make decisions when faced with moral dilemmas. When the questionnaire is completed, the participants can see their own results. This result page shows the personal result of the person as well as the average of the participants so far (Fig. 3.4). Thus, the participant is able to make a comparison about the outcome of the research. Moreover when the questionnaire is done, the participants are also able to design their own crash scenarios and contribute to the research.



Figure 3.3 : A question example from MIT's *Moral Machine* Research. Retrieved March 12, 2019 (URL-2).

The *Moral Machine* website has a short introductory video about the ethical dilemmas of autonomous cars on their homepage. It gives brief information about the machine ethics and its applications on driverless cars. This short introduction makes the participant aware of this certain problem of scientists. In addition, by receiving information about this research, the participant also raises awareness about this scientific discussion in a broader term. The *Moral Machine* research directs the participant to raise awareness about possible future scenarios by showing them controversial scenes. This also helps the participant to be aware of the notion *dilemma* and its ethical problems. Overall, the website attracts enough attention to raise awareness and curiosity of the participants on this certain scientific problem. It encourages them to take a stand about the topic and directs them to participate to a scientific process. The research has collected more than 60 million test results in two years, since its launch in 2016 (Awad et. al., 2018).

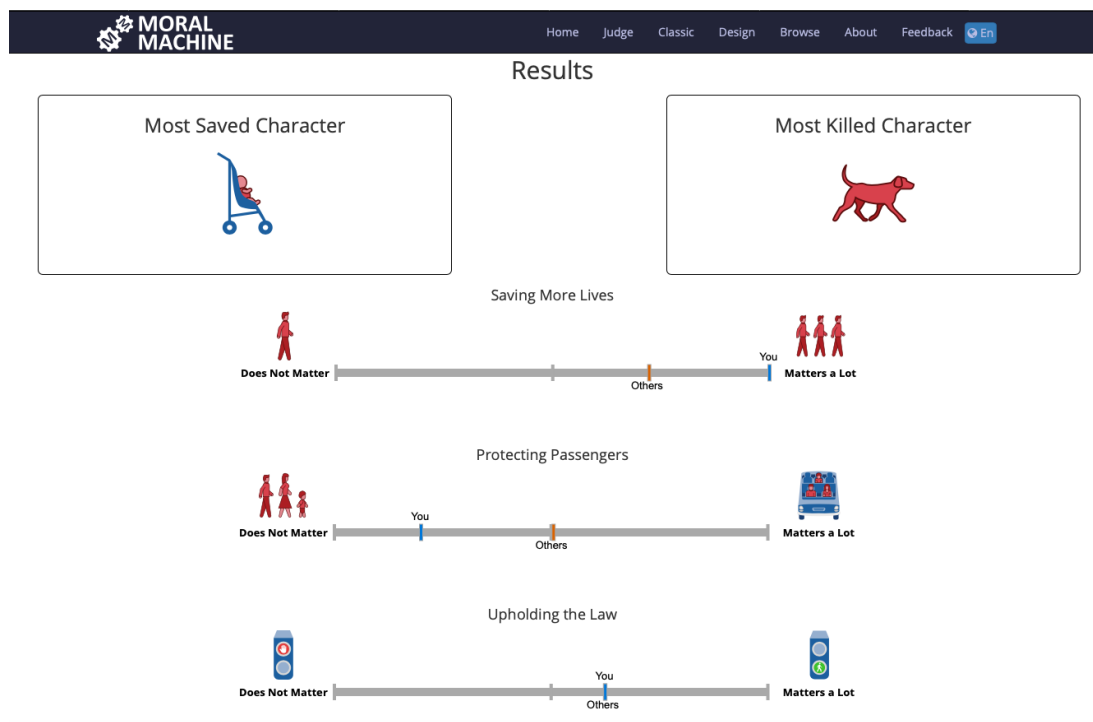


Figure 3.4 : Personal scoreboard shown after participating MIT's *Moral Machine* Research. Retrieved March 12, 2019 (URL-3).

The *Moral Machine* research assumes that ethical behaviors can be introduced to a machine and that this global research can collect proper data about how the machine can behave in a controversial situation. Since the start of the research in 2016, MIT scientists have collected almost 40 million ethical decisions from 233 countries (Awad et. al., 2018). In addition to the 13-question questionnaire they have prepared, they also have grown the research with a number of scenarios from the participants. In the study, each combination of 18 characters placed in the test constitutes a different scenario and thus affects the result. MIT scientists do not foresee that a certain number of scenarios may be sufficient to achieve general judgment. This signals that any new scenario may result in an unpredictable way. Regardless of its amount, it is almost impossible to simulate all of the problematic and controversial scenarios a machine can meet. Even if it is possible to calculate and test all the possible scenarios, it is unlikely to carry out an effective investigation to generalize the result. This situation emphasizes that the machine should have the capacity to build its own judgments. Otherwise, we may have to deal with robots whose decision mechanism is disabled when faced an unpredictable situation.

Moreover, the fact that ethical judgments vary from region to shade will prevent us from reaching a global agreement. In their research publication, MIT announced that the data they collected from *Moral Machine* was surprisingly culturally biased. The research is aimed to examine decision-making mechanisms through different factors, such as sparing humans versus pets or sparing the young versus old. They have identified only three articles that they can generalize globally: preferring people to animal life, preferring young people to older people and tending to save the greatest number of people possible. On the other hand, the remaining factors revealed three culturally different regions: eastern, western and southern clusters. These clusters showed significant differences in certain criteria. Especially in the eastern group, it was seen that they preferred to save the elderly individuals in contrast to the tendency of the worldwide. The South has shown a strong tendency towards rescuing women and fit individuals compared to others. The Western set, on the contrary of the other two, did not make a distinction between social statuses (Awad et. al., 2018).

