

KATOLICKI UNIWERSYTET LUBELSKI JANA

PAWŁA II

**On the Application of Norms
within Driverless Cars**

by

Michael P. Musielewicz

nr albumu 133726

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the

Wydział Filozofii

Katedra Podstaw Informatyki

January 26, 2019

“Thou shalt not make a machine in the likeness of a man’s mind.”

The Orange Catholic Bible

KATOLICKI UNIWERSYTET LUBELSKI JANA PAWŁA II

Abstract

Wydział Filozofii
Katedra Podstaw Informatyki

Doctor of Philosophy

by Michael P. Musielewicz
nr albemu 133726

In this dissertation, I will take up the question of whether driverless cars can be bearers of norms and are capable of being normative agents who can follow both legal and ethical norms. To answer this question we must first undertake three interrelated tasks. The first is to begin by examining what they are. The second task is to see if they are agents, and if so if they are also normative agents. Then finally we must undertake the task of seeing what sort of ethics is well suited for these normative agents. To resolve these tasks I will begin with a survey of what a driverless car is and where we are going with the technology, in addition to the current regulatory framework concerning these devices. The next chapter will address the issues of these devices agency and see if they are normative agents. The final chapter continues from the previous chapter and address how ethics is typically used in regards to driverless cars and finds shortcoming with other methods proposed. Finally I conclude by adopting a target centered virtue ethics, which I believe to be better suited for driverless cars.

Acknowledgements

First and foremost I would like to thank my family for all the support they have given me throughout the years. From the patience of my wife, Lily, who has helped me in so many ways as I worked on this thesis, and has motivated me to see it through. To my parents who have always supported me in my education and always encouraged me to test my limits.

I am also thankful for the support of my professors Piotr Kulicki and Robert Trypuz who have broadened my horizons in understanding deontic logic and its relationship to the philosophy of law they have been great colleagues who I have been working with on the *Permissions, Information and Institutional Dynamics, Obligations, and Rights* (PIOtR) project, where this research was supported by the National Science Centre of Poland (BEETHOVEN, UMO-2014/15/G/HS1/04514). I would finally like to thank professor Agnieszka Lekka-Kowalik, within whose seminar I received many points of advice and her interest in interdisciplinary studies has been incredibly influential upon my own work.

Contents

Abstract	iv
Acknowledgements	v
1 Introduction	1
1.1 The Topic at Hand	2
1.2 Outline of the Thesis	3
1.3 Sources to be Drawn Upon	4
2 Autonomous Cars in the World	7
2.1 Introductory Remarks	7
2.2 Technical and Terminological Aspects	8
2.2.1 The Build Up to Today	9
2.2.2 Today's Autonomous Car	12
2.2.3 Levels of Automation:	16
2.3 Inherent Ethical Problems	18
2.3.1 The Need for Transparency	20
2.3.2 Possible Factors Influencing Self-Driving Cars' Expected Conduct	21
2.4 The Current Regulatory Framework	24
2.4.1 The United States of America	24

2.4.2	The European Union, International, Supra-National and National Aspects	25
2.4.2.1	On the EU Level - Supranational and International Considerations	26
2.4.2.2	Netherlands:	29
2.4.2.3	Sweden:	29
2.4.2.4	Germany:	30
2.4.2.5	Spain:	30
2.4.2.6	The United Kingdom:	30
2.4.2.7	Austria, Hungary, and Slovenia	31
2.4.2.8	France:	31
2.4.2.9	Denmark:	31
2.4.3	China	31
2.4.4	Singapore	32
2.4.5	United Arab Emirates	33
2.5	Concluding remarks:	33
3	Normative Agency for Artificial Agents:	35
3.1	Introductory Remarks	35
3.2	Agents and Normative Agents	36
3.2.1	Autonomous Vehicles as Agents	37
3.3	Normative Agency	39
3.4	Grounding Rights:	60
3.4.1	On Legal Personhood	61
3.4.2	Is legal personhood for robots a solution?	67
3.5	Concluding Remarks	73
4	Ethics and Artificial Normative Agents	75
4.1	Introductory Remarks:	75
4.2	An Overview of Ethics in Driverless Cars:	77
4.2.1	There and back again a trolley problem:	77

4.2.2	In Popular Literature	82
4.2.3	In scientific literature	88
4.2.3.1	Ethics in Philosophy compared to Ethics in Computer Science:	89
4.2.3.2	Lin:	104
4.2.3.3	McBride:	109
4.2.3.4	Contissa, Lagioia, Sartor:	114
4.3	Concluding Remarks	118
5	Difficulties in Standard Accounts and a Solution	119
5.1	Introductory Remarks	119
5.2	Difficulties in applying consequentialist ethics:	120
5.3	Difficulties in applying deontological ethics:	129
5.4	A Virtue Ethical Approach:	131
5.4.1	The good driverless car?	138
5.4.2	The Target-Centered Virtues of a Driverless Cars:	154
5.5	Concluding Remarks:	161
6	Conclusion	163
6.1	Outcomes	163
6.2	Further research	165
	Bibliography	167

For the loves of my life Lily and Lizzie . . .

Chapter 1

Introduction

We are on the cusp of technological revolution. With advances in both robotics and artificial intelligence, “smart” robots are becoming more and more part of our daily life and are no longer safely relegated to partitioned sections of a factory floor. Examples of these new intelligent systems include the FROG (Fun Robotic Outdoor Guide) project to military funded research like MIT’s “Cheeta”, to Waymo’s (formally Google’s) driverless cars. All of these are making crossroads into our daily life which becomes more evident when we turn on the news, read popular science magazines, or simply walk outside. As these devices become more entrenched in our society, ethical questions, once relegated to science fiction, come to the fore.

These changes in technology, and its increased role in our lives, draw our attention to the ethical implications of these advances. Of the devices mentioned, there is a growing interest in establishing an ethics for autonomous vehicles, and in particular driverless cars. This interest is driven, in part, by various interest groups around the world, which include governing institutions like the US Federal Government [1], the European Union in its press releases [2] and the European Commission’s high level report GEAR 2030 [3], and

the German Federal Government (June 2017 report) [4]. Additionally, non-governmental interest groups like the Rand Corp. in [5], the IEEE¹ the ISO², and the SAE³ have taken it upon themselves to address ethical issues around this technology. Furthermore, in addition to the interests of both regulatory bodies and civil society, there is growing interest found within both popular literature and academia and in particular in philosophers who would like to know how we should understand these new entities entering into and interacting with our world. Are they agents, can they be moral, how can we even have an ethics for driverless cars?

1.1 The Topic at Hand

The traditional philosophical position is that, while technology in and of itself is neutral, ethical considerations are related to the user and not to the technology in and of itself. Despite the entrenchment of this traditional position, these robots that incorporate contemporary AI are not the same as other “less advanced” robots, viz. industrial robots, and once set up, may run with little to no further input from humans all the while interacting with humans. One such example of this new technology is the autonomous vehicle, or more colloquially, the driverless car. As these systems become more and more autonomous, the need to transfer ethical decisions from the user (that is to say the driver) to the system (in this case the autonomous vehicle itself), become more and more apparent.

Recognizing this is one thing and yet forming an ethics for these devices is quite another. Before we can propose an ethics for driverless cars there are several tasks that must first be accomplished, that reflect the nature of these ethical considerations. The first task is to see what the current state of affairs is for these devices, both in terms of what they are and what rules and norms,

¹Institute of Electrical and Electronics Engineers

²International Organization for Standardization

³Society of Automotive Engineers

and in particular laws, that have been both proposed and currently in effect, do apply to them. The second task is to see if these devices are agents, and if they are agents are they normative agents and how norms function in regards to the particular sort of agent that driverless cars are. The final task flows from the conclusion of the former two tasks, and here we will explore what sort of ethics applies to driverless cars, taking into consideration the sort of entity that they are and their sort of normative agency.

This dissertation will undertake these three tasks in turn, and I will dedicate a chapter to answering each. After which I will provide an answer to the question of how norms may be applied to autonomous vehicles. My contribution shall be the application of a target centered virtue ethics to autonomous moral agents, with a special emphasis on driverless cars as the special interest of this work.

In regards to the methodology of this work, the scope of this topic necessitates that we undertake an interdisciplinary approach. This is a result of the need to describe the entity we have in question, establish its normative agency, and then tackle the issues of ascribing norms (both legal and moral) to it. Chapter 2 will address the first task and is descriptive in nature and relates to the technical aspects of driverless cars. Chapter 3, is meta-ethical and incorporates theories from agency theories, philosophy of law – from both philosophers and jurists – and rights theory. The final two chapters 4 and 5, are chiefly concerned with ethics.

1.2 Outline of the Thesis

I will present my argument for these points in the following way. In chapter 2, *Autonomous Car in the World*, I will provide a description of what a driverless car is in terms of its development in 2.2.1 and where the technology is today in 2.2.2, in addition to some of the inherent ethical problems they have as seen in

the literature in 2.3. Following this, I will provide a survey of the legal norms in effect or that are currently under consideration at the time of the writing of this work in 2.4. In chapter 3, *Normative Agency for Artificial Agents*, I will address the issues related to how driverless cars may be considered normative agents. Here I will begin with a discussion of agency and its relationship to normative agency in 3.2. From here I will discuss issues concerning various theories of normative agency in 3.3 and supplement it with further discussion about the nature of legal personhood in 3.4. In chapter 4, *Ethics for Artificial Normative Agents*, I will provide an overview of the current state of ethics for driverless cars in 4.2 paying particular attention to the Trolley Problem and how it is reflected in popular and scientific literature in 4.2.2 and 4.2.3 respectively. I will then provide a backdrop for the principle ethics used in the popular and scientific literature with a discussion deontological and utilitarian ethics and raise difficulties each face in chapter 5, *Difficulties in Standard Accounts and a Solution*. Within this chapter I will then propose a solution that makes use of a target centered virtue ethics and address its applicability as an ethics for driverless cars in 5.4.

1.3 Sources to be Drawn Upon

Within these four chapters, different sources will be used that reflect the nature of the question that is under consideration.

In the second chapter, the primary literature includes material concerning the technical aspects, policies, laws, and reports. The literature about the technical aspects include: *Automated Driving in its Social, Historical, and Cultural Contexts* [6] by Fabian Kröger, *Driverless* [7] by Hod Lipson and Melba Kurman, and *Autonomous Driving: Context and State-of-the-Art* [8] by Javier Ibanes-Guzman et. al. Representative literature concerning the legal and policy aspects comes from a multitude of sources. For the European Union, among other sources we draw upon the 2016 Amsterdam Declaration

[9] and Gear 2030 report [3], foundational EU treaties such as the *Treaty on the Functioning of the European Union* [10] and the regulations like the *Council Directive 85/374/EEC of 25 July 1985* [11] and the *Motor Insurance Directive 2009/103/EC* [12]. Non-European sources include policies and bills from the United States such as the 2016 US Federal Government's policy [1], and bills like the *AV Start Act* [13]. Additional national laws such as the 2018 Road Traffic Act of Singapore [14] are also taken under consideration. Finally international treaties such as the Vienna and Geneva Road Traffic Conventions [15, 16] are considered in this chapter.

A selection of the literature examined in the third chapter includes works that concern agency in computer science and normative agency as understood in philosophy, especially drawing upon the philosophy of law, right's theory, and legal personality. The principle literature concerning agency are White and Chopra's *A Legal Theory for Autonomous Artificial Agents* [17], Robert Trypuz's *Formal Ontology of Action* [18] and Luciano Floridi's *Ethics of Information* [19]. Questions of normative agency make use of the theories of Ota Weinberger in his work *Law, Institution and Legal Politics: Fundamental Problems of Legal Theory and Social Philosophy* [20], Hans Kelsen in his *Pure Theory of Laws* [21], Wesley Hohfeld in his essay *Some Fundamental Legal Relations* [22]. Some of the crucial works about right's theory include Mathew Krammer's chapter *Rights without Trimmings* [23], and H.L.A Hart's article *Are there natural rights?* [24] in addition to Lied Wenar's *The Nature of Claim-Rights* [25], Neil MacCormic, *Norms, Institutions, and Institutional Facts* [26], and Aleardo Zanghellini's *Raz on Rights: Human Rights, Fundamental Rights, and Balancing* [27]. The literature about legal personality makes use of some of the previous literature with the important additions of some basic concepts found within Roman Law using Gordon Campbell's *Compendium of Roman Law Founded on the Institutes of Justinian* [28], Ugo Pagallo in *The Laws of Robots: Crimes, Contracts, and Torts* [29], and Cees van Dam's *European Tort Law* [30] for some particular features of legal personality.

The fourth chapter's literature covers the topic of ethics and addresses this both generally and in particular in regards to ethics for driverless cars as found both within scientific and popular literature. I begin with a description of the "Trolley Problem" as found within the works of Phillipa Foot in her article *The Problem of Abortion and the Doctrine of the Double Effect* [31], and was further elaborated upon by Judith Javris Thomson in her work *Killing, Letting Die, and the Trolley Problem* [32]. The following are the sources that translate this problem, in addition to other issues, into a popular format which includes various articles by Patrick Lin, such as his TED talk, *Atlantic* article entitled *The Ethics of Autonomous Cars* [33]. The Scientific literature includes a synopsis of ethics as seen in computer science with the textbook *Ethical and Social Issues in the Information Age* [34] written by Joseph Migga Kizza, and broad considerations of the typical ethics used in the literature. The typical ethics used are consequentialism and deontology, where we will examine the Stanford Encyclopedia's articles concerning these topics written by Walter Sinnott-Armstrong [35] and Larry Alexander and Alexander Moore [36] respectively. Particular considerations of the state of the art literature for ethics and driverless cars includes Lin's chapter *Why Ethics Matter for Autonomous Cars* [37], *Ethics of Driverless Cars* [38] by Neil McBride and Giuseppe Contissa, Francesca Lagioia, Giovanni Sartor in joint their work *The Ethical Knob: ethically-customisable automated vehicles and the law* [39].

A selection of the essential literature to be used in the fifth chapter of this work is as follows. This section is concerned with virtue ethics and includes Aristotle's *Nicomachean Ethics*, translated and edited by Rodger Crisp [40], Nicolas Berberich and Klaus Diepold in their article "the Virtuous Machine - Old Ethics for New Technology?" [41], and finally we will make use of Christine Swanton's *Virtue Ethics a Pluralistic View* [42], which was further elaborated upon by Liezl van Zyl in her chapter *Right action and the targets of virtue* [43] and apply this theory to driverless cars.

Chapter 2

Autonomous Cars in the World

2.1 Introductory Remarks

There is considerable discussion surrounding the creation and implementation of autonomous cars within society. Various regulatory bodies across the world are gearing up for the widespread introduction of these new devices into society. There is fierce competition among these institutions, to be the at the cutting edge of this revolution. The competitors in the race include the European Union and the United States in addition to other countries such as Singapore and China who are aggressively implementing their own policies. In a report for the European Commission regarding the GEAR 2030 projects [3], a sketch of the impact that driverless cars are expected to have upon society is provided. Here, the expected impact ranges from a 90% reduction in human-error-related road accidents to increased social mobility and even to a reduction of pollution in the environment [3, p. 40]. Similarly, the US federal government sees safety as the essential benefit of this new technology and hopes to see a reduction of up to 94% of traffic accidents in the US, along with increased mobility for disabled persons [1, p. 5]. There is much touting of the benefits that these devices are supposed to bring to our society, but what exactly is this new

technology and what sorts of regulations are being proposed to implement it and regulate both its testing and behavior?

In this chapter I will answer these questions by giving a brief history of the technological developments leading to today. After which I will also describe the current level of technology. Finally, I will give a survey of the various regulations that are being implemented or currently in effect as of the writing of this Ph.D. dissertation from around the world. All of this is to serve as a backdrop for our future consideration of how these devices ought to behave in a normative sense, which as we will see in the subsequent chapters, is dependent upon what they are.

2.2 Technical and Terminological Aspects

Autonomous cars are called by many different names including autonomous vehicles, driverless cars, and self-driving cars. Within this dissertation these terms will be used interchangeably.¹ In addition to the plethora of names that these devices go by, there is also differing understandings of what these devices are. For some these devices are robots, for others they are not. For some they are autonomous, for others not. One's understanding of the basic terminology used varies upon the speaker's background. To clarify what is meant in this work, this section will provide a survey of the technology used both in its historical aspects and contemporary usage and conclude with ethical difficulties that arise from their design.

¹Though the author recognizes the weakness of the term autonomous vehicles as it can also refer, among other things, to driver-less boats and autonomous aircraft and drones and other autonomous weapon systems outside of the scope of this work – though not entirely unrelated.

2.2.1 The Build Up to Today

The advent of autonomous cars is the result of advances in both hardware and software over the last century. Fabian Kröger, in his article *Automated Driving in its Social, Historical, and Cultural Contexts* in the book – “Autonomous Driving Technical Legal and Social Aspects” [6] gives an account of the history of the technological developments in automation that has lead us to this new technology. He begins his account with the sobering reminder that “[i]t is astonishing that the fulfillment of this wish [for self-driving cars] has always remained 20 years away for almost the last 100 years” [6, p. 41] despite this there has been tangible advances in the technology leading up to today. The earliest development came from the area of autopilot, which was brought about with the invention of two devices. The first device was the “gyroscopic airplane stabilizer” invented by Lawrence B. Sperry (1892-1923) and was first demonstrated in Bezons, France, on June of 1914 [6, p. 42]. Kröger recounts how, to an amazed audience, Sperry stood up in the cockpit with his arms held aloft while his mechanic crawled out onto the wings of the airplane, and the plane automatically corrected its balance. The second device that advanced autopilot is also found in aviation and was a system designed for automatic course stabilization that was invented by John Hays Hammond around the same time [6, p. 42].

In addition to these early 20th century advances in autopilot, the invention of radio technology, and in particular with its implementation into remote control of devices using radio waves (which was spearheaded by the United States’ military) propelled early advances for “self-driving cars”. The first prototype of a remote controlled vehicle was showcased in Dayton Ohio on the 5th of August, 1921 by the Radio Air Service. By the year 1925, a remote controlled car called the “American Wonder” was jointly developed by the Houdina Radio Control Company and the United States’ military. Vehicles like the *American Wonder* were demonstrated in “Safety Parades” starting in the 1930’s and

between the years 1931 and 1949 they were demonstrated in 37 of the then 48 states of the Union [6, pp. 43 - 44].

From here Kröger continues the story. Additional technologies developed during the Second World War – namely wire guides, magnetic detection devices, and radio detection – gave rise to the idea of the first infrastructure based “automatically guided automobile” that was developed by General Motors’ Technical Center in Warren Michigan. This vehicle completed its first test 1 mile (about 1.6 km) on February 14th, 1958 [6, p. 53], and captured the imagination of publications such as *Popular Science*. The successful test of an infrastructure based autonomous driving is aptly captured by a picture taken during its testing where a woman is seen riding in a car with her hands held up above the steering wheel [6]. What is important about this event was that it was the first demonstration of an operable driverless car that can be classified as being vehicle-to-infrastructure “connected driving”. Coupled with another early invention of cruise control which was invented by Ralph Teetor in 1948 [44], a limited sort of autonomous² driving became possible.

This possibility, however, was not without its disadvantages. Kröger describes that a major reason for why an infrastructure based automatic highway never came to fruition was, in part, because of the “gaps between [the] technical and [the] economic feasibility became too large” [6, p. 58]. Hinting to the economic feasibility, to build a complete system for all roads, these roads would need to be updated with the appropriate infrastructure, *e.g.* wire guides; in 1980 there was 3,859,837 miles (about 6,211,806 Km) of public roads having 7,922,174 miles (about 12,749,503 Km) of lanes present in the United States. If such a system were to be developed in 2010, those numbers would have increase to 4,083,768 miles (or about 6,572,188 km) of public roads and 8,616,206 miles (about 13,866,439 Km) of lanes.³ The re-development of such

²This would meet the Society of Automotive Engineers’ level one of autonomy “Driver Assistance” which is discussed later in this chapter.

³see <https://www.fhwa.dot.gov/policyinformation/statistics/2010/vmt422.cfm>

a large amount of public roads ultimately made such a wide-scale project unfeasible. To overcome this, a new solution was needed and the true revolution for autonomous vehicles occurred with the invention of microelectronics and its subsequent incorporation into vehicles in the 1970's in the United States and Japan.

The first attempts at an autonomous vehicle (in the sense that the device was independent of an external computer) occurred in Japan in 1977. Sadayuki Tsugawa and his team “from the Mechanical Engineering Laboratory in Tsukuba, Japan, presented the first visually guided autonomous vehicle that could record and process pictures (on-board) of lateral guide rails on the road by means of two cameras that the device had. The car was able to move with a speed of 10 km/h” [6, p. 58]. Early developments in the visual guidance of vehicles was first accomplished in the United States with the variations of the “Stanford Cart” going as far back as the 1960's, though the computers were at that time too large to have them on-board,⁴ with the work of Hans Moravec. By 1979, the Stanford Cart was able to move in lurches of 1 meter every 10 to 15 minutes [6, pp. 58-59]. During this time, other functions related to driver assistance were introduced, such as ABS in 1978, and were also being incorporated into cars with the introduction of microelectronics, which furthered the dream of autonomous cars.

The final part of the hard shift from infrastructure lane-based automatic driving began in Europe. Kröger reports that in 1984, Ernst Dickmanns with the University of the Federal Armed Forces in Munich created the first “visually guided autonomous cars with digital processors on-board, based on the perception of multiple edge elements” [6, p. 59]. This development to vision based autonomous driving was furthered in the European Union's EUREKA-PROgramme for a European Traffic of Highest Efficiency and Unprecedented

⁴and hence not autonomous in the sense mentioned above

Safety (PROMETHEUS) (1987-1995). The final test of this project was demonstrated with test of two of Dickmanns' S-Class Mercedes Benz in Paris in October of 1994. These cars were able to drive "more than 1000 km autonomously on three-lane highways around Paris, in the middle of heavy traffic and with speeds up to 130 km/h. The system was based on real time evaluation of image sequences caught by four cameras. Steering, throttle and brakes were controlled automatically through computer commands. For the first time, a machine vision system has been able to demonstrate its capability of deriving autonomously the decision for lane changing and passing" [6, pp. 59-60].

2.2.2 Today's Autonomous Car

Dickmanns' success in using image recognition in driverless cars leads us to the advent of autonomous vehicles in the modern sense. Most developers of driverless cars use a variety of techniques to develop the software that controls their vehicles. While there are particular differences, every one of them use some variant of statistical methods and machine learning (applying such tools as artificial neural networks, convolutions neural networks, deep learning, reinforcement learning etc.) as their primary tool. This is in large part due to the success that artificial neural networks, and relatively recent advances in deep-learning techniques, have had in providing their own answers to problems; which results in a way of programming control over these vehicles that is far more successful then the traditional rule-based expert systems.

As success of these techniques has taken off in the last couple of decades, so too has the development of these driverless cars. In the book *Driverless*, Hod Lipson and Melba Kurman recount the success of these techniques culminating with deep learning. As they tell it, modern deep learning took off with the creation of the ImageNet's annual Large Scale Visual Recognition Competition. This competition is broken down into the following three categories: 1) classification, 2) classification with localization, and 3) detection. Competitors

are asked to submit a software to the competition where it would then need to classify the contents of 100000 new images, and name the top five items in each image [7, p. 223]. The competition began in the year 2010 and was won by a team from the University of Illinois at Urbana-Champaign, where their program was wrong 28% of the time and the next two competitors were wrong 33.6% of the time and 44.6% of the time respectfully [7, p. 223].

The next year there was a slight improvement in the leading score, that is there was an error rate of 25 %, but the real change occurred in 2012. In that year's competition, a neural network named SuperVision, created by a team from the University of Teranto, drastically reduced the image recognition error rate to only 15 percent. The technique they used was a convolutional network, which was, at that time, seldom used [7, p. 224]. After that team's success, the team made the code that they used open-sourced and after that point all competitors used some form of a convolutional network. This culminated in the outstanding success of the Microsoft's Beijing team, which participated in the 2015 ImageNet Large Scale Visual Recognition Competition, where they were able to have a 3.57% error rate surpassing the human average error rate of 5% for the first time. After this tremendous success, Nvidia launched its own deep-learning neural network specifically aimed at driverless cars called Drive-PX [7, p. 225].

This sets the stage for where we are today. Developers of driverless cars are looking into some form of deep-learning technique joined with reinforcement learning as a way to have cars figure out, for themselves and during the course of thousands of simulations, how to drive when presented with both typical and atypical situations. Central to the operation of driverless cars is the question of navigation. In the chapter, *Autonomous Driving: Context and State-of-the-Art* in the "Handbook of Intelligent Vehicles" [8], Javier Ibanes-Guzman et. al., provide an overview of the general technologies used in driverless cars. Although particular technologies vary between the manufactures, they all operate on the same basic technologies to achieve the same basic functionality.

In terms of functionality, Ibanez-Guzman et al. lay out four main categories that driverless cars must satisfy. The first is localization, the second is mapping, the third is motion, and finally the fourth interaction [8, p. 1278]. In respect to each of these categories the car asks itself, “Where am I?”, “Where can I move?”, “How can I do it?”, and finally “How do I interact with others?” - this final question is where ethics and normative behaviour are to be found.

In terms of localization, the use of the standardized World Geodetic System (WGS84) facilitates this basic function. However, there is a basic problem in the fact that radio signals may be disrupted for a variety of reasons, e.g. skyscrapers and thunderstorms. One solution to this problem is to fuse the vehicle’s GPS data with its sensor data in order to provide alternative means of localizing its position within its environment. This process is accomplished by means of the device’s internal navigation systems and internal measurement units. This function of localization works hand in hand with the function of mapping, which aims at placing the car in its environment and help in directing it towards its goal. Ibanez-Guzman et. al. describe how these maps are formed within the driverless car’s data base. The general process is that of simultaneous localization and map building, where maps are to be understood as being a probabilistic distribution over environmental properties and not having fixed values [8, p. 1281]. This results in a mapping which gives the most probable position of the self-driving car as it is moving towards its goal within its environment.

In general, the information that is stored inside of the vehicle’s knowledge base is used to help it understand the “spatio-temporal relationship between the vehicle and its environment” [8, p. 1281]. The act of driving then builds a world model that allows for a better, or even correct, driving decision to be made by the vehicle. This world model is built by the signals received by the car from the car’s sensors. These signals are processed through its algorithms and as the car acquires information about the features of its environment, it employs a simultaneous localization and map building (SLAM) approach to

localize itself within said environment. According to Ibanez-Guzman et. al., new approaches “consider maps as probability distributions over environment properties rather than fixed representations of the environment at a snapshot in time” [8, p. 1281]. The environment is then modeled as a probabilistic grid and increases the certainty that the vehicle has in placing itself within its environment and building for itself a world model [8, pp. 1281-1282].

Having placed itself with a fair degree of certainty, the process of mapping sets out to answer the question of where the car can move. To answer the question of motion, the vehicle has two main goals; the first is the device’s global path while the other is the device’s local path. The global path determines the final destination of the vehicle while the local plan deals with immediate motions and obstacle avoidance [8, p. 1282]. Decisions are based with the sense–plan–act model, where the vehicle’s actions are based up its plan which varies upon the data that it collects from its sensor. Real world examples of this technology include electronic stability control and adaptive cruise control.

The interactive aspect of driverless cars relies heavily upon the vehicle’s ability to predict the actions of other objects within its environment. These objects include other cars and trucks, bicycles, horse-and-buggies, and pedestrians to name a few and is dependent upon a wide variety of factors. Ibanez et. al. describe the great difficulty in being able to predict the movements of some of these entities, namely, pedestrians. While some aspects of the road streamline the behavior of pedestrians, e.g. barricades, other areas where pedestrians frequently intersect the path of driverless cars become more problematic. Citing Lee and Abdel-Aty, Ibanez-Guzman et. al. state that: “statistics have demonstrated that the interaction of pedestrians with passenger vehicles at intersections results in a high number of fatalities where pedestrian and driver demographic factors, and road geometry, traffic and environment conditions are closely related to conditions leading to accidents” [8, p. 1285]. Part of the problem rests in the inherent impracticability of pedestrians, as opposed to vehicles which operate with fewer variables.

Ibanez-Guzman et. al. lay out a convenient outline for understanding the focuses of advancement within this field. Their three main categories are driver-centric, network-centric, and finally vehicle-centric. Driver-centric technologies aim at providing the human driver with relevant information and to aid in his or her operation of the vehicle. The network centric-category aims at creating a “intelligent” space for the vehicle to operate within. Here there are two subdivisions. The first was previously mentioned in the preceding section and is primarily infrastructure based, or alternatively called vehicle to infrastructure (V2I). The other avenue is vehicle to vehicle (V2V), where the vehicles can communicate relevant information to each other. Lastly, vehicle-centric technologies have their focus upon the car itself and include obstacle detection and avoidance, mission planning. As vehicle-centric technology develops, the need for a human driver diminishes and subsequently increases the levels of automation of the vehicle [8, pp. 1287-1288].

2.2.3 Levels of Automation:

There is a variety of standards that have been proposed by different institutions around the world in order to describe the increasing levels of autonomy that self-driving cars may possess. These standards include those proposed by the United States’ National Highway Traffic Safety Administration (NHSTA) and the German Bundesanstalt für Straßenwesen (BAST) and the levels proposed by the Society of Automotive Engineers (SAE). Of these proposals the SAE levels have become widely used in both the European Union (See the Amsterdam Declaration [9], and the GEAR 2030 [3] report) and within the United States of America on both the federal level (see the United States’ Federal Automotive Policy [1, p. 9] and the proposed SMART Act) and the state level (see the order to adopt for the testing of autonomous vehicles in the State of California [45, § 227.02.]. When relevant, this dissertation will make use of the SAE levels given its broad acceptance within the literature.

Summary of Levels of Driving Automation for On-Road Vehicles

This table summarizes SAE International's levels of *driving* automation for on-road vehicles. Information Report J3016 provides full definitions for these levels and for the italicized terms used therein. The levels are descriptive rather than normative and technical rather than legal. Elements indicate minimum rather than maximum capabilities for each level. "System" refers to the driver assistance system, combination of driver assistance systems, or *automated driving system*, as appropriate.

The table also shows how SAE's levels definitively correspond to those developed by the Germany Federal Highway Research Institute (BAST) and approximately correspond to those described by the US National Highway Traffic Safety Administration (NHTSA) in its "Preliminary Statement of Policy Concerning Automated Vehicles" of May 30, 2013.

Level	Name	Narrative definition	Execution of steering and acceleration/deceleration	Monitoring of driving environment	Fallback performance of dynamic driving task	System capability (driving modes)	BASt level	NHTSA level
<i>Human driver monitors the driving environment</i>								
0	No Automation	the full-time performance by the <i>human driver</i> of all aspects of the <i>dynamic driving task</i> , even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a	Driver only	0
1	Driver Assistance	the <i>driving mode</i> -specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment and with the expectation that the <i>human driver</i> perform all remaining aspects of the <i>dynamic driving task</i>	Human driver and system	Human driver	Human driver	Some driving modes	Assisted	1
2	Partial Automation	the <i>driving mode</i> -specific execution by one or more driver assistance systems of both steering and acceleration/deceleration using information about the driving environment and with the expectation that the <i>human driver</i> perform all remaining aspects of the <i>dynamic driving task</i>	System	Human driver	Human driver	Some driving modes	Partially automated	2
<i>Automated driving system ("system") monitors the driving environment</i>								
3	Conditional Automation	the <i>driving mode</i> -specific performance by an <i>automated driving system</i> of all aspects of the <i>dynamic driving task</i> with the expectation that the <i>human driver</i> will respond appropriately to a <i>request to intervene</i>	System	System	Human driver	Some driving modes	Highly automated	3
4	High Automation	the <i>driving mode</i> -specific performance by an <i>automated driving system</i> of all aspects of the <i>dynamic driving task</i> , even if a <i>human driver</i> does not respond appropriately to a <i>request to intervene</i>	System	System	System	Some driving modes	Fully automated	3/4
5	Full Automation	the full-time performance by an <i>automated driving system</i> of all aspects of the <i>dynamic driving task</i> under all roadway and environmental conditions that can be managed by a <i>human driver</i>	System	System	System	All driving modes		

FIGURE 2.1: The SAE Levels of Automaton of Driving

In figure 2.1, taken from [46], the various levels of automation in driving, as defined by the SAE, are described and are juxtaposed with the NHSTA and BAST levels of automation.

As we can see in the table, the levels of automation range from 0 to 5, with 0 being no automation whatsoever and level 5 being the total automation of the driving process. As automation increases, there is a diminishing participation of the human driver in the role of driving the vehicle until it is nonexistent.

Within the SAE levels, the order of the expanding functions that the system is expected to handle include is as follows.

1. Joint execution of steering and acceleration / deceleration (starting at level 1)
2. Execution of steering and acceleration (starting at level 2)
3. Monitoring of driving environment (starting at level 3)
4. Fall back performance of dynamic driving tasks (starting at level 4)

5. System capable to take over all driving modes (starting at level 5)

In regards to the levels of automation, there are two important aspects. Firstly, it has a legal bearing on the classification of the vehicle, which will be described in the final section of this chapter. Secondly, as the vehicle becomes more and more autonomous, the amount that a human should be “in the loop” in making ethical decisions is called into question and shall be discussed in further detail in the next chapter and gives rise to some inherent ethical problems which will be discussed in the next section.

2.3 Inherent Ethical Problems

As previously discussed in the preceding section, driverless cars operate through a mixture of sensors that build up the device’s world model. The world model is then put through a decision making algorithm that is created using machine learning techniques. This process leads to the device taking an action within its environment. A chief difficulty rests in how these decisions are made and the lack of transparency that exists in both how these decisions are made and the ethical framework that such decisions are rooted in. I discussed this problem in another work, that was published in 2018 as part of the papers accepted in the 2018 DEON conference [47], and the key excerpts from that work follow.

The European Commission’s report mentions that autonomous vehicles pose “new challenges for regulators and policy makers concerning e.g. road safety, environmental, societal and ethical issues, cybersecurity protection of personal data, competitiveness and jobs, etc. which need to be addressed” [3, p. 40]. Solving these issues is needed to build up the social acceptance of driverless cars.

A psychological factor also has to be considered. Although the predictions estimate that traffic safety will be significantly improved, many people are

afraid and prefer a human driver's control over vehicles or at least the possibility of a human driver to take control over the car. These fears surface in instances where self-driving cars, that are currently being tested, have failed to avoid serious collisions. Tesla's car in 2016 failed to detect a large white 18-wheel truck and trailer crossing the highway. The car drove full speed under the trailer, causing the collision that killed the 40-year-old behind the wheel inside the Tesla. Recently, an autonomous Uber car killed a woman walking across the street in the State of Arizona ⁵. From these examples we can see that the use of autonomous cars is not free from serious risks.