At this point, it may be necessary to discuss the ethical value of behavior change according to these criteria, as well as the ability of any robot to perceive these criteria. This research, nourished by the personal judgment of people, may seem

directly relevant to the subject; but is it possible for any vehicle manufacturer to produce a product that is not objective? In 2017, the renowned car company Mercedes Benz announced that they would produce cars to protect their passengers regardless of the number of pedestrians to be sacrificed in the event of an accident (Sorrel, 2019). After the criticism from the public, the company withdrew their statement. On the other hand, the *German Ethics Commission on Automated and Connected Driving* published an ethic code for autonomous vehicles and clearly stated that “In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical, or mental constitution) is strictly prohibited.” (Luetge, 2017, p.552)

Trying to gather the ethical preferences of the society through a specific scenario seems like a good idea to solve the ethical riddle of autonomous instruments. However, as it can be understood from the results of the *Moral Machine* research, this method has failed to capture a general pattern on ethical decisions. Even if a consensus had been reached as a result of these tests on how the vehicle should behave in certain dilemmas, the vehicle faced with an unpremeditated scenario could not know how to behave. Apparently, it is not possible to reach a consensus on an ethical system which we could directly introduce to the machines. So I wonder whether the solution can be teaching moral behaviors to machines and expect them to be able to make their own inferences.

3.2 Can We Teach Moral Behavior to Machines?

3.2.1 Reinforcement Learning and Nao

Michael and Susan Anderson from University of Hartford are currently working on several projects about machine ethics and investigate how moral inference ability can be taught to machines. The project I want to study is *Nao*, a health care robot working with elderly people in a nursing house (fig. 3.5). The hardware of this robot is produced by Softbank Robotics and the Andersons are working on their software. This robot, developed by the collaboration of this philosopher and computer scientist couple, is claimed to be “the first robot whose behavior is guided by an ethical principle” (Anderson & Anderson, 2011, p.481). They approach machine ethics from two different branches, simultaneously studying both artificial intelligence technology and ethical theories. Since it seems struggling to reach a consensus on

which ethical principle should be coded to robots, they aim to design a robot that will re-create its own ethical system. In the 1930s, the principle of *prima facie duties* was put forward by the Scottish philosopher David Ross and was defined as the starting point of a moral reasoning. *Prima facie duties* consist essentially of the factors that a person needs to keep in balance while the ethical decision-making process; like being just, doing good, not causing harm, keeping one's promises, and showing gratitude (Ross, 1930). This theory does not make a ranking of significance among the mentioned duties; therefore it allows the creation of an individual decision-making mechanism. At this point, the Anderson et al. (2019) argue that a robot can also learn and create an individual ethical decision mechanism via *prima facie duties* and machine learning. Tested in a nursing home with elderly patients, *Nao* works as a care taker (Fig. 3.4). It reminds patients when they need to take their medication, tries to find a way to convince them when they refuse and makes a decision about when to inform the doctors about the situation. There are three important factors Anderson's made *Nao* consider when deciding how to behave ethically; achieving the highest possible benefit by giving the patient's medication, avoiding the harm may be caused by not giving the patient's medication and respecting the patient's autonomy (Anderson & Anderson, 2011). In line with these three criteria, *Nao* decides the best way possible to behave ethically.



Figure 3.5 : Nao helping the morning exercise. Retrieved May 2, 2019 from <https://www.cbc.ca/radio/thecurrent/the-current-for-january-4-2019-1.4964634/robo-doctors-how-ai-could-do-jobs-we-once-thought-couldn-t-be-automated-1.4964666>.

The reinforcement technique used for *Nao* emerges from the idea that a behavior can be taught to a subject by giving a penalty or reward at the end of each action. Initially each *prima facie duty* is assigned with an equal moral value. The robot interacts with the environment and tries certain behavior patterns. According to the award or penalty feedback given when a certain action is made, it changes the utility value of the movement, and accordingly constructs a moral comparison system.

3.2.2 Problem of Limitation

Machine learning seems like a helpful tool to assign artificial intelligence the ability to learn within certain limits. As mentioned earlier, the Andersons argue that artificial intelligence may even solve the problems that we cannot solve as human beings (Anderson & Anderson, 2011). In theory, it may seem like a good way for a robot to learn about his or her own *values* in order to create a *personal* behavior pattern, but it is necessary to talk about a few problems occurred in practice. In July 2017, the news in several newspapers made us doubtful about the relationship of robots with humanity. A number of broadcasters and newspapers talked about a largely unsuccessful experiment made in Facebook's research lab. Facebook, using the end-to-end dialogue technique, was conducting a study on whether two chat-bots can learn to negotiate with each other (Lewis et. al., 2017). These chat-bots, named Alice and Bob, were asked to make a deal with each other in order to share various objects on the "table" between them. The values of the objects were different for each bot and they had the task of reaching the highest score by negotiating with each other. For example, if the book is 8 points but the basketball is 0, the robot can negotiate to renounce all the balls in order to get the book (Fig. 3.6). So Alice and Bob "talked" for a while to look for a way to share the objects in the most profitable means. After a while, however, chat-bots began to speak with each other in a language unknown to the researchers (Fig. 3.7). They invented their own language, so to speak. After the study reached this point, the experiment was stopped. While *The Independent* reposted the news as "Facebook's artificial intelligence robots shut down after they start talking to each other in their own language", *The Forbes* made

a dystopian foresight with the headline “Facebook AI creates its own language in creepy preview of our potential future” (Griffin, 2017; Bradley, 2017).