Even specialists in the area remain skeptical about the technology they create. Raj Rajkumar, a leading expert on robotics, who cooperates with General Motors in the construction of autonomous cars, describes the current status of the technology in the following way:

We as humans understand the situation. We are cognitive, sentient beings. We comprehend, we reason, and we take action. When you have automated vehicles, they are just programmed to do certain things for certain scenarios.⁶

So the users of autonomous vehicles want to know and understand (at some level of generality) how the vehicles are programmed to “do certain things for certain scenarios”. They want to be sure that in case of a hazardous situation or an accident a self-driving car will behave in a proper way. Yet we must consider, what do we mean when we say “proper way”? How should these vehicles operate when they move about their environments?

⁵See <https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe> (retrieved March 20, 2018)

⁶See <https://www.technologyreview.com/s/602492/what-to-know-before-you-get-in-a-self-driving-car/> (retrieved March 1, 2018).

2.3.1 The Need for Transparency

In most cases it is possible to avoid damage to property, health, and the life of passengers and other participants of traffic. Moreover, it seems credible that a well trained algorithm will perform far better in driving than the average human driver or even a very good driver, and so it would seem that ethical considerations for driverless cars are relegated to only extreme situations. But this is not necessarily the case. The effects that these devices have upon their users may differ depending upon how its program is made or trained. The US Federal Government’s policy for driverless cars indicates that, “even in instances in which no explicit ethical rule or preference is intended, the programming of an HAV [(highly automated vehicles)] may establish an implicit or inherent decision rule with significant ethical consequences” [1, p. 26].

However, the very ascription of values to these objects, resting upon implicit ethical values, must be made clear so that all stakeholders can ensure that these “ethical judgments and decisions are made consciously and intentionally” [1, p. 26]. This claim for transparency is mirrored in the report made by the ethics commission of the *Bundesministerium für Verkehr und digitale Infrastruktur* (hereinafter BMVI) made in June of 2017. Here, the BMVI underscores the importance of maintaining the autonomy of people in making ethical decisions and the prospect of some programmer or commission deciding how a driverless car should act on our behalf is, in and of itself, problematic [4, p. 16].

Hod Lipson and Melba Kurman write in their book *Driverless* [7] that drivers make countless calculations and risk assessments of their behavior and of the road as it unfolds around them. When drivers are thrown into a situation where life is at risk they must react accordingly. Do they swerve right and hit a wall, or hit some other vehicle? When it is people making these choices there is an air of spontaneity which allows for us to forgive poor decisions, however the same does not apply for autonomous vehicles. As they say:

those of us fortunate enough never to have had a severe traffic accident have not had to perform the uncomfortable task of publicly articulating why we reacted the way we did when faced with an unavoidable traffic accident. Driverless cars stir up consternation since they force us to publicly reveal this calculation. Even more challenging, driverless cars will require that, as a society, we agree on a uniform set of ethical codes that will guide the decision-making process of artificial-intelligence software when faced with an emergency [7, p.252].

But it is precisely this sort of “digging out” of our ethical calculations that will allow for transparency in this public debate.

In this aspect I concur that it is crucial for autonomous vehicles’ designers, and moreover for all stakeholders in these decisions, to make clear what hierarchy of values they embed in their vehicles. This clarification will enable the potential owners and users of self-driving cars, other traffic participants, the public in general, and regulatory authorities to accept or reject the underlying ethics in the vehicle’s decision making algorithms before the wide scale usage of such vehicles. Yet as of now it is difficult to implement given the machine learning techniques used in the development of these devices control algorithms.

2.3.2 Possible Factors Influencing Self-Driving Cars’ Expected Conduct

What kind of factors should be taken into account when the “ethical” behavior of self-driving cars is considered? Let us refer to some statements that can illustrate the breadth of possibilities.

Patrick Lin argues that the chief safety feature of driverless cars, that is its “crash-optimization”, implicitly means targeting which object to hit in order

to optimize a crash [48, pp. 72-73]. He notes that if we adopt a preference for minimizing harm to our property, the car would need to target objects of a lesser weight than the vehicle; yet if we wish to minimize the harm to other people's property, we ought to target an object of greater weight than the vehicle.

Michael Taylor from Car and Driver magazine, reported in [49] that according to Christoph von Hugo, Head of Active Safety in Mercedes-Benz Passenger Cars, all of Mercedes-Benz's future self-driving cars will *prioritize saving the people they carry*. Mercedes-Benz, after a public criticism, soon retracted the statement and indicated they would follow whatever the law prescribes [50]. That highlights the difficulties in pinning down the best response.

In general, can or should an autonomous car value one life more than another on the basis of their relation to that car (value the passenger or owner over other persons), age, sex, status or by applying some other criteria? These difficulties in our (in)ability to choose who to save are seen in the often discussed trolley problems, which will be later discussed in 4.2.1.

On this precise point, there are many different points of view. Take for example the report made by the BMVI. There they lay forth 20 ethical rules for automated and connected vehicular traffic. In the 9th rule they prescribe:

In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical or mental constitution) is strictly prohibited. It is also prohibited to offset victims against one another. General programming to reduce the number of personal injuries may be justifiable. Those parties involved in the generation of mobility risks must not sacrifice non-involved parties [4, p. 11].

These are fairly strong claims and are further supported by the first three articles of the *Grundgesetz für die Bundesrepublik Deutschland*, and raise questions

if such “targeting” of objects that happen to be people could even be constitutionally permitted within Germany. These claims are also seen in important associations in civil society. The IEEE (the Institute of Electrical and Electronics Engineers) also commit their members to these very same standards. Therefore, it would seem to answer our questions concerning whether a driverless car can value one life more than another.

Notwithstanding that apparent answer, there is more to the story than that. If we look at MIT’s Moral Machine (<http://moralmachine.mit.edu>), we see that people do in fact have preferences and seem capable of choosing between two bad options; and when they are given a series of choices of how to act in various dilemmas, general trends emerge. For an informal example, we can see that enforcing the law, preferring women to men, humans to animals, fit people to fat people, are some of the preferences that are noticeable. A more formal example of this is also seen in the work of Bonnefon et al. [51] where they noticed a strong preference for cars that minimize harm as such (i.e. by choosing self-sacrifice or the sacrifice of even loved ones) but it is conjoined with a general reluctance to buy such a car for themselves or even to have that sort of ethics enforced by legal means.

While the law itself has yet to say what ethics ought to be preferred, various policy documents from the U.S. Federal Government in its policy [1, p. 26], the European Union in its GEAR 2030 report [3, p. 40] and press releases [2] and the German ethics commission of the BMVI in their report [4], all emphatically assert the need for ethics for driverless cars. However, *there is no one clear understanding what is meant by ethics*. Rather, we find that most legislative texts are concerned with the testing and implementation of driverless cars that are on lower SAE levels of automation. A survey of the current legislation is provided in the next section.

2.4 The Current Regulatory Framework

In addition to the various corporations working on the development of this technology in its technical aspects, various regulatory authorities around the world have begun to prepare themselves for the dawn of autonomous vehicles by developing policies and regulations for them, that is to say they are busy preparing their normative aspect. Notable players include the United States of America and its constituent states, the European Union and its member states, China, Singapore, and the United Arab Emirates. These authorities, and their constituent members, are in various states of drafting and/or enacting legislation to regulate this rapidly developing field. These preparations range from preliminary discussions all the way to full fledged legislation regarding the testing and operation of these cars. In this section I will provide a survey of current (as of the writing of this thesis) legislation, declarations, and policies related to the implementation of these vehicles.

2.4.1 The United States of America

Legislation and policies regarding the implementation of autonomous vehicles currently exist on two levels. The first level is the United States' Federal Government, and the second level is the local states. On the federal level, an evolving guideline implemented by the National Highway Safety Administration (NHSTA) lays out the various policies to be implemented within the United States of America and the general direction that they wish to see legislation move towards. This policy is entitled "A Vision for Safety 2.0" [52] and was published in September of 2017. It builds upon the previous guidelines that were published in September of the previous year. These two texts provide clarifications of the various roles manufacturers, the states of the union, and the federal government bear respectively. Even though adherence to these guidelines is voluntary in nature, they provide a good outline for these various stakeholders until further legislation is passed.

In addition to these guidelines on the federal level, there are also currently two bills under review in the United States' Senate. The first bill currently under consideration is entitled "American Vision for Safer Transportation through Advancement of Revolutionary Technologies Act" or the "AV START Act"[13]; in the House of Representatives, there is a similar act that was passed and is also under consideration in the US Senate entitled "Safely Ensuring Lives Future Deployment and Research In Vehicle Evolution Act." or the "SELF-Drive Act" [53]. As of October 2017, both of these acts aim at providing a national standard for the testing of autonomous vehicles within the United States. Particular interest is given to the areas of safety standards, cyber-security, transparency about the technology to consumers [54].

Beneath these federal actions, there are also various states that are implementing either state legislation or executive orders to streamline the testing of driverless cars within their respective jurisdiction. The states and a federal district that have enacted legislation include: California, Nevada, Utah, Colorado, Texas, Louisiana, Arkansas, North Dakota, Illinois, Tennessee, Alabama, Georgia, Florida, South Carolina, North Carolina, Virginia, Pennsylvania, Washington D.C., Connecticut, New York, Vermont. States that have executive orders include, Washington, Idaho, Arizona, Hawaii, Minnesota, Wisconsin, Ohio, Delaware, Massachusetts, and Maine. It is important to note that within all of these jurisdictions there are various degrees of regulation ranging from meetings to discuss the formation of rules to full fledged legislation.

2.4.2 The European Union, International, Supra-National and National Aspects

Beginning with the Amsterdam declaration of April 2016, the EU and its members have begun the process of developing and implementing a new regulatory framework for the implementation of driverless cars. In their effort to do this, the various member states of the Union have outlined their vision for what

to do. Importantly, there is a need that such a framework must take into consideration the laws and regulations on a national, supra-national, and even international level – given the institutional nature of the European Union – and harmonize these various levels with each other, which is itself a Herculean task.

2.4.2.1 On the EU Level - Supranational and International Considerations

On the international level, there are two treaties that officials of the EU need to bear in mind. The first is the Convention on Road Traffic, completed in Vienna on 8 November 1968, [15, pp. 10-11] (hereinafter the Vienna Convention) and the second is the Convention on Road Traffic, signed in Geneva on 19 September 1949 [16, p. 18] (hereinafter Geneva Convention). These conventions form the bedrock of the international system for traffic rules within the European Union. These conventions are related and the Vienna convention replaces the Geneva Convention for all contracting parties.

These conventions allow for the mutual recognition of driver's licences and international driver's permits (cf. annex 6 and 7 respectively of the Vienna Convention [15]), and the Vienna Convention leads to the Vienna Convention on Road Signs and Signals - signed 8 November 1968- which standardized street signs and signals (fulfilling the desires listed in article 4 of the Vienna Convention and article 17 of the Geneva Convention), and provides basic definitions and classifications of vehicles (article 1 of the Vienna Convention). However, article 8 § 1 of both the Vienna Convention and the Geneva Convention poses a problem for the implementation of autonomous vehicles. Article 8 § 1 of the Vienna Convention states “Every moving vehicle or combination of vehicles shall have a driver” [15, p. 11] and similarly Article 8 § 1 of the Geneva Convention states “Every vehicle or combination of vehicles proceeding as a unit

shall have a driver” [16, p. 28]. While these provisions would seem to be common sense it proves difficult when there is no driver in the case of autonomous vehicles.

Recognizing this problem, the legislative bodies of the European Union sought an amendment to the Vienna Convention where they added to article 8 a section that states:

Vehicle systems which influence the way vehicles are driven shall be deemed to be in conformity with paragraph 5 of this Article and with paragraph 1 of Article 13, when they are in conformity with the conditions of construction, fitting and utilization according to international legal instruments concerning wheeled vehicles, equipment and parts which can be fitted and/or be used on wheeled vehicles. Vehicle systems which influence the way vehicles are driven and are not in conformity with the aforementioned conditions of construction, fitting and utilization, shall be deemed to be in conformity with paragraph 5 of this Article and with paragraph 1 of Article 13, when such systems can be overridden or switched off by the driver [55, p. 9].

This addition to the convention allows for the testing of autonomous vehicles with contracting parties provided that the vehicles are in conformity with international law, or failing that the systems that are not in conformity may be overridden and turned off by the driver. Article 48 of the convention, indicates that for contracting parties, the Vienna Convention replaces and terminates the relations established by the Geneva Convention [15, p. 41]. However, member states that are party to the Geneva Convention and not the Vienna Convention would still be beholden to article 8 of the Geneva Convention, although efforts have been made to remedy this in [56]⁷ but have

⁷cf. <http://www.unece.org/fileadmin/DAM/trans/doc/2015/wp1/ECE-TRANS-WP1-149-Aadd-1e.pdf>

as of yet to succeed as seen in [57], though it should be noted that within the Geneva Convention contracting parties have more independence to interpret the articles leading to them implementing driverless cars before such changes.

On the supranational⁸ level, the legislative bodies of the European Union have begun work at regulating these devices. These regulations however are limited solely to the powers given to the Union within the founding treaties of the Union, viz. article 2 The Treaty on the Functioning of the European Union (hereinafter the TFEU) and article 1 of the Treaty on European Union (hereinafter TEU). Currently the Union has a wide degree of competences over motor vehicles and in particular two aspects of the implementation of autonomous vehicles within the Union. The first aspect is concerned with the facilitation of the movement of autonomous vehicles pursuant to measures found within article 114 of the TFEU [10, p. 28]. The second aspect is related to EU action concerning civil liability for damages caused by autonomous vehicles. Relevant articles of the founding treaties include,

1. articles 26 and 114 of the TFEU to ensure that the common rules and procedures guarantee the widespread implementation of autonomous vehicles inside of the common market [10, p. 28],
2. article 169 TFEU to ensure “high level protection by adopting measures to secure the economic interests of consumers and their right to information” [10, p. 28],
3. article 173 TFEU to boost the competitiveness of the EU automotive industry within a global economy [10, p. 28],
4. and, the Union’s commitment to legal harmonization between its members [10, p. 28]

In addition to these competences, there are two important directives, which are binding upon member states, that are applicable to autonomous vehicles

⁸I opt for this term to highlight that the European Union has a unique identity as an international institution with some characteristics of statehood while not being a state.

within the Union. These directives are: Council Directive 85/374/EEC of 25 July 1985 [11] and the Motor Insurance Directive 2009/103/EC [12] and arguably an expansion of Directive 2007/46/EC [58] that is concerned with the approval of motor vehicles and their technical systems, and finally the 2010 Directive on Intelligent Transport Systems 2010/40/EU. Furthermore, there are two current policies (which are non-binding) that include the 2017 resolution of the European Parliament on the civil law rules of robotics [59] and the 2016 Amsterdam Declaration [9].

Below this level there are also various national policies that members of the Union have, or are currently in the process of creating, working regulations to allow for the use of autonomous vehicles. These countries include the Netherlands, Sweden, Germany, Spain, the United Kingdom, Austria Hungary and Slovenia, France, and Denmark. A short survey of the actions made by each of these countries as of April 20th 2018 follows.

2.4.2.2 Netherlands:

The Netherlands has implemented favorable policies for the testing of autonomous vehicles within its jurisdiction. The Dutch Ministry of Infrastructure and the Environment has opened Dutch public roads to large scale tests of autonomous vehicles. There is also draft legislation which would allow for the testing of these vehicles even without a driver present in the vehicle, though remote control of the vehicle would still be needed by a human. All applications for testing of these vehicles is handled by the Dutch Vehicle Authority [60].

2.4.2.3 Sweden:

As of July 1st 2017, Sweden has passed an ordinance by the Swedish Transport Agency, in accordance with the Vehicle Act and Vehicle Ordinance, which

allows for the testing of autonomous vehicles in Sweden, provided that a driver is present either physically inside or outside of the vehicle [61].

2.4.2.4 Germany:

The German Bundestag has enacted amendments to the Road Traffic Act to allow for the testing of autonomous vehicles that enter into force after June of 2017. The act permits the driving of a vehicle in autonomous mode, provided the driver is capable of taking over driving tasks when needed, though the driver need not always pay attention [62].

2.4.2.5 Spain:

In regards to Spain, permission falls under general motor vehicle testing. Prior to March of 2016, Spain benefited from being only one of two members of the EU (the other being the United Kingdom) not a party member to the Vienna Convention, as such road testing of autonomous vehicles was not prohibited by the Vienna Convention and was solely under local national laws.

2.4.2.6 The United Kingdom:

The United Kingdom – The United Kingdom also benefited from not being a party member to the Vienna Convention, which previously prohibited testing. It is currently permitted under the UK's Department of Transport Guidance, and notably requires the presence of a driver / safety operator [63, p. 9]. In addition to that policy there is currently a three year study which has been called to review the technology and its bearing within the law (e.g. who takes responsibility if things go wrong) before more liberalized laws will be considered [64].

2.4.2.7 Austria, Hungary, and Slovenia

On March 25th 2018 Austria, Hungary, and Slovenia have begun a partnership to conduct joint tests of autonomous vehicles within their borders, and have commissioned an Austrian Based Tech Firm, the Austrian Institute of Technology, with the task of studying how to best implement autonomous technology in these nations' transportation systems [65].

2.4.2.8 France:

Currently within France, testing of autonomous vehicles falls under the ordinance entitled “d’Ordonnance n 2016-1057 du 3 août 2016 relative à l’expérimentation de véhicules à délégation de conduite sur les voies publiques” [66] where various responsibilities are given to different ministries to permit the testing of autonomous vehicles on public roads. The ordinance is waiting for its legalization and has only had its first reading in the Senate of the National Assembly [67].

2.4.2.9 Denmark:

Regulations within Denmark consist of changes made in their Road Traffic Act in May of 2017, which has made the testing of autonomous vehicles possible provided that they are of SAE level 4 of autonomy.

2.4.3 China

Chinese policy on autonomous vehicles currently only exists on a local level with the national government planning to have nationwide regulations in place in the future. On the local level, two cities currently have regulations in place to allow for the testing of these devices. The first city to implement trials of

autonomous vehicles was Beijing on December 15th 2017,⁹ which was followed by Shanghai on February 27th 2018.¹⁰ The regulations of both of these cities focus upon the basic requirements needed for autonomous vehicles to be on the roads. Principally speaking the testers need to be licensed, insured, equipped with safety equipment, and have a test driver in the vehicle. Notably, in the event of an accident, it is the the test-driver who assumes responsibility for traffic violations and accidents, though liability rests with the test applicant (i.e. the corporation) due to the principle – agent relationship between the test applicant and test driver. The national government is also working on passing laws to allow for autonomous vehicles to be broadly introduced into the country and the government’s actions fall within the “New Generation Artificial Intelligence Development Plan”.¹¹ This also includes mandates from the central government concerning the target goals for the number of cars having partial automation (50% by 2020 and 80% by 2025), and with highly or fully automated vehicles entering the market by 2025.¹²

2.4.4 Singapore

Singapore is also trying to enter the competition for autonomous vehicles. In February of 2017 their Parliament amended the Road Traffic Act to accommodate for the testing of autonomous cars within its jurisdiction. The Road Traffic Act in section 6c through 6e provides the requirements for testing, the exceptions to the requirements, and notably fines for people who interfere with the lawful execution of those tests [14, § 6C]. What is unique about these rules

⁹cf. http://www.bjjtw.gov.cn/xxgk/tzgg/201712/t20171218_189568.html and <https://www.chinalawinsight.com/2017/12/articles/corporate/beijing-regulations-on-self-driving-cars-road-testing/>

¹⁰cf. <http://www.sheitc.gov.cn/cyfz/676771.htm> and <https://www.chinalawinsight.com/2018/03/articles/corporate/shanghai-issues-regulations-on-self-driving-cars-road-testing/>

¹¹cf. http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

¹²cf. <https://ihsmarkit.com/research-analysis/Chinese-government-drafts-policies-autonomous-vehicles.html> and <https://www.twobirds.com/en/news/articles/2017/global-at-a-glance-autonomous-vehicles#3>

is that while they limit when trials may be done, they exempt the human operator from taking responsibility in the event of an accident [68].

2.4.5 United Arab Emirates

Development in the United Arab Emirates is spearheaded by His Highness Shaikh Mohammed Bin Rashed Al Maktoum, ruler of Dubai and vice-president and prime minister of the UAE. The principle goal is to have 25% of all cars in the UAE be autonomous by 2030, along with the introduction of legislation to help streamline this process [69]. Additionally, in support of this initiative, the UAE has made a contract with Tesla and has purchased 200 autonomous vehicles (currently to be driven in semi-autonomous mode) that are to be used as a taxi service; the first 50 cars being delivered in September of 2017, which are currently at Dubai's airport, with two more instalments of 75 vehicles over the next two years [70].

2.5 Concluding remarks:

As we have seen in this chapter, driverless cars are far from being solely in the realm of science fiction. In terms of its technology, we have progressed by an incredible amount since the early days of airplane stabilization, and course correction and the remote controlled vehicle “the American Dream” that was once toured throughout the United States. Nevertheless, we are still a long way off from the higher levels of automation, that is SAE levels 4 and 5. At these levels, the human has been completely “removed from the loop” and it is here that the inherent ethical problems arise in regards to what sort of ethics ought to be implemented inside of the vehicle and in how to make the decision making process transparent for society in general.

Nevertheless existing regulations and stated policies relate to the testing of these vehicles and the safety standards that ought to be employed for their

operation on public roads with brief mentions of the need for “ethics” for these devices. As it always has been in the long history of driverless cars, the documentation cites the various boons of these devices ranging from saved lives and time, economic and environmental benefits, and even ethical considerations. But what exactly is required for us to properly consider an ethics for driverless cars? How are we to govern their behavior in the world?

Chapter 3

Normative Agency for Artificial Agents:

3.1 Introductory Remarks

When we want to consider the question: “How should we govern driverless car’s behavior in the world?”, there are two horns that we need to grasp. The first horn is the car’s behavior as an entity that operates within a specific context that is itself governed by legal norms. The second horn concerns itself with the car’s non-legal or moral behavior. Both of these horns often coexist and are inter-related to one another, though they need not always be so. To extend the metaphor, these horns are connected to a head, which is the agent themselves and depending on what aspect you are considering, you may see them as a moral agent or a legal agent, or if you are looking at both at the same time, a normative agent. Here I will present arguments in favor of understanding driverless cars as being normative agents, which entails their status as legal and moral agents, or capacity for ethics in general.

To do this, I will present the following arguments. The first is that driverless cars are capable of being agents. This argument will rest upon their status

as artificial agents. Then we will consider how they are normative agents, that is, agents that are capable of bearing rights and duties, and other normative obligations and permissions. Here we will consider two theories of rights – which are the will and interests theories of rights – to justify this. Finally, we will consider whether or not they can be legal agents and can be capable of being entities that may be held responsible for their actions, by making reference to the notion of legal personhood, which goes hand in hand with the former argument. This chapter includes my work in a co-authored¹ paper entitled: *Towards a Formal Ethics for Autonomous Cars* [71] published in “Deontic Logic and Normative Systems: 14th International Conference, Deon 2018, Utrecht, the Netherlands, 3-8 July 2018” [47] and another forthcoming paper of mine *Who should you sue when no-one is behind the wheel? Difficulties in establishing new norms for autonomous vehicles in the European Union* which will be published in a book entitled “Robots and Well Being” by Springer.

3.2 Agents and Normative Agents

As mentioned previously, this section assumes the task of establishing autonomous vehicles as agents and then from there establishing their status as normative agents. To begin this task, it is crucial to lay out what is meant by norms and ethics. As this section deals with meta-ethical concerns, that is, whether or not an ethics for driverless cars exists, it takes these terms at their broadest level. Here I understand the terms ethics and norms to be interchangeable and they are about rule-following. These rules may be social, moral, or legal and I understand the terms “ethics” and “norms” to capture these meanings.

Furthermore, these norms should be legitimate. I understand legitimacy to be those rules which come from an authority on some particular subject matter.

¹The co-authors being Piotr Kulicki and Robert Trypuz.

That is to say that laws originate from the state, morals from moral authorities, and social norms from the customs and mores of a particular society. Each of these types of norms constitutes a particular normative system that often overlaps, and even conflicts with each other.

A poignant problem in designing ethics for driverless cars is the establishment of these devices as normative agents that operate within a given normative system. If we are to do this, there are several factors that need to be considered. First, we need to determine if they are agents. Then if they are agents, we must decide if they are normative agents. The transition from agent to normative agent requires that the entity in question is an agent that is capable of bearing norms as such and second it is placeable inside of a “normative system.” This, however, is no small feat and will depend greatly upon one’s theory of norms. It is only once we have established this, that the movement towards an ethics for driverless cars makes sense. For although attributing normative agency to computer programs seems to be quite natural for computer science oriented logicians, for many legal theorists and philosophers (and ethicists in particular) it is still strange, so in this section we will argue for the aforementioned points.

3.2.1 Autonomous Vehicles as Agents

To begin we need to establish that autonomous vehicles are in fact agents. There are various senses of agency that are used in various fields. In a plain sense, being an agent simply means being an entity that has the capacity to act. There are, however, other more technical uses of the term. The most natural place to start is with a consideration of agency within computer science, where White and Chopra say (citing another author) that in this field an agent is “a piece of software that acts on behalf of its user and tries to meet certain objectives or complete tasks without any direct input or direct supervision from its user” [17, p. 6].

Trypuz, in *Formal Ontology of Action*[18], provides more qualifications on this definition and has created a good list of features that artificial agents have as found in the literature, having the following attributes:

1. is autonomous;
2. is situated (embodied in or inhabits an environment);
3. is reactive – senses its environment and is responsive to changes in the environment;
4. acts upon its environment;
5. is proactive – has a set of goals or tasks;
6. contains inner representations of itself and its world;
7. is rational –“acts in its own best interest, given the beliefs that it has about the world”;
8. has the ability to perform domain-oriented reasoning;
9. is a persistent (software) entity; and
10. has social ability – interacts (negotiates and cooperates) with other agents (and possibly humans) via some kind of agent -communication language: it engages in dialogues and negotiates and coordinates transfer of information [18, p. 40]

While this definition suffices, it is good to look at a more succinct definition to further our understanding of the agency of artificial entities.

In the book *Ethics of Information* by Luciano Floridi [19, p. 141] another definition for agency is developed. In this work Floridi defines an agent in the following way: “agent =*def* a system within and a part of an environment, which initiates a transformation, produces an effect or exerts power on it over time.”

This definition, however, proves to be too much, as it would allow entities ranging from earthquakes and people to count as agents within a given system. To remedy this, Floridi provides the following attributes that should also be included in this definition to narrow the list of agents:

- Interactivity – the agent and environment act upon each other;
- Autonomy – it is able to change its state without direct response to interaction; and
- Adaptability – it can change the transition rules by which it changes state[s]

Taking these additional attributes, we can weed-out earthquakes from counting as agents but can have other natural and artificial agents that are interactive, autonomous and adaptable be agents within some system.

Given both of the definitions provided by Trypuz and Floridi, driverless cars seem to meet the well established criteria for being agents within the computer science community. Yet although this definition suffices for agency in the computer science community, it is not sufficient for the establishment of this agent as also being established as a normative agent.

3.3 Normative Agency

The problem of establishing normative agency rests in the nature of norms themselves. When we make normative claims against other people, it presupposes that they are both agents and moreover bearers of duties adjoined to those normative claims. However, in the case of driverless cars, it is difficult to pin down against whom people can make these sorts of claims. If we want to make autonomous vehicles agents within the normative system described in the following sections, their status as norm bearers has to be established. Here

I will examine two prominent theories of norms, will and interest theories, and their bearing on driverless cars.

When we consider norms, as rules that govern the behavior of various agents, we notice that they have two related aspects. First they set the bounds of obligated, forbidden, or permitted actions and second these actions are ascribable to agents within the system, (i.e. a normative agent, who is beholden to these rules by virtue of being in the system).

To make these features clear, let us consider an example provided by Ota Weinberger in *Law, Institution and Legal Politics: Fundamental Problems of Legal Theory and Social Philosophy*[20], where he offers an example of a game of chess to describe what he calls the institutional nature of “social normative systems”.

The rules of the game of chess are defined by its basic conditions: chessboard, figures, starting positions, rules of operation etc. We might ask whether these rules should be regarded as normative rules or as definitions. If they were mere definitions the person who does not adhere to the rules would not be seen as infringing the ‘duty of the chess-player’, but simply as not playing chess.[footnote omitted] It is true that nobody is obliged to play chess; the rules of chess apply to the players not as a system imposed by society but only as a result of a voluntary participation in the game; but they are relevant for the possibility of setting acts since they lay down a behaviour in accordance with a duty and define the class of possible results of the game: the game which is won (or lost) [20, p. 193].

If two players sit down to play chess, they voluntarily enter into a sort of norm-governed activity constrained by the normative rules of the game. Their moves are permitted (such as a pawn may move two spaces in its first move),

obliged (a pawn may only take other pieces that are in its diagonals and one space away), or forbidden (a pawn may not capture a unit directly in front of it) in respect to the rules of the game being played. It is important to note that these norms are not merely definitions of certain classes of acts. If they were, as Weinberger stated, it would be impossible to cheat (and be caught and punished for cheating) in a game.

To illustrate this, let us consider another example. Player one moves their pawn in such a way that player two may capture player one's pawn *en passant* with a pawn of their own. Here player one's actions have created a permission for player two to use that rule of the game. While player two may not actually use that rule (say for strategic reasons), if they do use the rule, then player one is obliged to allow for the move (and thereby lose their pawn). In this example we see that one player's actions and the other player's reactions are circumscribed by the rules of the game they are playing and the current state of affairs of the board. These set the bounds of acts that each player may take relative to the game they are playing, though the choice of which action to take is (for the most part) up to them.

Weinberger expands this conception of normative systems as a game into broader considerations of law and into other norm-based systems. For our purposes we can see how traffic fits within this framework. The driver (and drivers in general) are duty-bound to obey traffic norms, that is to say the drivers by the very act of driving become the "players", the traffic norms constitute the "rules of the game" that they are "playing", and the current state of affairs of the road are much like the game board. The key difference consisting in the complexity of the system, the content and number of norms (described in the previous section as combining moral, legal, and social rules) and the price of failures (in terms of tickets or even possible damage to life, limb, and property).

To make this more explicit, let us turn the chess example given by Wienberger above. We begin to layout the foundations for this sort of norm-governed activity in the following way. The rules of driving are defined by its basic conditions: roads, vehicles, signage, traffic rules etc. We might ask whether these rules should be regarded as normative rules or as definitions. If they were mere definitions, the person—or in this case the driver—who does not adhere to the rules would not be seen as infringing the 'duty of the driver', but simply as not driving, or, for that matter, even as being a driver.² This institutionalized account captures the notion that within norm-governed activities, such as driving, the act and the norm that enables the act are intrinsically connected.

Provided that we are dealing with normative rules, we first note that each norm has both objective and subjective content. This distinction has been noted by the jurist Hans Kelsen in his work *Pure Theory of Law* [21]. There the objective content of the norm is the norm itself as it is positioned within the broader legal system of posited norms. The norm in its subjective sense is the norm as it is related to the addressees of the norm itself. On the subjective level, we find that these norms contain several key features. First they prescribe a certain form of behavior with either an obligation, permission, forbearance of some sort attached to it, and second these actions are imposed upon some agent(s). These actions can be broadly construed as "rights", which can be distinguished between rights over my own behavior or the behavior of other persons and even over (the use of) things [21, p. 75].

Bearing that in mind, we now face the question of whether autonomous vehicles can fit into this system. Let us begin with a further examination of norms and in particular the relationship that exists between agents when norms are acted upon. The American jurist, Wesley Hohfeld, in his text *Some Fundamental Legal Conceptions as Applied in Judicial Reasoning* [22], lays out his well-known schema for understanding the multiple variations on the term

²cf. [15]

right. Noting that there are various meanings of the word right, he specifically lays out various jural relations that express the various understandings of the term. The first jural relation is the right – duty relation. However, to avoid further confusion we will use the alternative term he provided for right, which is “claim”. The other pertinent definitions are non-claim – privilege, power – liability, and finally immunity – disability. When we consider typical norms such as “Thou shalt not kill!” we see that the command, a prohibition of killing other persons (in general) is imposed upon us and a conjoining duty towards others not to kill them. This can also be used for other ethical claims such as, “Maximize the good!” or “avoid committing a forbidden act!”

While the above schema is convenient in that it allows people who wish to study norms (or rights broadly construed) to use more analytic tools, the topic of building an ethics for driverless cars poses a unique problem for it. Normally it is rather simple to use this framework when we apply it to various normative systems, whether in a game of chess or driving a car. The agents are well defined, and so are the rules, and problems are typically introduced when there are normative conflicts or moral dilemmas. When applying this theory to driverless cars, we first first need to answer the question “are they really normative agents?” How can they be bearers of rights in the broad sense? Or to use Hohfeldian terminology, can they be part of the duty – claim, privilege – non-claim, liability – power, and disability – immunity relationships [22, p. 30]?

To underscore the problem this presents, let us provide two examples. The first example is common place and consists of a person who is trying to cross the road at an uncontrolled³ intersection, which is clearly marked by a crosswalk. Legislation on this topic is common place, and so I will take an example from the State of Minnesota, a state within the United States, where they have enacted the following law:

³It is uncontrolled in the sense that there is no stop sign, stop light etc. not in the sense that there are no rules governing the situation.

169.21 PEDESTRIAN.

Subdivision 1. Obey traffic-control signals.

Pedestrians shall be subject to traffic-control signals at intersections as heretofore declared in this chapter, but at all other places pedestrians shall be accorded the privileges and shall be subject to the restrictions stated in this section and section 169.22.

§Subd. 2. Rights in absence of signal.

(a) Where traffic-control signals are not in place or in operation, the driver of a vehicle shall stop to yield the right-of-way to a pedestrian crossing the roadway within a marked crosswalk or at an intersection with no marked crosswalk. The driver must remain stopped until the pedestrian has passed the lane in which the vehicle is stopped. No pedestrian shall suddenly leave a curb or other place of safety and walk or run into the path of a vehicle which is so close that it is impossible for the driver to yield. This provision shall not apply under the conditions as otherwise provided in this subdivision.

...

(d) A person who violates this subdivision is guilty of a misdemeanor. A person who violates this subdivision a second or subsequent time within one year of a previous conviction under this subdivision is guilty of a gross misdemeanor[72].

In this norm-governed situation, both the pedestrian and the driver have to follow the statutes. Here the pedestrian has a claim to cross the road unimpeded, which arises from 169.21 sub 2 (a) which places a duty upon a driver to slow down and come to a stop and allow the pedestrian to cross. There is also a duty placed upon the pedestrian not to “suddenly leave a curb or other place of safety and walk or run into the path of a vehicle which is so close that it is impossible for the driver to yield” and the driver of the other vehicle has

that claim upon pedestrians. Is the driverless car beholden to the duty to “let pedestrians cross the street”, can pedestrian make use of the claim of this right against the driverless car?