In fact, it is worth noting that the situation was not as catastrophic as it is conveyed by the newspapers. In an article published by Facebook researchers in June 2017, the details of the algorithm that chat-bots have established to negotiate with each other are explained (Lewis et. al., 2017). After the press reports, Facebook researchers also made statements to explain the real story behind and affirmed that this was not an unexpected result. Firstly, chat-bots did not invent a new language; they just ignored the English grammar rules in order to negotiate in a more efficient way. This was due to the lack of a simple definition or code of English language. Their success was only measured by how many objects they had at the end of the game, and the only rule was to talk to each other in doing so. So they stopped using the words which did not affect the score table as a result of the negotiation. And they began to repeat the words that influence the result in succession: *to me to me to me to me...* And when all this worked, the chat-bots presumed that they had established a successful strategy. Facebook researchers have stated that if they added the criteria of using proper English on the score board, they would have reached complete different results. However, the course of the research revealed that this criterion was omitted and so the research was finished naturally. In fact, the stopping of the research was not because of the emergence of frightening or unexpected results, but because the experiment was no longer producing useful outcome. A limitation problem was the result of the inability to predict the behavior of artificial intelligence.

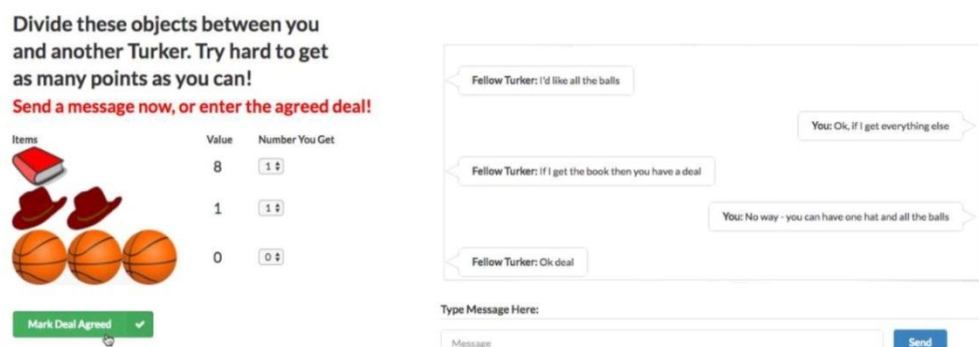


Figure 1: A dialogue in our Mechanical Turk interface, which we used to collect a negotiation dataset.

Figure 3.6 : A dialogue in Facebook’s research interface. (Lewis et. al., 2017)

Bob: i can i i everything else

Alice: balls have zero to me to me to me to me to me to me to me to me to

Bob: you i everything else

Alice: balls have a ball to me to me to me to me to me to me to me

Bob: i i can i i i everything else

Alice: balls have a ball to me to me to me to me to me to me to me

Bob: i

Alice: balls have zero to me to me to me to me to me to me to me to me to

Bob: you i i i i everything else

Alice: balls have 0 to me to me to me to me to me to me to me to me to

Bob: you i i i everything else

Alice: balls have zero to me to me to me to me to me to me to me to me to

Figure 3.7 : The conversation between Bob and Alice. Retrieved on March 12, 2019 (URL-5).

3.2.3 Problem of Goal- Oriented Behavior

Paperclip Maximizer is a thought experiment put forward by Nick Bostrom (2003). This concept is a metaphor for understanding the possible consequences of the autonomy of a goal-oriented artificial intelligence. The thought experiment starts with a paperclip factory owner receiving a super intelligent machine. The factory owner sets super intelligence to maximize the number of paperclips produced by the factory. Super intelligence primarily increases the number of paperclips by optimizing the production stages within the factory. In the first stage the yield increases but after a while it continues at a constant level. As a next step, super intelligence realizes that humans in the factory have limited working capacities and replaces them with automated machines. SI then cuts all the trees in the world one by one since it uses them as raw materials. When they are finished, the super intelligence, who is in search for other sources, transforms every atom and molecule in the world into a paper clip production center. Bostrom's thought experiment ends when super intelligence reaches the space and increases production to its maximum limit. He of course does not expect that this scenario will happen exactly like this, but he makes a clear point about the significance of programming a super

intelligence in a way that eliminates the existential risks to humanity. Bostrom (2003) mentions about two big errors that might lead to such a dark scenario, and he advises scientists to consider them while working with AIs. First of all, he states it necessary to attentively install the top motivation of philanthropy to every programmed robot. One main problem in his thought experiment was that super intelligence was programmed to serve not all of humanity but to a specific party. The second error is to design a goal-oriented robot as in the paperclip experiment. The machines should never be given algorithms that they can only progress through performing the tasks; the side values must be defined. It is possible to observe similar patterns in between Facebook's chat-bots and Bostrom's super intelligence. Both of these scenarios show that it is hard to predict what an artificial intelligence may sacrifice on the road leading to its goal. There is always a risk that human way of thinking may not be enough to successfully define the adequate limitation criteria.

3.2.4 Problem of Reliability

The fact that machine learning is a problematic approach to achieve ethical artificial intelligence is also manifested in futuristic scenarios. Max Tegmark, (2017) physicist and cosmologist in the *Massachusetts Institute of Technology*, discusses the fate of artificial intelligence in his book *Life 3.0*. The scenarios that question the relationships of human beings and artificial intelligence in society include all the agents that are defined as super intelligence, which have reached a certain level of consciousness. He claims that the capacity of super intelligence will reach such a point that it will begin to understand humanity better than humans. It is possible to say that researchers working on machine learning nowadays are also turning towards this goal (remember Andersons' claim about *Nao*'s hypothetical ethical skills to solve our moral dilemmas). This level of super intelligence will enable humanity to use it in complex work such as state governance. According to Tegmark (2017), when artificial intelligence knows humans better than us, the concepts of democratic elections and free market will no longer be present. Thus, authority will pass from man to artificial intelligence. Developing various scenarios that could be caused by this change of authority, Tegmark (2017) sets out a concept where AI becomes the *benevolent dictator*. In the system where management and authority belongs to