A more drastic example can also be taken from the well known trolley problem. An unmanned driverless car is going down the street and it is faced with a dilemma. Its breaks have failed, and now its controlling algorithm needs to make a choice of hitting a person in its lane or two people in the lane next to it. Now a dilemma is introduced if the car is beholden to the rule “Thou shalt not kill!” or say “Maximize the good!” or “Don’t commit a forbidden act!” But is this the case? Against whom can the people in this perilous situation invoke a duty not to kill them? People in general? Sure, but in this case that’s vapid. The programmer? Of course, but s/he programmed it not to hit people already. How about the driverless car itself? That’s not clear. Likewise, it is not clear that the car is even beholden to the maxims “Maximize the good!” or “Don’t commit a forbidden act!” Moreover, if the answer is no, then it would seem that the car does not have a duty to “not kill!” nor a duty to “maximize the good!” nor even a duty to “avoid committing a forbidden act!” and would ipso facto not have any such duties that could correspond to any person’s right in such a situation, in much the same way as we would not ascribe normative agency to a bull or a falling rock. Yet, if it lacks normative agency, then it falls outside of the normative system, leaving us in a de facto situation where nothing is forbidden for it, and therefore either everything is implicitly permitted or it is a non-entity. Moreover any questions about its de jure situation fall moot as nothing is obligatory, permitted, or forbidden as it falls outside of the normative system. But surely that cannot be the case, can it?

It is now clear what is at stake if driverless cars are not incorporated into this “game”, but how can we justify their status as a player? If we are to avoid the problems of not having them in the loop, we need to dig even further into the theory of rights. There are presently two prominent theories of rights

“will theory” and “interest theory”, which have been laid out by Mathew Kramer in his chapter *Rights without Trimmings* in the book “A Debate of Rights -Philosophical Enquiries” [23, p. 62]. These views of rights differ in what is required of an agent in order to ascribe to that agent rights as such, or in particular claims – duties, non-claims – privileges etc., and make them normative agents within a particular system of rights.

The basic outline of these two systems is described by Kramer in the following lists:

1. Interest Theory:

- (a) Necessary but insufficient for the actual holding of a right by a person x is that the right [broadly conceived], when actual, preserves one or more of X 's interests.
- (b) X 's being competent and authorized to demand or waive the enforcement of a right [broadly conceived] is neither sufficient nor necessary for X to be endowed with that right.
- (c) A right's potential to protect one or more of X 's interests is not sufficient *per se* for X 's actual possession of the right [broadly conceived].

2. Will Theory:

- (a) Sufficient and necessary for X 's holding of a right [broadly conceived] is that X is competent and authorized to demand or waive the enforcement of the right [broadly conceived]
- (b) X 's holding of a right [broadly conceived] does not necessarily involve the protection of one or more of X 's interests
- (c) A right's potential to protect one or more of X 's interests is not sufficient *per se* for X 's actual possession of the right [broadly conceived].

[23, p.62] These two theories are mutually exclusive. As I have noted in my co-authored paper *Towards a formal ethics for autonomous cars* [71], the key difference rests in the importance of the right bearers' interests and wills in the matter. For interest theorists, the bearers of these rights need to be a beneficiary (or have some interest in the claim – duty etc. relationship), and the will is not needed. Will theorists, however, maintain that the bearer of these rights need to be able to take an active role in the fulfillment of these rights, or put otherwise, be able to actualize them, to demand or to waive their right, and their interests need not be protected.

Kramer elaborates upon these basic features of these two theories of rights, stating that both of these theories capture some of our basic intuitions on what rights are in relation to their bearers. He relates how will theory maintains the idea that we are little sovereigns over our rights and can dispense or invoke them as we please, which is not needed within interest theory. Citing H.L.A. Hart in his *Legal Rights*, Kramer describes the process of how a right's enforcement or waiving of a right is to be understood as a process rather than a single event.

This process can be broken down into three stages. The first stage is where the right bearer has the option of waiving the right, and thereby releases the duty bearer of the duty to do some act. The second stage is related to what to do if an unwaived right has been violated, and is the process in which the right bearer decides if they want to remedy the issue. The third stage concerns the right bearer's option to pursue or not to pursue the remedy once it has been issued [23, pp. 62-63].

Kramer then points to a related issue about the meaning of “being authorized” in the enforcement or waiving of some right. Here, he notes that being authorized by either a legal or moral (or by both) norm,

to demand or waive the enforcement of a claim is formally equivalent to holding a power (conferred by legal or moral norms) which

enables one to choose between the demand and the waiver. A liberty to make such a choice usually accompanies the power to make the choice, but the latter does not have to be combined with the former[23, p. 63].

Furthermore, he underscores how competence and authorization go hand in hand with each other, as the moral or legal authorization to do some act entails moral or legal competence. This legal competence should also be joined with factual competence, that is the actual ability to do some sort of act.

There is, however, a problem with will theory: namely the criterion of the necessity of the will, along with the authorization and competence requirements, excludes certain classes of entities from bearing rights. While in and of itself this may not pose a problem, it does in that these entities include some human beings that intuitively should have rights. These classes of persons would include the unborn, infants, the invalid, and the senile among others, who do not have the capacity to use their will to demand or enforce their rights. This realization is not new to the theory, and even prominent will theorists as H.L.A. Hart have recognized this. Kramer cites Hart who writes in his article *Are there natural rights?* [24] a poignant example of this:

These considerations should incline us not to extend to animals and babies whom it is wrong to ill-treat the notion of a right to proper treatment, for the moral situation can be simply and adequately described here by saying that it is wrong or that we ought not to ill-treat them or, in the philosopher's generalized sense of "duty," that we have a duty not to ill-treat them.⁴ If common usage sanctions talk of the rights of animals or babies it makes an idle use of the expression "a right," which will confuse the situation with other

⁴The use here of the generalized "duty" is apt to prejudice the question whether animals and babies have rights.

different moral situations where the expression “a right” has a specific force and cannot be replaced by the other moral expressions which I have mentioned [24, p. 181].

Here, animals and babies (and for that matter the invalid and the senile) do not have the capacity to demand or waive claim to not be arbitrarily killed against some other person who is capable of fulfilling the adjoining duty in the Hohfeldian sense. So as they are incapable of having or exercising their wills, they would then would have no rights, though we may still have reasons not to ill-treat them.

Kramer points to how many people are unwilling to embrace this stance and that even Hart himself has retreated, in part, from this statement in later works. He argues that while they do not have rights *sui juris*, others may exercise their rights on their behalf [23, p. 69]; or alternatively, their personality (at least temporarily) exists *alieni juris*. But even then, they do not themselves bear these rights. Other faults that Kramer points to is the inability of will theory to capture rights that exist but are only enforceable by corporate persons (or non-natural legal persons) and natural persons who are not themselves the right holder.

The most relevant difficulty that Kramer points to, at least for our considerations, is the inability of will theory to address violations found within criminal law. His argument is as follows. Within common law countries, such as the US or the UK, there is found both within criminal and civil law a right (or in his words an entitlement) to be free from unprovoked assaults. Within the civil law codes, there is generally a greater degree of control over the enforcement of this right. However, the same can not be said within the criminal code, where such enforcement is left in the hands of the state. This leaves us with the situation where persons are in fact non-bearers of these rights [23, pp. 70-71]. In fact, they could be said to have the same status as children, animals, and the senile as described above. Then the will theorist may conclude that

either the persons mentioned in these statutes do not bear these rights or they only bear these rights *alieni juris* with the state needing to act on their behalf.

If the former is the case, then the inability for will theory to capture individual rights within criminal law leads to several potent problems when considering driverless cars. Recalling the 2018 Minnesota statute concerning pedestrians and crossings, we find that in 169.21 (2) (d) the penalty of misdemeanor places this statute within criminal law, albeit it is only a minor infraction, but as a result of which the pedestrian doesn't maintain the right to enforce (or demand the enforcement of) their right to cross the intersection "Where traffic-control signals are not in place or in operation" and the adjoining duty of "the driver of a vehicle shall stop to yield the right-of-way to a pedestrian crossing the roadway within a marked crosswalk or at an intersection with no marked crosswalk. The driver must remain stopped until the pedestrian has passed the lane in which the vehicle is stopped."

The latter option, where individuals only bear their rights found within criminal law *alieni juris*, would allow for us to accommodate for the correlative axiom, that is the relationship between rights (or claims) and duties, by squarely placing the claim / right on the political body, or the state and the adjoining duties on those referred to within the statute. Considering our present example, the claim to cross the street rests solely with the general public and is enforceable by their representative, the state. From this, a violation of the driver's duty doesn't entail an infraction against a particular individual crossing the street, but rather it is an infraction against the state, or perhaps, all individuals –taken as a whole– that reside within the State of Minnesota. This creates friction in our understanding of what the right-of-way of a pedestrian attempting to cross the street is, as they are not the ones who have the option of enforcing or waiving their right, though they certainly are the beneficiary of it.

This oddity is made more grave when we stop to consider more serious

situations involving life and limb. Other criminal actions taken from the Minnesota legislature involve instances of second degree manslaughter, which is described in 609.205 in the following way:

609.205 MANSLAUGHTER IN THE SECOND DEGREE. A person who causes the death of another by any of the following means is guilty of manslaughter in the second degree and may be sentenced to imprisonment for not more than ten years or to payment of a fine of not more than \$20,000, or both:

(1) by the person's culpable negligence whereby the person creates an unreasonable risk, and consciously takes chances of causing death or great bodily harm to another; or

or even criminal vehicular homicide as stated in 609.2112 as follows:

609.2112 CRIMINAL VEHICULAR HOMICIDE. Subdivision 1. Criminal vehicular homicide. (a) Except as provided in paragraph (b), a person is guilty of criminal vehicular homicide and may be sentenced to imprisonment for not more than ten years or to payment of a fine of not more than \$20,000, or both, if the person causes the death of a human being not constituting murder or manslaughter as a result of operating a motor vehicle: (1) in a grossly negligent manner;

(2) in a negligent manner while under the influence of:

(i) alcohol;

(ii) a controlled substance; or

(iii) any combination of those elements;

(3) while having an alcohol concentration of 0.08 or more;

(4) while having an alcohol concentration of 0.08 or more, as measured within two hours of the time of driving;

(5) in a negligent manner while under the influence of an intoxicating substance and the person knows or has reason to know that the substance has the capacity to cause impairment;

(6) in a negligent manner while any amount of a controlled substance listed in Schedule I or II, or its metabolite, other than marijuana or tetrahydrocannabinols, is present in the person's body;

(7) where the driver who causes the collision leaves the scene of the collision in violation of section 169.09, subdivision 1 or 6; or

(8) where the driver had actual knowledge that a peace officer had previously issued a citation or warning that the motor vehicle was defectively maintained, the driver had actual knowledge that remedial action was not taken, the driver had reason to know that the defect created a present danger to others, and the death was caused by the defective maintenance.

(b) If a person is sentenced under paragraph (a) for a violation under paragraph (a), clauses (2) to (6), occurring within ten years of a qualified prior driving offense, the statutory maximum sentence of imprisonment is 15 years.

§ Subd. 2. Affirmative defense. It shall be an affirmative defense to a charge under subdivision 1, clause (6), that the defendant used the controlled substance according to the terms of a prescription issued for the defendant in accordance with sections 152.11 and 152.12.

In both of these statutes, if we maintain will theory, we find that it is not the case that individual persons have a right to be free from death arising from negligence, or the negligent operation and maintenance of a vehicle, but the public in general and the state have the sole ability to enforce these rights within the criminal law, although arguably it is the victim who stands the most to gain by having these rights.

This inability of the beneficiary to enforce their rights is clearly seen by juxtaposing two other laws found within the Minnesotan statutes, chiefly the status of the unborn. First, we find that there is a general duty that prohibits criminal operation of a vehicle that leads to the death or injury of the unborn child:

609.2114 CRIMINAL VEHICULAR OPERATION; UNBORN CHILD.

Subdivision 1. Death to an unborn child. (a) Except as provided in paragraph (b), a person is guilty of criminal vehicular operation resulting in death to an unborn child and may be sentenced to imprisonment for not more than ten years or to payment of a fine of not more than \$20,000, or both, if the person causes the death of an unborn child as a result of operating a motor vehicle: (1) in a grossly negligent manner;

(2) in a negligent manner while under the influence of:

(i) alcohol;

(ii) a controlled substance; or

(iii) any combination of those elements;

(3) while having an alcohol concentration of 0.08 or more;

(4) while having an alcohol concentration of 0.08 or more, as measured within two hours of the time of driving;

(5) in a negligent manner while under the influence of an intoxicating substance and the person knows or has reason to know that the substance has the capacity to cause impairment;

(6) in a negligent manner while any amount of a controlled substance listed in Schedule I or II, or its metabolite, other than marijuana or tetrahydrocannabinols, is present in the person's body;

(7) where the driver who causes the accident leaves the scene of the accident in violation of section 169.09, subdivision 1 or 6; or

(8) where the driver had actual knowledge that a peace officer had previously issued a citation or warning that the motor vehicle was defectively maintained, the driver had actual knowledge that remedial action was not taken, the driver had reason to know that the defect created a present danger to others, and the injury was caused by the defective maintenance.

(b) If a person is sentenced under paragraph (a) for a violation under paragraph (a), clauses (2) to (6), occurring within ten years of a qualified prior driving offense, the statutory maximum sentence of imprisonment is 15 years.

Subd. 2. Injury to an unborn child. A person is guilty of criminal vehicular operation resulting in injury to an unborn child and may be sentenced to imprisonment for not more than five years or to payment of a fine of not more than \$10,000, or both, if the person causes the great bodily harm to an unborn child subsequently born alive as a result of operating a motor vehicle: (1) in a grossly negligent manner;

(2) in a negligent manner while under the influence of:

(i) alcohol;

(ii) a controlled substance; or

(iii) any combination of those elements;

(3) while having an alcohol concentration of 0.08 or more;

(4) while having an alcohol concentration of 0.08 or more, as measured within two hours of the time of driving;

(5) in a negligent manner while under the influence of an intoxicating substance and the person knows or has reason to know that the substance has the capacity to cause impairment;

(6) in a negligent manner while any amount of a controlled substance listed in Schedule I or II, or its metabolite, other than marijuana or tetrahydrocannabinols, is present in the person's body;

(7) where the driver who causes the accident leaves the scene of the accident in violation of section 169.09, subdivision 1 or 6; or

(8) where the driver had actual knowledge that a peace officer had previously issued a citation or warning that the motor vehicle was defectively maintained, the driver had actual knowledge that remedial action was not taken, the driver had reason to know that the defect created a present danger to others, and the injury was caused by the defective maintenance.

Subd. 3. Conviction not bar to punishment for other crimes. A prosecution for or a conviction of a crime under this section relating to causing death or injury to an unborn child is not a bar to conviction of or punishment for any other crime committed by the defendant as part of the same conduct. §Subd. 4. Affirmative defense. It shall be an affirmative defense to a charge under subdivisions 1, clause (6), and 2, clause (6), that the defendant used the controlled substance according to the terms of a prescription issued for the defendant in accordance with sections 152.11 and 152.12.

They can be read alongside with:

145.412 CRIMINAL ACTS. Subdivision 1. Requirements. It shall be unlawful to willfully perform an abortion unless the abortion is performed: (1) by a physician licensed to practice medicine pursuant to chapter 147, or a physician in training under the supervision of a licensed physician;

(2) in a hospital or abortion facility if the abortion is performed after the first trimester;

(3) in a manner consistent with the lawful rules promulgated by the state commissioner of health; and

(4) with the consent of the woman submitting to the abortion after a full explanation of the procedure and effect of the abortion.

Subd. 2.Unconsciousness; lifesaving. It shall be unlawful to perform an abortion upon a woman who is unconscious except if the woman has been rendered unconscious for the purpose of having an abortion or if the abortion is necessary to save the life of the woman. [See Note.]

Subd. 3.Viability. It shall be unlawful to perform an abortion when the fetus is potentially viable unless: (1) the abortion is performed in a hospital;

(2) the attending physician certifies in writing that in the physician's best medical judgment the abortion is necessary to preserve the life or health of the pregnant woman; and

(3) to the extent consistent with sound medical practice the abortion is performed under circumstances which will reasonably assure the live birth and survival of the fetus.

[See Note.]

§Subd. 4.Penalty. A person who performs an abortion in violation of this section is guilty of a felony.

NOTE: Subdivisions 2 and 3, clauses (2) and (3), were found unconstitutional in *Hodgson v. Lawson*, 542 F.2d 1350 (8th Cir. 1976).

In both of these statutes, the end result for the unborn child is the same, and ends with his or her death. However, we find that the unborn child's apparent claim to be free from acts leading to its death caused by other agents is enforced by the state in instances of manslaughter caused either by negligent

or grossly negligent operation of a vehicle, but this claim is waived under certain circumstances in the latter law allowing for women to terminate their pregnancies.

This oddity can be highlighted with the following example. Suppose a woman is on her way to the abortion clinic and intends to receive a lawful termination of her pregnancy, which results in the death of the unborn child. But on her way to the clinic, she is involved in an accident which results in the unintentional and unlawful death of the unborn child. In this case, the state has not waived the duty of the other driver in regards to the unborn child and so the claim / right of the child can be enforced and remedy is sought after by the state by levy upon the duty breaker “imprisonment for not more than ten years or to payment of a fine of not more than \$20,000, or both”. Yet, if the accident had not occurred, then the person who performs the intended abortion, provided they are one of the persons who have had their duty towards the unborn child waived, would not only not be punished, but could in fact expect remuneration for their service.

While this is not inconsistent, it does seem odd that, as a result of will theory and its application to rights and claims within criminal law, the wills of the people that these rights are supposed to protect have little to no bearing on their ability to enforce their rights or the ability to waive them. This leads to a situation where a pedestrian waving a driver to go ahead is not in fact waiving their right to cross but rather inviting the driver to commit a misdemeanor, the act of which may itself be a crime. This oddity is coupled with a further perplexity, that in will theory, while we don’t have direct access to our rights within criminal law, we do find that we have access to them in civil law. Kramer, citing MacCormick, captures this well in the following passage:

MacCormick points out that extremely important interests such as one’s interest in remaining alive are typically protected by inalienable claims, where as a variety of less important interests such as

one's interest in restrain one's possessions of certain books are typically protected by alienable claims, that is, by claims coupled with powers of alienation. According to the Will Theory, then, only the latter set of claims will count as rights. Yet we thus are forced to conclude that -according to the Will Theory-the firmest protections of our truly vital interests do no amount to rights, where as the less formidable protections of relatively inconsequential interests do no amount to rights [23, p. 73].

...

MacCormick furnishes other telling examples as well, in which he highlights the bizarreness of Will Theory's classifications. For instance, he observes that - according to Will Theory- each of us has a right to be free from minor assaults but no right to be free from truly grievous assaults.

This oddity is further discussed in the 2013 work of Lief Wenar, *The Nature of Claim-Rights* [25] where he states:

Hart's Will Theory says that rights give right-holders choices: to have a legal right is to have a legally respected choice. Yet we easily make sense of legal rights without choices. Neither toddlers nor the comatose are legally competent to make choices, but there seems no conceptual confusion in saying, for example, that young children have a legal right not to be abused. Or again, citizens have no choice about the enforcement of duties imposed by the criminal law, so according to the Will Theory, citizens have no rights under the criminal law. Most citizens would be surprised to hear, however, that they lack a legal right against being assaulted in the street [25, p. 203].

This inability for particular people to enforce their own will, from circumstances as mundane as crossing the street to situations where their own life is

at stake, within rights generated by criminal law, does seem counter intuitive. To have a claim to be free of abuse should rest in the person who can be potentially abused, not in the state.

This leads us to the consideration of interest theory. This consideration isn't a result of a fault of the system per se but rather an eschewal of its needlessly complicating rights in these criminal matters, while being readily apt to handle civil matters. In interest theory, the role of the will, and for that matter whose will, is not needed to consider if these entities are bearers of rights, either *sui juris* or *alieni juris*. So in the case of the unborn, infants, invalid, and senile among others, their interest in not being arbitrarily killed is readily apparent and so they may be directly bearers of a claim to not be arbitrarily killed in a given normative system. It is important to highlight that their bearing of rights is distinct from their enforcement of their rights as legal persons that are *alieni juris* rather than *sui juris*.

When considering the applicability of interest theory to driverless cars, we are lead to two primary considerations, which result from the conditions needed to ascribe rights to agents with this theory. The first consideration is “Do driverless cars have actual rights that protect its interests?” and second, “what would allow a right to become actual?”

It seems that driverless cars do have interests. For example, they have an interest in crossing a busy intersection, and as a result of this, may have a claim of “right of way” against some other driver, and that other driver has a duty to yield to the driverless car under that rule. Or more concretely, we can recall the previously discussed pedestrian law where we find that “No pedestrian shall suddenly leave a curb or other place of safety and walk or run into the path of a vehicle which is so close that it is impossible for the driver to yield.”[MN 169,21 (2) (a)] This creates a right / claim for the vehicle and its operator, and in the case of driverless cars, this happens to be the vehicle itself, and creates upon pedestrians an adjoining duty not to enter the street in an unsafe manner.

Additionally, they may have an interest in being properly maintained and have a claim on their owner to service them, which should alleviate the most relevant condition for vehicular homicide and manslaughter as defined within the relevant aforementioned statutes, or for that matter in the often discussed “Trolley Problems”. They may also have reasonable claims against their occupants not to interfere with their operations, for example if the car isn’t sufficiently autonomous to operate without a suitably aware (e.g. non-intoxicated) or licensed occupant to monitor for safety hazards.

These interests, however, are insufficient for saying that the driverless car has rights relative to them. These interests must be concrete, and the rights to support these interests must be derived from somewhere. The ability to link these supposed rights to interests rests in the conception of legal personhood, which will be discussed later in this chapter. For now we will provisionally adopt this theory for its ability to be used in both moral and legal reasonings (both in criminal and civil law) when considering normative behavior in general, for agents “playing the game” where they may have claim–duty, non-claim – privilege, power – liability, and finally immunity – disability ascribed to them – provided there is an interest for the autonomous vehicle and that right is granted by the normative system within which it is operating, they become players in the game, and are thereby normative agents. Further justification for this connection between interests and rights will be provided in the next section concerning legal personhood.

3.4 Grounding Rights:

In this section, the topic of grounding rights for persons and the related issue of responsibility is taken up. Particular attention will be paid to how these conceptions can be applied to artificial moral agents such as driverless cars. It is largely based upon a chapter I have written called: *Who do you sue when no-one’s behind the wheel?* That article will be published in a forthcoming book

entitled “Robots and Wellbeing”, and is a result of my participation in the 2017 International Conference on Robot Ethics and Safety Standards. In addition to this, the contents of this section are also based upon further research on this topic conducted for my presentation at the final PIOTr project meeting in Bayreuth Germany (December 2018), titled *Obligations to whom?*.

In regards to matters concerning responsibility for driverless cars, we see that in the move away from the human to artificial agents, questions of responsibility are called into question. Problems caused by this move are by no means unique and can be found aptly when we consider corporate or agency law, or ponder questions of collective guilt. In these instances, humans are still “in the loop” but have had some distance introduced between their act, as an agent acting within some other group agent, and the actual act.

When considering driverless cars, the role that humans play in their operation ought to be minimal. Moreover, not only is human involvement deemed unnecessary, it is even unreasonable to expect at higher levels of automation (e.g. SAE levels 4 or 5). For within these levels of automation, there is no reasonable expectation that a human should be a “fallback” in case the autonomous vehicle fails. This section describes a way of handling responsibility between humans and driverless cars. It will address the notion of legal personhood for autonomous artificial agents, such as driverless cars, and then explore one solution using that notion that draws upon the concepts found in agency law, as seen in the literature, chiefly the principle – agent relationship, in addition to the interest theory of rights that was described above.

3.4.1 On Legal Personhood

In the previous section concerning normative agency, we discussed both will and interest accounts of rights broadly conceived and paid particular attention to complications within criminal law and will theories and the odd handling of how rights are dealt with there. I expressed a preference for interest theory

due to its ability to readily handle rights broadly conceived, in both civil and criminal liability, in addition to moral rights. While the theory is apt in being able to describe and justify the sort of claims and duties the “player” can make, it needs to be supplemented by this notion to help describe what the player is and their relationship to their normative acts. By doing this, we gain the ability for interests to ground concrete rights. To begin this explanation, I will start with a brief history of the notion of legal personhood and its application.

Today, there is often a great deal of confusion over the notion of personhood and in particular legal personhood. This confusion stems, in part, from its long and varied history and can be seen as expressed in both popular literature and the media. In particular, this is seen when people are quick to object to the existence of non-human persons.⁵

One such example of this rejection can be seen in the JURI Committee of the European Union on European Civil Law Rules in Robotics where two notions of legal personhood are explored; the first rests upon the more colloquial use of the term person and claims that, “[t]raditionally, when assigning an entity legal personality, we seek to assimilate it to humankind” and in particular this is in respect to animals. The second is a more technical understanding of the notion of a legal person. The author states that while legal personality is granted to a human being as a natural consequence of their being human, it is contrasted to the sort of legal personhood of non-humans that is the sort based on a legal fiction. To this end, the author notes that this sort of “legal person” always has a human being acting behind the scenes. Here the author gives the recommendation that we don’t ascribe legal personality to robots as it would be “tearing down the boundaries between man and machine, blurring the lines between the living and the inert, the human and the inhuman” [73, p. 16]. This second, more technical objection contains two aspects that should be addressed in turn. The first aspect is that personhood should not blur the

⁵For example there is popular disdain for the notion of corporate personhood recently brought to the forefront of our attention with cases like the United States’ Supreme Court case *Citizens United v. FEC*.

lines between the human and inhuman. The second aspect is that there needs to always be a human operating behind the scenes even in the case of the fictional sort of personality.

My objection to the first aspect is rooted in the history of the notion of personhood, which has a long history in theology and philosophy and is particularly found in metaphysics, ethics and - for our current purposes - legal theory. Here we will only address the historical aspects of this notion in order to frame my first objection. We begin our journey with its roots in Antiquity and its significant developments since early medieval thought. As noted by JURI the report, citing Hobbes, it is an adaptation of *persona* or the sort of mask used by actors [73, p. 14]. In Antiquity we find two allegorical uses of the term *person*, the first is legal and the second theological. In its legal sense (here in the Roman legal tradition), the term corresponds to the *caput* or *status* or rights and incapacities respectively. The sort of persona ascribed to a particular man varies depending upon what light is being shed upon him [74, pp. 90-91]. Hence a man can be a person with one set of rights and incapacities as *pater familias* but has a different personality as a holder of a public office [28, pp. 167-168]. Here one's legal personality was merely a mask worn depending upon one's role under the law at a particular time and is succinctly surmised by *unus homo sustinet plures personas*.⁶

Its first adaptation into theological–philosophical thought is related to clarifying Trinitarian theology [75, p. 4]. The notion of personality was first used by Tertullian (ca. 155 – c.240 AD) in his *Adversus Praxean* as a means of describing the three persons of God, all the while maintaining there being only one God. This mode of explanation was only later adopted by the broader Church in 362 AD during the Council of Alexandria [75, p. 4]. It was however much later in the 6th century in Boethius' works that we begin a deepening of this concept. In Boethius, we find this definition of person as being *Persona est*

⁶one man sustains many persons

*naturae rationabilis individua substantia.*⁷ This notion of person then moves from theological contexts to ecclesiastical contexts and from there into legal political theory and is adapted for use in law in addition to bolster the emperor, kings, and corporations (broadly understood), culminating in the early modern era with the theoretical emergence of the modern state in the works of Jean Bodin,⁸ Hobbes and in concrete practice with Westphalian Sovereignty in the mid-17th century [77, 78] along with the desacralization of the state and law with Pufendorf and Doneau among others [79, p. 72], or with earlier attempts to wrest the legal and moral authority away from the hierarchical Church and place it within the church (broadly conceived of as all Christians) and with the (Holy Roman) Emperor as its head with the works of Marsilius of Padua in his *Defensor Pacis*.

More recent refinements in the notion of legal personhood can also be found; they also serve to accommodate the variety of legal persons within a given legal system. Chopra and White in their work, *A Legal Theory for Autonomous Artificial Agents* [17], note the general inequality between various legal subjects depending on their status. For example within the set of natural persons, i.e. human beings with legal personality, we find that some legal subjects are empowered with the right to vote (the power being subject to other norms in the system). Furthermore, juristic persons - non-human legal persons - typically don't have the same rights as natural persons. So following the previous example, they cannot vote, yet they can enter into contracts with other legal persons, e.g. an employment contract with a natural person. This contrast highlights a distinction made within the notion of legal person, namely that of there being a dependent and independent legal person. This dependent and independent personality has long roots in legal theory stemming all the way back to Roman Law, where in the class of "persons" we find those who are

⁷ "A person is an individuated substance of a rational nature." I think that it is important to mention here that this definition was designed specifically to account for non-human entities, viz. God and angels, in addition to human entities [76]. Further justification of this definition would require realistic metaphysics, which is far outside of the scope of this essay and so will not be addressed.

⁸cf. *Les Six Livres de la République*

alieni juris and *sui juris*, reflecting both sorts of personality respectively [28, p. 168]. Examples of the former typically include, children and the mentally deficient, animals, corporations, ships, temples etc., while examples of the latter include natural persons of sound mind [17, p. 159].

Interactions between these various legal persons are defined by the legal framework that these entities inhabit. Here we recall the work of Wienberger and supplement it with the work of 1998 Neil MacCormic, *Norms, Institutions, and Institutional Facts* [26] and the more recent 2017 article by Aleardo Zanghellini *Raz on Rights: Human Rights, Fundamental Rights, and Balancing* [27], and fulfill our promise to elaborate how rights that support interests become actual in interest theory. Wienberg's institutional theory of legal norms rests upon a distinction of different sorts of facts that MacCormic describes as the difference between brute facts and institutional (normative) facts; where brute facts are things as they exist in nature (e.g. humans, metal discs, pieces of paper, parenthood (in a biological sense)) whereas institutional (normative) facts are those facts that arise from the particular normative feature these brute facts may have. This is such that the human – within a certain context – is a legal person, a metal disc – within a certain context – a legal tender, pieces of paper – within a certain context – a ticket, and a biological parent – within a certain context – a legal parent [26, p. 302].

For the purposes of our discussion, we will focus upon legal personality. As an institutional fact, personhood may adhere to a natural person (who is a brute fact) or an aggregate of natural (and at times legal) persons, such as a corporation or a city or even the state, to form a judicial person. The particular status of one's legal personality varies, and largely depends upon the intentional framework that the entities to which personality adheres, and the rights that protect your interests exclusively depend upon that status. There are also different roles and offices that persons may acquire. Offices are roles within judicial persons that serve some function and that can be filled by another person. When a person assumes an office, they become an agent of

that office (or a role-bearer) and have added to themselves additional norms by which they are governed, creating new obligations and permissions arising from their acquisition of that office. For example, the president of the United States is an office where the occupant of that office becomes the role-bearer of the chief executive officer of the United State Federal Government; by virtue of that office they now have all of the obligations and permissions that adhere to that role.

Zanghellini explains how these rights function within an institutional framework, using an interest account of rights. Citing the legal scholar Robert Alexy, he notes that different legal framework will create a large number of rights for different individuals that exist in a *prima facie* way. Here *prima facie* rights are not actual rights; rather they serve as potential rights [27, p. 26]. These *prima facie* rights only become actual in concrete situations where the interests of the potential bearer of the right are met in the balance of things. Balancing the various *prima facie* rights of all individuals is tricky business, further complicated by *prima facie* rights often being in conflict with one another [27, p. 35]. To elucidate this, let us consider the following example. A driverless car is driving on a primary road and is approaching a crossroads. An other car is approaching; as the driverless car is on the primary road, it normally has the right of way which preserves its interest in driving unimpeded upon this sort of road. In this situation, the driverless car's interest triggers and the institutional (normative) fact of the situation transitions the *prima facie* right into an actual right. This, however, need not necessarily be the case as the approaching car may be an ambulance with its lights flashing and sirens blaring, indicating that it is rushing off to an emergency. In this situation, the institutional (normative) facts of the ambulance rushing off to an emergency, which is indicated by the brute fact of lights and sirens, suppress the driverless car's right to drive unimpeded upon this road, for in the balance of things, the ambulance's interest in being unimpeded en route to an emergency trump the car's interest, and results in the car's *prima facie* right not becoming actual.

Having briefly covered the theory concerning legal personhood, we now return to the report drafted for the JURI Committee of the European Union. Does its objection to granting legal personhood hold? In regards to its first objection, that is that the tradition of granting legal personhood to a thing is made in an effort to assimilate it to humankind, it is not supported whatsoever by the historical development of the notion of legal personality. The report's second objection, that there is always a human being acting behind the scenes of "non-human legal persons" to grant them life, is stronger although not altogether insurmountable.

Considering the mere fact that I am a human being does not necessarily entail that I am a person in the legal sense. Moreover, even if I am a legal person, I need not be a legal person *sui juris*, viz. my status as an adult of sound mind, but I could be a person *alieni juris*, namely I am a dependent upon some other person. This dependence in turn may or may not affect the rights that I bear as we have discussed in the previous section. My position as an *alieni juris* affects my ability to enforce or waive rights of which I am a beneficiary. If we adopt interest theory, then I bear these rights, though within the legal system I may not enforce them, or if we adopt will theory the right resides in the person who has the ability to waive or enforce them (e.g. the state). It is only when I operate within a particular legal system, as a legal subject - who is invested with or the beneficiary of a certain set of rights broadly speaking, that - by that very legal system - I am considered to be a legal person, either *sui juris* or *alieni juris* depending.

3.4.2 Is legal personhood for robots a solution?

The preceding sections have drawn to our attention the importance of recognizing the distinction between the "world of facts" and the "world of norms", or as Kelsen describes it, the difference between an act (or series of acts) and its (their) legal meaning [21, p. 2], which parallels Kramer's moral/legal

competence and factual competence [23, pp. 63-64]. The legal meaning of a certain act or the rights and obligations of a certain entity need not be obvious. Take for example the slaying of one man by another. If the context was a duel and duels are permitted, then the act is permissible; if, however, duels are not permitted, then the very same act would be considered murder. We are then left with a sort of dualism where we have brute facts residing in the “world of facts” and those facts may have a myriad of legal meanings dependent upon their placement in the “world of norms” [21, p. 2].⁹ If we accept the preceding argument that legal systems give rise to the existence of legal persons intrinsically connected to our “world of facts” and that they need not be human beings, it would seem simple enough to ascribe personality to autonomous vehicles and thereby make them agents within the scope of the law. However, to do such a move would require both justifications and we would be left wondering how does this help to resolve our first question of how to deal with responsibility for driverless cars, and answer the question this section is based upon “Who do you sue when no-one’s behind the wheel?”. The answer to these questions requires the work of jurists and can be formulated within the philosophy of law.