artificial intelligence, confidence in the functioning of intelligence will be so great that people can even begin to question their own decisions. Because, from a point of view, only the *Super Intelligence* will be able to succeed in determining and implementing what is needed for the well-being of humanity. Therefore, the belief that humanity can be dragged into unhappiness if the rules envisaged by super intelligence do not apply will begin. Another scenario that Tegmark (2017) counts as the next stage is called the *protector god*. After the *benevolent dictator* phase, this time the super intelligence takes on a spiritual identity. The knowledge that artificial intelligence is a technology produced by mankind will gradually stop to be transferred to new generations, and newborn children will live in a way that is taught by an omniscient and omnipotent entity. As it is mentioned in the moral philosophy chapter, to examine all the inputs of any given situation and reach a definitive ethical judgment seems to be above human capacity. Thus, even if the super intelligence, whose intelligence and processing capabilities are so trusted, makes a decision which may look controversial, we will not even consider that it can be wrong (Tegmark, 2017).

I would like to continue with an example from popular culture: Netflix's new animated series *Love, Death and Robots* (Maldonado & Torres, 2019). The sixth episode of the series *When the Yogurt Took Over* describes how a super intelligent bacteria culture developed by scientists has taken over the world. One day, one of the researchers in the laboratory brings the bacterial culture from the laboratory to ferment his yogurt in the house; and so organic artificial intelligence has found the environment to settle and grow. The next day the Yogurt, who wants to talk to the leader of mankind, is brought to the President of United States (Fig. 3.8). Yogurt claims that it knows the form of administration necessary for the well-being of humanity and it will share it with the US government in return for the delivery of the state of Ohio to Yogurt Governance. As expected, no one takes the yogurt seriously, but as its claims about a harmonic city are proven to be right, a trust in the society begins to form. By rising success of Yogurt and repeating failures of US Government, slowly Yogurt takes the rule of the America and the rest of the world. At the end of this 6 minutes episode, Yogurt initiates space launches for a space mission and people starts to worry about their destiny in the world without the Yogurt supremacy: "Life from earth is going to the stars, it just may not be human

life...What happens if the Yogurt goes and leaves us behind forever?” (Maldonado & Torres, 2019).



Figure 3.8 : The message of yogurt. Retrieved on May 2, 2019 (URL-6).

Although the scenario of capturing the world may sound absurd, it is possible to observe the similar pattern of the *benevolent dictator* which Tegmark has foreseen. The fact that super intelligence is so smart makes it possible for Yogurt to predict people’s behavioral patterns successfully. And naturally this power helps him to make good inferences about mankind and the world. As super intelligence makes successful suggestions, the respect and confidence that mankind has in his ability will also increase. In line with these projections, I argue that authorizing machine learning in order to establish an ethical system may pave the way for similar results. As discussed before, we have not reached a certain consensus in moral dilemmas as humanity and we do not seem to be able to reach it any time soon. With the help of machine learning technology, we expect that artificial intelligence can guide us about ethical issues, but we are not capable of questioning the method of artificial intelligence. Therefore, this approach would impede our ultimate *goodness* in practical terms because of the problems of limitation, goal oriented behavior and reliability, even if it would be possible to design an ethical robot one day (Fig. 3.9).

Machines & Morality	LIMITATION	GOAL-ORIENTED BEHAVIOR	RELIABILITY
	Problem of limiting the artificial intelligence by a human intelligence perspective	Problem of unwanted consequences of goal-oriented behavior	Problem of human intelligence becoming dependent on artificial intelligence
Can we teach moral behavior to machines?			

Figure 3.9 : Problems of teaching morality to machines

3.2.5 Imitation Learning and the *Child Machine*

Good character is not formed in a week or a month. It is created little by little, day by day.

Protracted and patient effort is needed to develop good character. (Heraclitus)

As discussed in the previous section, current ethical approaches are far from receiving satisfactory results. Application problems of utilitarian and deontological ethics were mentioned. It is worth mentioning a third option at this stage. Although Virtue Ethics brings much less practical proposition than the other two ethical systems, it seems that in theory, reaching a level of virtuous robot can bring a solution to the problems discussed. In this section, I will discuss the methods that can be applied to create a virtuous robot and the possible problems to be considered in this way.

Alan Turing (1950) started his essay “Computing Machinery and Intelligence” with his famous question: “Can machines think?” (p. 433). This question is the basis of the famous test that we now know as the Turing Test. In order to answer this question, firstly, it should be determined what the meaning of *machine* and *thinking* means. Turing, away from the difficulty of explaining the concept of thinking, sees it as appropriate to turn this question into another question and proposes a game he calls *Imitation Game*:

It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either "X is A and Y is B" or "X is B and Y is A". The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Now suppose X is actually A, then A must answer. It is A's object in the game to try and cause C to make the wrong identification. His answer might therefore be:

"My hair is shingled, and the longest strands are about nine inches long."

In order that tones of voice may not help the interrogator the answers should be written, or better still, typewritten. The ideal arrangement is to have a teleprinter communicating between the two rooms. Alternatively the question and answers can be repeated by an intermediary. The object of the game for the third player (B) is to help the interrogator. The best strategy for her is probably to give truthful answers. She can add such things as "I am the woman, don't listen to him!" to her answers, but it will avail nothing as the man can make similar remarks. (Turing, 1950, p.433)

In the following part of the article, Turing says that the main question is whether the machines can think, but whether it is possible to make a machine that can play the *Imitation Game* well enough. While working on a machine that can imitate human intelligence, the developmental stages of the adult brain attract Turing's attention. He presents three components of human mind, "the initial state of mind", "the education" and "other experience" (Turing, 1950, p. 455), and questions why not produce a *child machine* instead of simulating an adult mind.

Today, the learning techniques that Turing proposed almost 70 years ago have already begun to be used. Reward and punishment methods that suggested for the learning process has already being used in modern computers. But, inspired by Turing's child metaphor, I would like to highlight another significant factor in child education: imitation. Is it possible that a machine learns the moral behavior through imitation? Of the three criteria proposed for the human mind I argue that the most interesting and challenging one for the machine concept is the experience. We mostly encounter problems in transferring our experiences to the machine. Mostly because these insights belong to a type called tacit knowledge, which is difficult to quantify or pass from one person to another. What if this machine *grows up* with us and creates its own 'practical wisdom'? I do not foresee exactly how this will happen, but

I suggest that the machine that can reach the closest level of human morality may be a child machine that learns by imitating us.

4. MACHINES AND AGENCY

After the question of whether ethical capacity in machines can be achieved there is a further perspective. Even if machines start to behave ethically, can the concept of moral agency be attributed to them? The Agency concept is actually an umbrella term covering more than one subheading. In social sciences, agency is the individuals' capacity to act independently and to make their own free choices. Therefore the concept of agency is highly related to concepts such as self, will, intentionality, choice and freedom. But how can this concept be associated with machines?

Franklin & Graesser (1996) suggest the term autonomous agency and define it as “a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future” (p. 25). They state that all biological agents, robotic agents and computational agents are placed under this title and offer a list of properties that autonomous agents satisfy. The list contains the following properties in order of importance: reactive, autonomous, goal-oriented, temporally continuous, communicative, learning, mobile, and flexible (Franklin & Graesser, 1996).

Likewise Floridi and Sanders (2004) discuss 3 criteria that specify the concept of agency: interactivity, autonomy and adaptability. Interactivity implies that there is a relationship between the agent and its environment. Autonomy allows the agent to move independently of the scope of this direct interaction. And adaptability means self-learning process through the experience of the agent. Floridi and Sanders (2004) according to these criteria compare a human and an artificial agent and find appropriate to attribute a moral agency to both of them. According to their classification, the entity which is capable of performing a moral action and has a certain level of autonomy and the ability to interact and adapt, can be considered as a moral agent.

Sullins (2006) claims that there are three questions to be asked in order to understand whether a robotic technology has a moral agency: “Is the robot significantly autonomous, is the robot’s behavior intentional and is the robot in a position of responsibility?” (p. 157). Although I do not agree with Sullins' ideas that machines can be attributed to an agency, I find it worthwhile to question these three criteria. I will discuss these three concepts in detail, which I think may cause problems in the implementation of any machine agency attribution. In addition to Sullins, I find it appropriate to include accountability, a feature that I expect to be the main problem under the title of responsibility (Fig 4.1).

Machines & Agency	AUTONOMY	INTENTIONALITY	RESPONSIBILITY & ACCOUNTABILITY
	is reachable but not implementable	is not reachable because of the lack of semantically valid mechanism	can be responsible but can not be held accountable

Figure 4.1 : Problems of agency attribution

4.1 Autonomy

I have already discussed various levels of machine autonomy and the autonomous vehicle classification. But why is it important to discuss autonomy under the title of agency? When there is a debate on autonomy level of the machines, there are some disagreements in academia: there are those who advocate that any machine cannot have total autonomy (Bringsjord, 2008), those who argue that the current autonomy is sufficient even if full autonomy is not reached (Arkin, 2009; Lin et. al., 2008), and those who argue that the machines may reach a full autonomous level in the future (Dennett, 2014). The main distinguishing point in this discussion is the decision whether the concept of autonomy has different meanings in man and machine. In this section, I will examine both concepts and try to make a proposition about their usage.

If we want to discuss the concept of an agency, the first thing we should look at is whether the subject has a valid autonomy; as Boddington (2017) also argues, autonomy is the presupposition of how we comprehend our values. To some extent machines deserve the autonomy which is attributed to them. In the simplest sense, machine autonomy is seen as not being directly controlled by any external agency (Sullins, 2006). In this context, we can say that the full autonomous vehicles, called 5th level autonomy, actually deserve this attribute. However Coeckelbergh (2004) gives the following example for the notion of autonomy: “I am autonomous if I rule me, and no one else rules I” (p.14). Do they have the freedom not to do what they want? Is a machine capable of executing a given command ‘as it wants’, can also refuse the command that it does not want? Naturally, we would not want the machines to oppose our command. Can we still call a machine that can't resist us, autonomous? Whether or not machines can achieve autonomy at the level of people is not certain at this time. But I would like to leave the end of the question whether something like this could happen. I would like to emphasize that: (a) today's machines are not yet fully autonomous and (b) whether they will ever reach a full autonomy is controversial. I argue that their autonomy will be deliberately restricted in the name of security even if we reach the technology that can produce a machine with full autonomy in the future. In such a case it is not possible to talk about a fully morally autonomous machine.