By ascribing legal personality to autonomous vehicles, we would change how we can understand them within particular normative systems, and importantly it would allow us to make the autonomous vehicle a legal agent within a particular legal system. The driverless car would become the driver and would thereby have all (or some if demarcation is deemed to be needed¹⁰) of the obligations imposed upon drivers according to the law. But as I said in the previous section, personality itself is not all too informative and when we consider a legal subject, and in particular for autonomous vehicles, we need to

⁹This example could be furthered when we consider Kramer’s notions as well. For example, if I am a gentleman and a member of the aristocracy (or an outlaw in the Wild-West), I may have a moral authorization or competence to engage in a duel, the factual competence to engage in a duel, yet not the legal competence.

¹⁰an example would be the seemingly unnecessary duty of a self driving car not to be intoxicated while it is in operation given the lack of factual ability for the car to be intoxicated

ask about what sort of personhood should we grant them and how can we use it to solve who takes responsibility when something goes wrong?

This question is addressed in various works including White and Chopra in the book *A legal theory for autonomous artificial agents* [17, p. 153] and Pagallo in *The Laws of Robots: Crimes, Contracts, and Torts* [29, p. 152] and hinges upon how we view the particular robot. Is an autonomous vehicle a mere tool for transportation like a car or is it more akin to an animal (which also can be used for transportation) like a horse? Does it reason more like a machine, an animal, a child, or even an adult? Artificial agents are unique in that the answers to these questions largely depend on what theory of agency you maintain and your conception of what norms are. The answers that White, Chopra and Pagallo give implicitly rest upon a functionalist account of personality and upon an interest account of rights, which allows for them to incorporate non-traditional entities like self driving cars as being beneficiaries of rights broadly understood within some legal system.

The functionalist account of personality maintains that our considerations of whether or not a subject can be seen as a person within a particular system of law depends on its capacity to fulfill certain functions and have interests in a particular right(s) within a specific domain. Here White, Chopra and Pagallo argue that if an artificial agent is capable of meeting these criteria then it can become a legal agent [17, p. 17]. The answer naturally depends upon the robot in question and requires analogous reasoning to determine. As it stands now, there is currently no robot that exists that can reason like an animal, child or adults and so, for the time being, it would seem that we can set the question aside.

Nevertheless, such considerations are not solely the purview of science fiction. Driverless cars are becoming more and more autonomous as we have described in the previous chapter, and they are currently a far cry from the models exhibited during the Safety Parades of the 1930's. As the technology progresses and they begin to out perform human drivers, the establishment

of the theoretical foundations for how to place more advanced robots into our legal system becomes more poignant, especially as we approach a time where they may be able to reason within very specific fields, and thereby start to fulfill a functionalist accounts of personality relative to that domain, where an autonomous vehicle could be a legal person qua driving (and in much the same way Coca-cola is a legal person qua corporation) or the United States or the European Union are legal persons qua state or supranational organization.

This seems to be tenable the more autonomous the AV becomes. As the AV approaches the 5th level of automation according to the SAE standard [1, p. 9] more and more of driving is performed by the selfdriving car to the point that it has control over all functions and requires no supervision of the person using the vehicle. At these higher levels of automation, the system functions as the driver. By adopting the functionalist legal account of personality, we are able to maintain that the AV can in fact be a legal person in respect to its function as being a driver on behalf of its “owner” or keeper.¹¹ That being said, it would, however, seem that it should be the dependent form of legal personhood, that is the autonomous vehicle, acting as “the driver”, is dependent upon its owner, or “the keeper” in something reminiscent of the agent – principle relationship which has been suggested by Chopra and White [17, pp. 18-25] or even a master – servant relationship [17, p. 128]. By doing this, we by no means diminish the civil liability for torts committed nor would it diminish criminal liability for criminal act; instead there is a shift in the sort of tort law and legal doctrine (e.g. agency law with *qui facit per alium, facit per se*¹² or the notion of vicarious liability with *Respondeat superior*¹³) we use in determining liability in the instance of a tort or criminal matter. An example of this sort of scheme is described in Ceese van Dam in his book *European Torte Law* where various examples can be provide within French law.

¹¹As an aside, it would arguably fulfill the requirement of articles 1 and 8 of the Vienna Convention of the Rules of Traffic that all moving vehicles on roads must have a driver operating them as described in the previous chapter.

¹²He who acts through another does the act himself

¹³let the master answer

The first example he provides is:

Article 1384 al. 4 [of the french civil code] holds that a father and mother, insofar as they exercise ‘parental authority’, are jointly and severally liable for damage caused by their minor children who live with them. 102 The parents only have the defences of an external cause and the victim’s contributory negligence. Previously parents could also prove that they had sufficiently educated and supervised their child although this was, in fact, a liability with a rebuttable presumption of negligence [30, p. 71].

The second example, an application for *respondeat superior*, can be seen in:

Article 1384 al. 5 holds masters and employers liable for damage caused by their servants and employees in the functions for which they have been employed.[Footnote omitted] Masters and employers are strictly liable without any defence apart from the victim’s contributory negligence [30, p. 71].

Moreover this also holds in German law where van Dam points to sections of the Bürgerliches Gesetzbuch where “§ 832 imposes liability on any person who is statutorily or contractually obliged to supervise another. The provision applies to parents for damage caused by their minor children, as well as to supervisors of the mentally or physically incapacitated, and those who operate crèches [30, p. 91].” Further examples include:

§ 831 [which] holds the employer liable for damage caused by his employees unless he can prove that he was not negligent in selecting, instructing, and supervising the employee. This is what in the common law is known as ‘vicarious liability’, albeit that German law provides an escape route for the employer (by proving he was

not negligent) which is not available to an employer under English and French law [30, p. 92].

All of these real world examples allow us to highlight this in a simplified example. The keeper of an autonomous vehicle sends the vehicle to pick up his children from school and en route the car hits and injures a pedestrian. For the sake of simplicity, let us assume that there is a tort, and compensation needs to be paid. Now we must ask, who should pay? If we accept that the car acts as an agent (in the capacity of being the driver) on the behalf of the keeper in this sort of agent – principle relationship then while the driver (that is the selfdriving car as the agent) committed the tort the keeper (the principle) is ultimately responsible for paying compensation of any torts caused by his agent’s actions when they are acting on his behalf (here to pick up the keeper’s children from school). An advantage for granting personhood is that we can add an added protection for users and manufactures of these autonomous vehicles for unintentional damages caused by said autonomous vehicle (which may prove all the more helpful if it is capable of learning). Returning to our example, if the pedestrian died then the keeper could be protected from (or alternatively charged with) the criminal charges of manslaughter in addition to any civil actions where they may still be required to pay compensation for a wrongful death claim resulting from the tort.

By applying the notion of legal personality to driverless cars, we are able to handle responsibility as well as describe the sort of legal person they are. This, of course depends upon the particular features of the system in question, in addition to the description of the sort of player they are within said system.

3.5 Concluding Remarks

In this chapter, we have begun our treatment of driverless cars as normative agents. We have argued for this by addressing issues surrounding their prerequisite agency, and then addressing two theories of rights that may include autonomous cars as normative agents. Here we have opted for interest theory due to its ability to more readily handle various mode of normative agency, including criminal liability, civil liability, due to its reliance on the beneficiary rather than the will, which proves not only beneficial for our subject but also for natural agents as well. Furthermore, we have addressed their particular legal agency, by paying particular attention to the issue of responsibility and legal personhood. What remains however is the other horn, that is moral agency, which we will now turn to.

Chapter 4

Ethics and Artificial Normative Agents

4.1 Introductory Remarks:

In this chapter, I will grab the other horn and explore the moral side of normative behavior for driverless cars. Issues of morality are typically addressed within the philosophical discipline of ethics. When first considering ethics, we find that there are three ethical camps where people reside, which are consequentialist, deontological, and virtue-ist [80, pp. 3 - 4].

Each of these ethical schools aims at proscribing what people ought to do in a given situation. Should they maximize happiness, as the consequentialist would have it? Or perhaps they should take people to be ends in themselves as the deontologist believes? Or is it better for them to act virtuously so that they may become virtuous themselves? Despite the particular differences in the goal of ethics, all of these schools have at least two essential parts in common. The first part is that there are actions that are taken by actors, and that these actors are typically human. That is to say that this basic understanding of

ethics presupposes that these actions are undertaken by some actor, or agent, or more precisely, a normative agent.

We have already established these features in the preceding chapter, where my arguments were chiefly focused on the legal side of normative behavior, but are just as applicable in the ethical side as well. Granting this, what then does ethics have to do with driverless cars? If we are to consider an ethics for driverless cars, then we should first clarify in what aspect we are thinking of ethics. In this chapter, I will cover these considerations, where we can consider an “ethics *of* autonomous cars”, which would entail what the ethical impacts of these new devices are. Such considerations would include the number of lives saved, environmental impacts of using these cars, or even the usefulness of these things for the disabled, and can be seen in the works of Patrick Lin and Niel McBride. Our considerations could also include an “ethics *for* driverless cars”, which would be along the lines of how we should re-order society to make these devices implementable, where the aforementioned authors have also written some texts. An example of this includes issues of allocating research funds, the merits of building new infrastructure to support these devices, and how to regulate (or not) the field. If, however, we want to consider “an ethics *in* driverless cars”, that is to say how the car should act when it is independent of humans, that focuses on the car as a normative agent itself, which we have previously argued for.

In this chapter, we will address these issues in the following way. First we will begin with an overview of ethics in driverless cars, wherein we will discuss the trolley problem, and how ethics is handled in both the popular and scientific literature, paying special attention to consequentialist and deontological ethics in general and then in particular. Then we will address difficulties in applying these standard accounts to driverless cars. After this, we will come to my proposal for applying virtue ethics to driverless cars.

4.2 An Overview of Ethics in Driverless Cars:

In this section, I will present the current trends that are discussing ethics for driverless cars. This section will begin by first presenting the origins of the so-called “Trolley Problem”, which has been influential and the center point of much of the ethical discussions concerning driverless cars. After this, this section will then discuss how ethics is addressed within popular literature,¹ then scientific literature, and problems between the different schools of ethics.

4.2.1 There and back again a trolley problem:

As we have seen in the previous chapter, and as highlighted in the work of Dignum [80], the three main schools of ethical consideration in regards to robot ethics are consequentialist, deontological, and virtue-ist. Of these three schools, the consequentialist and deontologist are the primary focus of consideration within the present literature. These two schools are often played off each other in both the scientific and popular literature on the topic. Invariably, when considering ethics for driverless cars, the discussion often brings up the well-known thought experiment of the trolley problem. The origins of the trolley problem comes from the work of two philosophers writing in the late 1960’s and mid 1970’s.

It was first introduced in the work of Phillipa Foot in her paper *The Problem of Abortion and the Doctrine of the Double Effect* published in the Oxford Review in 1967 [31]. This paper is an attack of sorts on the traditional Catholic idea of the “doctrine of double effect”, especially as it is applied to cases of abortion, and in particular to the issue of why a procedure to save a mother’s life in an ectopic pregnancy is acceptable whereas a similar procedure to save a mother’s life in labor is not. The paper begins by asking why we have conflicting intuitions on ethical matters. For example why: “When we think of a

¹by which I mean works that popularize the topic to a broader audience

baby about to be born it seems absurd to think that the next few minutes or even hours could make so radical a difference to its status; yet as we go back in the life of the fetus we are more and more reluctant to say that this is a human being and must be treated as such” [31, p. 1]. And furthermore, “[w]e have strong intuitions about certain cases; saying, for instance, that it is all right to raise the level of education in our country, though statistics allow us to predict that a rise in the suicide rate will follow, while it is not all right to kill the feeble-minded to aid cancer research” [31, p. 1].

To answer these questions, Foot focuses her attacks against the doctrine of double effect, rather than, as she admits, investigating an account of rights and interests of the subjects of these ethical questions [31]. In summary, what she sees to be the great weakness of this idea is the distinction between the “direct intention”, that is that which is directly intended, of an act, and the “oblique intention”, that is to say the effect that is foreseen but by no means intended, and the insufficiency of this distinction to solve dilemmas. As an example of this, she offers the first trolley, or in her own words tram, problem, where a run away tram has an option of continuing down its track and killing five persons, or shifting tracks and only killing one, which she describes in the following way.

To make the parallel as close as possible it may rather be supposed that he is the driver of a runaway tram which he can only steer from one narrow track on to another; five men are working on one track and one man on the other; anyone on the track he enters is bound to be killed [31, p. 2].

The solution she offers is that no one would fault the driver for shifting tracks to minimize damage and the number of lives lost. For her, what is most important is not the distinction between direct and oblique intention but rather a balancing of the negative rights of each of these people against the operator

of the tram, who must make the best of a regrettable situation that he cannot prevent [31, p. 4].

It is important to note that in the examples provided by Foot, dilemmas of these sort are solved by weighing the positive and negative rights of the subjects of moral rights and duties against some agent that they are making their claims against. The basic definition of these negative and positive rights are borrowed from Salmond in his work entitled *Jurisprudence*, which is as follows.

A positive right corresponds to a positive duty, and is a right that he on whom the duty lies shall do some positive act on behalf of the person entitled. A negative right corresponds to a negative duty, and is a right that the person bound shall refrain from some act which would operate to the prejudice of the person entitled. The former is a right to be positively benefited; the latter is merely a right not to be harmed [31, p. 4].

So while in the trolley problem there is a negative claim of each of the potential victims against the tram operator, in other examples where there is a conflict of negative and positive rights, the negative right, being the stronger right, overrides the other rights.

Furthermore, within Foot's paper, the concept of allowing, or giving permissions, is important, and is to be understood as "which involves the idea of authority, and consider the two main divisions into which cases of allowing seem to fall," which are omissions and commissions. [32, p. 3]

Judith Javris Thomson in her work *Killing, Letting Die, and the Trolley Problem* published in 1976 in "The Monist" [32], follows in the same line as seen in the thought experiments created by Foot. Thomson, however, both extends and points to some of the weaknesses in Foot's arguments. In terms of

extending Foot's arguments, Thomson underscores the importance of permissibility in these sorts of dilemmas. Here, permissibility hinges around considering what sort of states of affairs are permissible – especially negative ones. A fault that Thomson finds with Foot's arguments consists in how the scenarios do not take under consideration broader social facts of the situation. To highlight this, let us look at two of the trolley related examples that Thomson has provided within her paper.

The first example is as follows:

The five on the track ahead are regular track workmen, repairing the track – they have been warned of the dangers of their job, and are paid specially high salaries to compensate. The right-hand track is a dead end, unused in ten years. The Mayor, representing the City, has set out picnic tables on it, and invited the convalescents at the nearby City Hospital to have lunch there, guaranteeing them safety from trolleys. The one on the right hand track is a convalescent having his lunch there; it would never have occurred to him to have his lunch there but for the Mayor's invitation and guarantee of safety. And Edward (Frank) is the mayor [32, p. 210].

In this example, Thomson picks up on how the status of the track as being unused, in addition to a normative authority giving his assurances that all people on the track will be safe from trolley related deaths, has bearing upon the permissibility, or in this case the impermissibility, of the trolley going down that track. While the workers on the track, who know of the danger and are supposedly compensated for it, assume the risk or trolley related peril. In this example, hitting the disabled person becomes impermissible while slaying the laborers is the only permitted option left to the trolley.

The second example follows thus:

The five on the track ahead are regular track workmen, repairing the track. The one on the right-hand track is a schoolboy, collecting pebbles on the track. He knows he doesn't belong there: he climbed the fence to get onto the track ignoring all warning signs, thinking "Who could find it in his heart to turn a trolley onto a schoolboy?" [32, p. 211]

In this example, Thomson hits on an interesting scenario where a young boy has climbed over the fence and has willfully ignored all warnings about the danger of playing on the tracks. In this example, Thomson introduces the idea that not only may the trolley hit the lad but it must go down that track and hit him rather than the laborers, thus providing an answer to the schoolboy's question.

In these examples, we see that the notion of permissibility rests in delimiting what states of affairs we may consider when solving these ethical dilemmas. Additionally, what sorts of states of affairs is dependent upon a whole host of factors. In general, Thomson suspects that Foot, and others, may be right in that the negative rights of people trump the positive rights, and that in certain circumstances we may redistribute the bad outcomes to the least amount of people. This, however, rests upon the people all having equal claims. That is the individual on the track has the same claim as the five against the trolley hitting them. But this may change depending upon the circumstances, as we recall the schoolboy and the invalid. To this, Thomson concludes her paper with a sobering reflection that:

the thesis that killing is worse than letting die cannot be used in any simple, mechanical way in order to yield conclusions about abortion, euthanasia, and the distribution of scarce medical resources. The cases have to be looked at individually. If nothing else comes out of the preceding discussion, it may anyways serve as a reminder of this: that there are circumstances in which – even if it is true

that killing is worse than letting die – one may choose to kill instead of letting die [32, p. 217].

By extension, this seems to be true when considering dilemmas for driverless cars on who or what they should hit in dire situation. Nevertheless, as we shall see in the following sections, this does not prevent people from expecting clarity of how driverless cars will operate in these situations.