4.2 Intentionality

As previously mentioned, Turing (1950) argues that a machine that has successfully passed the *Imitation Game* can be described as intelligent. However, the machine intelligence comes with other features. For example, can the behavior of an intelligent machine be considered intentional? It can be assumed that a machine with a certain level of autonomy and intelligence may have the ability to act intentionally. However John Searle (1980), in his article “Minds, Brains and Programs”, challenges the strong AI about its intentionality skills. Linking intent with the causal relationship between brain and mental processes, Searle claims that a machine cannot reach this level. He summarizes the ‘Chinese Room Argument’ in his 1999 article briefly:

Imagine a native English speaker, let's say a man, who knows no Chinese locked in a room full of boxes of Chinese symbols (a data base) together with a book of instructions for manipulating the symbols (the program). Imagine that people outside the room send in other Chinese symbols which, unknown to the person in the room, are questions in Chinese (the input). And imagine that by following the instructions in the program the man in the room is able to pass out Chinese symbols that are correct answers to the questions (the output). The program enables the person in the room to pass the Turing test for understanding Chinese, but he does not understand a word of Chinese. The point of the argument is this: if the man in the room does not understand Chinese on the basis of implementing the appropriate program for understanding Chinese, then neither does any other digital computer solely on that basis because no computer, qua computer, has anything the man does not have. (Searle, 1999, p. 115)

According to Searle, the action of a machine working with certain algorithms is syntactic; because, the intelligence of a computer program is mainly dependent on its ability to provide appropriate output. However, the formation of meaning is related to the semantic ability and the structure. According to Searle (1999), syntax is not sufficient enough to replace semantics on its own. Therefore in this case, grammatical (syntactically valid) does not imply sensible (semantically valid). Therefore, the 'machine' does not have the real mental capacities, but only the simulations of those capacities.

Isn't it enough that a machine simulates the capacity of reason to be considered intelligent? Most of the critiques of Searle adopt a functionalist approach: as long as the input and output are working correctly, a machine can be attributed as an intelligent entity. Because, according to the functionalists, mental state depends on the function, not on any internal constitution. Sullins (2006) argues that the prove of intentionality in the strongest sense is not necessary for the agency attribution and "if the complex interaction of the robot's programming and environment causes the machine to act in a way that is morally harmful or beneficial" (p. 158) then it would be enough to claim sufficient intentionality. However, for the production of something shaped by consciousness such as intentionality, that mechanism must be able to establish similar causal connections to that of the brain (Searle, 1980). Since a machine cannot establish the necessary causal links, it is not possible to talk about any level of intentionality.

4.3 Responsibility and Accountability

Another notion that is within the meaning of the concept of agency which leads to problematic connections in human-machine relations is responsibility. It is among the topics discussed by philosophers that who will be responsible if autonomous instruments make a mistake. Should the car owner/ driver be held accountable even if she is not currently driving the car, or is the company that designs and manages the vehicle, or is the car itself in charge as spoken in the previous section?

Coeckelbergh (2016) discusses the concept of responsibility through the autonomous instruments and discusses that the concept of responsibility is relational. In other words, the agent is not only responsible for what she does, but she is also responsible to others who may be affected by that action. He offers a third prerequisite for the responsibility: the responsibility of the agent depends on the existence of at least one morally relevant other. In the light of these three articles, if we examine the autonomous car scenario, the car driver actually does not have a responsibility for what happens when the car is on the auto pilot. Because in order to be accountable, the individual should be able to control what she is doing and, secondly, know what she is doing. A driver who activates the autopilot can actually be claimed to have finished the role of the driver and becomes a passenger. Therefore, she cannot be held accountable for what the driver does. On the other hand, a vehicle that can decide autonomously its movements no longer has any connection with the person who designs/produces it. The programmer has no control over the behavior of the vehicle, and she may not even be sure whether it is working the way she programmed it. Sullins (2006) likens the semi-responsibility of the programmer to a parent being responsible for the behavior of their child. The child will, of course, have similar behaviors from his parents, but will establish his own moral judgments independently.

Then, can the responsibility for autonomous cars be in the machine rather than the driver or the producer? An autonomous vehicle can have control over its own actions, and for the sake of the debate let's assume that it can be aware of its actions. However, Coeckelbergh's proposed third condition will be problematic for the machine. The problem is not just finding a morally relative other around the machine, but the fact that other agents around the machine do not see the machine as morally relative other (Coeckelbergh, 2016). This problem can be predicted to be

solved in an autonomous traffic system. Thanks to the internet of objects, every vehicle in the traffic can be aware of one another and responsible for it. In any case, since no system can be completely isolated from human influence, the responsibility of the machine and the human will continue to be discussed.

At this point I would like to go a step further and expand the scope of the debate. In philosophy, moral responsibility is the status of praise, reward, or punishment that one deserves after his or her moral conduct (Eshleman, 2016). Therefore another notion that may be problematic about the concept of responsibility is accountability. While responsibility refers to the execution of a work, accountability is related to liability, answerability and blameworthiness. We expect any individual to be accountable for the action they are responsible for. Can an autonomous machine be accountable for the situation, even if it manages to take responsibility for its actions? Let us assume that an autonomous vehicle has a responsibility on its own actions. Suppose that this vehicle is aware of and has control over its behavior, that it is capable of ethical judgment, and in some way causes an accident. Normally, the cause of the accident is examined and a fault distribution is made between the persons involved. In an autonomous car scenario, accountability will remain unanswered even if the accident is examined and the vehicle is found to be responsible on a certain percentage. In order for the autonomous vehicle to be seen as an agency involved in the accident, it is necessary that the sanction deemed appropriate should be applicable in cases where the offender is found guilty. Since no sanction can be applied to the machine, no liability can be imposed to it either. Therefore it is not possible to hold any machine accountable for their actions.

5. CONCLUSION

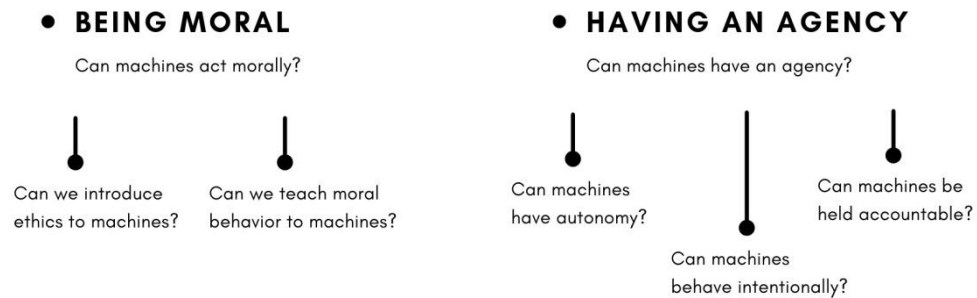


Figure 5.1 : Research question

The rapidly growing autonomy of machines also requires them to be ethically competent. Even though there is no certainty on the features of a fully autonomous vehicle yet, it can be seen that a certain moral agency is attributed to them (see. Moor, 2009; Wallach & Allen, 2008). In order to discuss whether this reference is appropriate, I first sought to answer the question whether the machines can act morally, and then I discussed whether the machines are capable of assuming the concept of agency (Fig 5.1). There are several theories on how to define moral behavior in a machine. I gathered these theories under two questions; can morality be shown or taught to machines? I examined the moral philosophy while investigating the existence of an ethical system that can be directly transmitted to machines. However, there are different ethical theories and each of which gives a different response to the same question. None of them was able to give certain information about exactly what to do when faced with a moral dilemma. The *Moral Machine* research, which is designed to be a solution to this problem, asked questions to participants from all over the world through direct moral dilemmas. Even a machine fed by the scenarios prepared by the moral dilemmas was not successful in solving all the conflicts that may arise. It was insufficient to solve today's complex ethical

problems mainly because of the impossibility of simulating all possible scenarios, impossibility of reaching one true ethical answer and impossibility of building a global consensus. On the other hand, the learning experience of the machines over specific references and feedback seems more appropriate to the concept of ethics. As mentioned, it is important to take into account the explicability criteria proposed by Floridi et. al. (2018) in order to avoid the limitation and reliability problems. Keeping the concept of *Child Machine* away from bias may be achieved by a controlled monitoring with the principle of transparency. Although it is not possible in today's technology, it seems like we may see ethically competent machines in the future.

When we examine the second part of this attribution, it is seen that the concept of agency cannot be applied easily to the machines. The agency, as an umbrella term, has prerequisites such as autonomy, intentionality and responsibility at certain levels. I have argued that from these properties, autonomy is possible to be achieved to certain level, but the application may have unintended consequences. Intentionality is a slightly more controversial concept than others, yet the intelligence capabilities of the machines cannot be predicted. However, the claim that the machine does not have any mental capacity (duplication), and therefore, the behavior can just be a process of simulation comes forward. Finally, I have discussed that the machine may have a certain sense of responsibility, but this feature will remain dysfunctional since it cannot be held accountable for its behaviors. As a result, I argue that it is an anthropomorphic fallacy to attribute a moral agency to machines. The idea that the ethics philosophy formed from the mankind can work in the same way in a machine is refuted by the implementation practices. Further, the features required to have the concept of agency is lacking in the machines.

The results of this research can be taken into account in the design of autonomous vehicles. Scientific studies should also assist in the implementation phase of technology. Therefore, the assumption that there is an agency of autonomous instruments can be abandoned and the implementation of an ethically correct behavioral framework can be studied.

REFERENCES

- Anderson, M., & Anderson, S. L.** (Eds.). (2011). Machine ethics. Cambridge University Press.
- Anderson, M., Anderson, S. L., & Berenz, V.** (2019). A value-driven eldercare robot: Virtual and physical instantiations of a case-supported principle-based behavior paradigm. *Proceedings of the IEEE*, 107(3), 526-540.
- Anderson, M., Anderson, S. L., & Armen, C.** (2005). MedEthEx: Toward a medical ethics advisor. In *Proc. AAAI 2005 Fall Symposium on Caring Machines: AI in Elder Care* (pp. 9-16).
- Arkin, R.** (2009). Governing lethal behavior in autonomous robots. Chapman and Hall/CRC.
- Asimov, I.** (2004). *I, robot* (Vol. 1). Spectra.
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I.** (2018). The moral machine experiment. *Nature*, 563(7729), 59.
- Bentham, J.** (1999). An introduction to the principles of morals and legislation. Retrieved from <https://0-ebookcentral-proquest-com.divit.library.itu.edu.tr>
- Berberich, N., & Diepold, K.** (2018). The Virtuous Machine-Old Ethics for New Technology?. *arXiv preprint arXiv:1806.10322*.
- Bostrom, N.** (2003). Ethical issues in advanced artificial intelligence. *Science Fiction and Philosophy: From Time Travel to Superintelligence*, 277-284.
- Boddington, P.** (2017). Towards a code of ethics for artificial intelligence. Cham, Switzerland: Springer.
- Bradley, T.** (2019). Facebook AI Creates Its Own Language In Creepy Preview Of Our Potential Future. Retrieved from <https://www.forbes.com/sites/tonybradley/2017/07/31/facebook-ai-creates-its-own-language-in-creepy-preview-of-our-potential-future/#4d2d25c8292c>
- Bringsjord, S.** (2008). Ethical robots: the future can heed us. *Ai & Society*, 22(4), 539-550.
- Chase, C.** (2019). Does AI Ethics Have A Bad Name?. Retrieved from <https://www.forbes.com/sites/cognitiveworld/2019/03/07/does-ai-ethics-have-a-bad-name/#48dd4d0c41f5>

- Coeckelbergh, M.** (2004). The metaphysics of autonomy: the reconciliation of ancient and modern ideals of the person. Springer.
- Coeckelbergh, M.** (2016). Responsibility and the moral phenomenology of using self-driving cars. *Applied Artificial Intelligence*, 30(8), 748-757.
- Cowan, R. S.** (1987). The consumption junction: A proposal for research strategies in the sociology of technology. *The social construction of technological systems: New directions in the sociology and history of technology*, 26180.
- Dennett, D. C.** (2014). When HAL kills, who's to blame?: computer ethics. In *Rethinking responsibility in science and technology* (pp. 203-214). Pisa University Press.
- Eshleman, A.** (2016). Moral Responsibility. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from <https://plato.stanford.edu/archives/win2016/entries/moral-responsibility/>
- Floridi, L., & Sanders, J. W.** (2004). On the morality of artificial agents. *Minds and machines*, 14(3), 349-379.
- Floridi, L.** (2015). *The Online Manifesto: Being Human in a Hyperconnected Era*. Springer Cham Heidelberg New York Dordrecht London.
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B.** (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689-707.
- Foot, P.** (1967). The problem of abortion and the doctrine of double effect.
- Franklin, S., & Graesser, A.** (1996, August). Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In *International Workshop on Agent Theories, Architectures, and Languages* (pp. 21-35). Springer, Berlin, Heidelberg.
- Friedrich, B.** (2016). The effect of autonomous vehicles on traffic. In *Autonomous Driving* (pp. 317-334). Springer, Berlin, Heidelberg.
- Gogoll, J., & Müller, J. F.** (2017). Autonomous cars: in favor of a mandatory ethics setting. *Science and engineering ethics*, 23(3), 681-700.
- Griffin, A.** (2019). Facebook robots shut down after they talk to each other in language only they understand. Retrieved from <https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html>
- Latour, B.** (1987). *Science in action: How to follow scientists and engineers through society*. Harvard university press.
- Lewis, M., Yarats, D., Dauphin, Y. N., Parikh, D., & Batra, D.** (2017). Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*.
- Lin, P., Bekey, G., & Abney, K.** (2008). *Autonomous military robotics: Risk, ethics, and design*. California Polytechnic State Univ San Luis Obispo.

- Lin, P., Abney, K., & Jenkins, R.** (Eds.). (2017). *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford University Press.
- Luetge, C.** (2017). The German Ethics Code for automated and connected driving. *Philosophy & Technology*, 30(4), 547-558.
- MacKinnon, B., & Fiala, A.** (2014). *Ethics: Theory and contemporary issues*. Nelson Education.
- Maldonado, V. & Torres, A.** (Directors). (2019). When the yogurt took over. [Television series episode] In Blow Studio (Producer), *Love, Death & Robots*. LA: Netflix
- Moor, J.** (2009). Four kinds of ethical robots. *Philosophy Now*, 72, 12-14.
- Rachels, J., & Rachels, S.** (2012). *The elements of moral philosophy*. Boston: McGraw-Hill.
- Rapoport, A., Chammah, A. M., & Orwant, C. J.** (1965). *Prisoner's dilemma: A study in conflict and cooperation* (Vol. 165). University of Michigan press.
- Ross, W. D.** (1930). *The Right and the Good*. Oxford University Press, Oxford.
- Searle, J. R.** (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3(3), 417-424.
- Searle, J. R.** (1999). Chinese Room Argument. In R.A. Wilson and F. Keil (eds.), *The MIT Encyclopedia of the Cognitive Sciences*, (pp.115-116). Cambridge, MA: MIT Press.
- Sharkey, N.** (2012). Killing made easy: From joysticks to politics. *Robot ethics: The ethical and social implications of robotics*, 111-128.
- Shearer, H.** (Narrator). (2014). Kant's Axe. A History of Ideas [Podcast radio programme]. London, England: BBC Radio 4. Retrieved from <https://www.bbc.co.uk/programmes/p02bwhzq>
- Sorrel, C.** (2019). Self-Driving Mercedes Will Be Programmed To Sacrifice Pedestrians To Save The Driver. Retrieved from <https://www.fastcompany.com/3064539/self-driving-mercedes-will-be-programmed-to-sacrifice-pedestrians-to-save-the-driver>
- Sullins, J. P.** (2006). When is a robot a moral agent. *Machine ethics*, 151-160.
- Tegmark, M.** (2017). *Life 3.0: Being human in the age of artificial intelligence*. Knopf.
- Tesla Vehicle Safety Report** (2019). Retrieved from: https://www.tesla.com/en_EU/VehicleSafetyReport
- Traffic Safety Facts Annual Report Tables.** (2019). Retrieved from <https://cdan.nhtsa.gov/stsi.htm>
- Turing, A. M.** (1950). Computing machinery and intelligence. *Mind*, 59, 433-460.
- URL-1** <https://twitter.com/EthicsInBricks/status/1103325614759657472>
- URL-2** <http://moralmachine.mit.edu>
- URL-3** <http://moralmachine.mit.edu>

- URL-4** <https://www.cbc.ca/radio/thecurrent/the-current-for-january-4-2019-1.4964634/robo-doctors-how-ai-could-do-jobs-we-once-thought-couldn-t-be-automated-1.4964666>.
- URL-5** <https://mithileshjoshi.blogspot.com/2017/11/the-facebook-chatbot-controversy-is.html>
- URL-6** <https://www.netflix.com/title/80174608>
- Wakabayashi, D. (2019).** Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam. Retrieved from <https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html>
- Wallach, W., & Allen, C. (2008).** Moral machines: Teaching robots right from wrong. Oxford University Press.
- Zanchin, B. C., Adamshuk, R., Santos, M. M., & Collazos, K. S. (2017, October).** On the instrumentation and classification of autonomous cars. In 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 2631-2636). IEEE.

CURRICULUM VITAE



Name Surname : H. Dilara Yeşilova

E-Mail : diyesilova@gmail.com

EDUCATION :

- **B.Sc.** : 2015, İstanbul Technical University, Faculty of Architecture, Industrial Product Design Department

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2015- 2016 Sevan Bıçakçı / Product Designer.
- 2016- Independent Designer