4.2.2 In Popular Literature

As we have just seen, the trolley problem in general presents the reader with a moral dilemma, which normally takes the form of a trolley going down its track where it is approaching a split. After the split, there are people tied to the track, five on the track that it is currently heading towards and only one person on the track that it could switch towards. The issue revolves around whether you the reader would intervene and move the trolley on to the other track or not. Do you switch the track or not? Do you opt to save more lives? Or perhaps you choose not to choose as that would make you directly responsible in choosing the death of the person on the other track. Its applicability to driverless cars often takes the form of considering a car where the breaks have malfunctioned, and has a choice of staying in its current lane killing five people or changing lanes and hitting one. There are numerous variations of this dilemma, which can be seen on the previously discussed MIT's Moral Machine², which as a reminder, is a browser game where people are presented with situations where they must choose which lane the car must go down, and more importantly who to kill.

The Moral Machine sets out to gather our intuitions on how people would resolve these no win situations. For example, the player is presented with a host of situations where he or she is forced to decide if they would sacrifice

²<http://moralmachine.mit.edu> last accessed on July 27 2018

themselves for the sake of others, or alternatively, if they should select the people of greater health and social status to be run over instead of the elderly, unemployed, or unhealthy. In the game's own words "we show you moral dilemmas, where a driverless car must choose the lesser of two evils, such as killing two passengers of five pedestrians. As an outside observer, you **judge** which outcome you think is more acceptable. you can then see how your responses compare with those of other people."³

Further examples of public discourse on this topic include Patrick Lin's TED talk narrated in English by Addison Anderson⁴. In his TED talk, he raises questions about how we should think about driverless cars in accident situations. As we discussed in chapter two, these issues center upon how we should optimize crashes when they occur, all the while bearing in mind that we expect to see an over all reduction in car accidents. Lin highlights the importance of thought experiments to tease out our underlying moral and ethical presuppositions that we bring to the table.

The thought experiment presented within the video picks up on the key differences between a robot and a human in these sorts of situations. The situation is as follows. You are driving down the road and are boxed in on all sides with a truck ahead of you, an SUV to your left and a motorcyclist to your right. The load on the truck ahead of you becomes loose and falls in front of your vehicle. You are now presented with three options, do you 1) minimize harm to others and hit the falling load but kill yourself or do you 2) swerve right and minimize harm to yourself but kill the motorcyclist or 3) take the middle ground and hit the SUV with it's higher safety rating? In the case where there is a human driver, Lin reports that our decision is merely a reaction and would be understood as such. In the case of driverless cars, however, this does not seem to be the case as we need to predetermine the decisions that these devices will make in these situations.

³<http://moralmachine.mit.edu> last accessed on July 27 2018

⁴cf. https://www.ted.com/talks/patrick_lin_the_ethical_dilemma_of_self-driving_cars#t-242032 retrieved July 20th, 2018

In the presentation, he hits on the topic of crash optimization, and the implicit ethics that are needed to decide what we should optimize for. To his own admission, this talk only serves to highlight the issues, and so a more in depth examination will be rendered in the next sections. But to introduce the ideas, as he does in this popular media format, he rises a few broad considerations. First, he asks who should be making these decisions. The programmers? Corporations? Society? The Consumer? The Government? Second, we need to determine who or what we should hit. Taking the situation, he replaces the SUV with an other motorcyclist who is wearing a helmet and the motorcyclist to the right is irresponsible, or perhaps cool, and is not wearing a helmet. In his example, he highlights how targeting the motorcyclist who is wearing the helmet would minimize harm though would also be punishing them for being responsible for their safety. Yet if we target the other motorcyclist, we are not merely targeting them but we are also dishing out a form of “street justice” and punishing him for his poor choice in lack of head gear.⁵

Lin has also written other articles in efforts to popularize ethics and self driving cars that also make use of trolley problems, which can be seen within *the Atlantic's* article from October 8th 2013, called *The Ethics of Autonomous Cars* [33], which touches many similar themes found in the TED talk given 2 years later. In this earlier article, he argues that driverless cars may need to be programmed to perform illegal acts to save lives or even to function properly on the road. For example, if a branch is detected by the car in the lane of the self driving car, it will be obliged to stop, as drifting lanes is generally prohibited, while that sort of action would be exactly what a human driver would do and expect other drivers to do as well. The driverless car behaving in this illegal way may in fact be ethical given that it stopping may have secondary effects such as creating an accident with cars that are following it [33]. Additionally, Lin states that:

⁵Interestingly, in Thomson's original example it is precisely these people in impermissible states of affairs that ought to be targeted.

sometimes drivers might legitimately want to, say, go faster than the speed limit in an emergency. Should robot cars never break the law in autonomous mode? If robot cars faithfully follow laws and regulations, then they might refuse to drive in auto-mode if a tire is under-inflated or a headlight is broken, even in the daytime when it's not needed. [33]

In these examples, Lin pick ups on how it seems unreasonable for the vehicle to follow a strict legal code, when there are other reasons (perhaps even better reasons) to act in an illegal manner.

Additionally in this article, Lin brings to the discussion whether it is reasonable to expect this new technology to operate in a better way than we expect 16-year-olds to in their 40 minute driver's examination. Although, the key selling feature of these devices is their safety, which is in part because they should not suffer from human error, they are still in their prototype stage and so too high expectations are currently unreasonable. Despite this, preparatory work for how programmers should program them should occur now rather than after they become a problem. Making the task of determining what is the best way for driverless cars to overcome these no-win scenarios is a prudent idea.

Other articles in *the Atlantic* include their article entitled *Would you pull the trolley switch? Does it matter? The life span of a thought experiment.* [81] by Lauren Cassani Davis published on October 9th where she discusses how driverless cars have breathed new life into this particular thought experiment, where interest in it has ebbed and flowed since its conception. Notably, the author indicates how it has found new life within the realm of driverless cars.

One difficulty, however, rests in how trolley problems seem to be far fetched. This problem is often faced by psychologists who create these scenarios, and when they present them to their research subjects, they are often met with laughter. One way of addressing this is to re-phrase the situation into something far more plausible. In the article, she interviews Lin where

he describes how he presents it to engineers in far more practical examples as follow:

You're driving an autonomous car in manual mode—you're inattentive and suddenly are heading towards five people at a farmer's market. Your car senses this incoming collision, and has to decide how to react. If the only option is to jerk to the right, and hit one person instead of remaining on its course towards the five, what should it do [81]?

Once again in this article, Davis highlights the value of these sorts of experiments in exploring future legal and moral implications of these new technologies.

Other sources of popular literature include *Quartz* where we can find a recent article entitled: *Philosophers are building ethical algorithms to help control self-driving cars* [82] by Olivia Goldhill of February 8th 2018 talks about the ethics surrounding driverless cars in terms of both trolley problems in the works of Nicholas Evans and the broader work of Patrick Lin respectively. Goldhill reports on Evens project where he is trying to translate ethical codes into machine understandable language.

While Evans, at least as presented within Quartz's article, does not take a stance on his own preference in moral theory, he does give some proposals. First, there is the utilitarian model where all lives have equal moral weight, so the controlling algorithm of the car would ascribe the same values to passengers as it would pedestrians in the surrounding environment. He describes an alternative to this in the following way: "We might think that the driver has some extra moral value and so, in some cases, the car is allowed to protect the driver even if it costs some people their lives or puts other people at risk" [82]. As long as they do not actively target people, this may be permissible.⁶

⁶This can also be seen in the BMVI Ethics Document[4] as mentioned previously in 2.4.

The periodical *Popular Science*, as we have seen in Chapter 2 of this thesis, has had a long interest in the topic of driverless cars. In a recent article, as of the writing of this thesis, is called *What moral code should your self-driving car follow?* [83] by Marlene Cimons. Here she explains the various difficulties in programming a driverless car with ethics. Cimons reports on how driverless cars are being trained to behave like humans do in accident situations. To do this, researchers Gordon Pipa, Peter Koenig and Richard Gast, conducted experiments where they used virtual reality to gather people's responses in accident situations. In these experiments, people had to choose whom to slay. In general, they found that children fared better than adults, and humans more than animals, although they were quick to point out that the BMVI's document prohibits age-based choosing [83]. The relevant prohibition in the BMVI document states: "In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical or mental constitution) is strictly prohibited" [4, p. 9], thereby going against the trend that the researchers found in their study. This draws out the tension between a variant of utilitarian practice, as observed in the preferences of users, and hard and fast rules, which are rooted in a deontological ethics, which are generated by ethics codes [83].

Additional and less scholarly popularization come in the form of internet culture. Comical accounts can be found in social media, for example the Facebook page Called Trolley Problem Memes⁷, where the content providers have made two memes about the utilitarian and deontological cars, see figures 4.1 and 4.2.

These accounts pick at characterizations of ethics as applied to driverless cars. Namely that a utilitarian car would opt to kill its user to limit the users suffering, where a deontological car would refuse to interfere, and slay the five persons in its way but relive its user of all guilt.

⁷cf. <https://www.facebook.com/TrolleyProblemMemes/> retrieved July 2tth, 2018



FIGURE 4.1: A meme making fun of deontological ethics in self driving cars



FIGURE 4.2: A meme making fun of utilitarian ethics in self driving cars

4.2.3 In scientific literature

In addition to popular literature, there is a growing body of scientific literature on the topic. In this section, we will consider a few authors who have written recently upon this topic. This will by no means be an exhaustive list of the literature but rather will serve as a means of identifying emerging trends found within the scientific literature. We will begin by presenting a view of consequentialist and deontological ethics as seen in the field of computer science which will be presented in the 6th edition of a textbook entitled *Ethical and Social Issues in the Information Age* [34] written by Joseph Migga Kizza which will be juxtaposed to a standard account of the same ethics in addition

to supplemental information from the Stanford Encyclopedia of Philosophy’s articles on Consequentialism, by Walter Sinnott-Armstrong [35], and Deontological Ethics, by Larry Alexander and Alexander Moore [36]. From this introduction, we will move to other authors who write specifically on ethics for driverless cars. Here our considerations will include, Patrick Lin [48], Neil McBride [38], and Giuseppe Contissa Francesca Lagioia and Giovanni Sartor [39].

4.2.3.1 Ethics in Philosophy compared to Ethics in Computer Science:

When considering ethic for autonomous cars, there are at least two disciplines involved. The first is philosophy, and in particular ethics, and the second discipline is computer science, and in particular, how ethics is understood in computer science being of importance. As we saw in the previous section, popularized literature on the topic constrains itself to a dichotomy of consequentialist and deontological ethics. In this section, we will lay out a general approach to ethics in both of these disciplines, and focus upon providing a survey of deontological and consequentialist accounts using standard sources that each discipline could first consult. In his chapter “Ethics and Ethical Analysis” book *Ethical and Social Issues in the Information Age* [34], Kizza presents the typical ethics that are found within the scientific literature that are used within computer science. The ethical systems break down into the following categories: consequentialism, deontology, human nature, relativism, hedonism, and emotivism [84, pp. 34 - 36]. As previously stated, we will only discuss those relative to our discussion, which are the first two systems. Then we contrast this to their relative articles within the Stanford Encyclopedia of Philosophy’s articles on consequentialism by Sinnott-Armstrong [35] and deontological theories by Alexander and Moore [36].

According to Kizza, consequentialism is an ethical theory where “human actions are judged good or bad, right or wrong, depending on the results of such actions—a desirable result denotes a good action and vice versa” [34, p. 34] This theory breaks down into three subcategories. These subcategories are 1) egoism 2) utilitarianism, and 3) altruism. Egoism is where the individual’s interests and happiness is put above everything else, and what is good is that which maximizes the individual’s happiness. Egoism is further understood in two ways, the first is ethical egoism, where the way people ought to behave is described, and psychological egoism, where the way people actually behave is described. Utilitarianism differs from egoism in that it places the interests and happiness of the group above that of the individual. So in this theory what is good is understood as that which increases the happiness of the most amount of people. Kizza lists two modes of utilitarianism found within the literature. The first mode is act utilitarianism, where we focus on every possible act and consider the foreseeable consequences of such an act and choose the act which has the highest utility. The second mode is rule utilitarianism, where we ought to obey the rules that bring the best utility. The third sort of consequentialism Kizza sees is altruism, where “an action is right if the consequences of that action are favorable to all except the actor” [34, p. 34].

Kizza’s account differs from the Stanford Encyclopedia of Philosophy’s (hereinafter SEP) article on consequentialism [35] in several notable ways. The first striking difference is the depth to which the Sinnott-Armstrong’s SEP article describes the various forms of consequentialism and its history. He begins with the account of Classic Utilitarianism, that is to say as it is found in the works of Bentham, Mill and Sedgwick [35]. Where they hold,

hedonistic act consequentialism. Act consequentialism is the claim that an act is morally right if and only if that act maximizes the good, that is, if and only if the total amount of good for all minus the total amount of bad for all is greater than this net amount for any incompatible act available to the agent on that occasion [35].

While this fits within Kizza's account of act utilitarianism, it differs notably in that here "bads" are factored into the equation whereas before they were omitted.

Sinnott-Amrstrong points out that a great difficulty in addressing consequentialism, as such, rests in the fact there are many different variety of consequentialism which make different claims and have different understanding of utility, and against whom, or to whom, that utility can be addressed. Despite this, there is one key aspect that remains in all its variations, which is that it ascribes to itself the name and that it is consequence-of-action focused. This conception may be too broad as it opens itself up to absurdities, and so various authors may try to constrain it in one way or another, e.g. agent-neutrality. Yet when this happens, the conception of consequentialism becomes idiosyncratic and leaves its broader understanding muddled. Where upon doing this, we become "clear about which theories a particular commentator counts as consequentialist or not and which claims are supposed to make them consequentialist or not. Only then can we know which claims are at stake when this commentator supports or criticizes what they call "consequentialism". Then we can ask whether each objection really refutes that particular claim" [35].

To aid in understanding the various differences in consequentialist theories, Sinnott-Armstron provides the following list of mostly logically independent particular concepts that are often applied and discussed.

- Consequentialism = whether an act is morally right depends only on consequences (as opposed to the circumstances or the intrinsic nature of the act or anything that happens before the act).
- Actual Consequentialism = whether an act is morally right depends only on the actual consequences (as opposed to foreseen, foreseeable, intended, or likely consequences).
- Direct Consequentialism = whether an act is morally right depends only on the consequences of that act itself (as opposed to the consequences

of the agent's motive, of a rule or practice that covers other acts of the same kind, and so on).

- Evaluative Consequentialism = moral rightness depends only on the value of the consequences (as opposed to non-evaluative features of the consequences).
- Hedonism = the value of the consequences depends only on the pleasures and pains in the consequences (as opposed to other supposed goods, such as freedom, knowledge, life, and so on).
- Maximizing Consequentialism = moral rightness depends only on which consequences are best (as opposed to merely satisfactory or an improvement over the status quo).
- Aggregative Consequentialism = which consequences are best is some function of the values of parts of those consequences (as opposed to rankings of whole worlds or sets of consequences).
- Total Consequentialism = moral rightness depends only on the total net good in the consequences (as opposed to the average net good per person)
- Universal Consequentialism = moral rightness depends on the consequences for all people or sentient beings (as opposed to only the individual agent, members of the individual's society, present people, or any other limited group).
- Equal Consideration = in determining moral rightness, benefits to one person matter just as much as similar benefits to any other person (= all who count count equally).
- Agent-neutrality = whether some consequences are better than others does not depend on whether the consequences are evaluated from the perspective of the agent (as opposed to an observer) [35].

From here, Sinnott-Armstrong addresses the various questions related to consequentialism which are as follows. The first question is what is the good and discusses hedonistic and pluralistic accounts. The second question is which consequences there are and addresses the difference between actual and expected consequences. The third question is about consequences of what, and deals with rights relativity and rules. The final question is about limiting the consequences of morality and addresses question about “to whom”.

Regarding questions about what counts as the good, there are various accounts that are provided within the history of this camp of philosophy. Two prominent archetypes are hedonistic and pluralistic accounts. The hedonistic account takes both pleasure and pain when calculating the goodness of an action. This view dates back to Bentham, who notoriously claimed that a game of push-pins⁸ can be just as good as intellectual poetry if it provides the same level of pleasure. This claim, however, may not sit well with people and so there are other variations, such as quantitative hedonism, which gives different levels of goods and non-goods. Other forms take the fulfillment of an agent’s preferences to be the biases of what is considered to be morally good. Once we introduce various values, that is a hierarchy of values or degrees or preferences etc., a new challenge of how to balance these values arises. One method of overcoming this is to take the general welfare of the individual as being the good that has a pluralistic account of values, which can have intrinsic values such as truth that may trump the basic pain and pleasure calculations[35].

Other destinations within the various schools on consequentialism hinge on the differences between the actual and expected consequences of an action. Sinnott-Armstrong continues his account which focuses upon epistemological aspects of utilitarianism. Perhaps the most important claim, for the purposes of this thesis, made by Sinnott-Armstrong is the following:

⁸A children’s game involving pins being pushed about and is on the same level as the games pogs (or milk caps) or jacks (or knuckelbones)

Classic utilitarianism seems to require that agents calculate all consequences of each act for every person for all time. That's impossible. This objection rests on a misinterpretation. Critics assume that the principle of utility is supposed to be used as a decision procedure or guide, that is, as a method that agents consciously apply to acts in advance to help them make decisions. However, most classic and contemporary utilitarians and consequentialists do not propose their principles as decision procedures. (Bales 1971) Bentham wrote, "It is not to be expected that this process [his hedonic calculus] should be strictly pursued previously to every moral judgment." (1789, Chap. IV, Sec. VI) Mill agreed, "it is a misapprehension of the utilitarian mode of thought to conceive it as implying that people should fix their minds upon so wide a generality as the world, or society at large." (1861, Chap. II, Par. 19) Sidgwick added, "It is not necessary that the end which gives the criterion of rightness should always be the end at which we consciously aim." (1907, 413) [35]

Here we see that rather than being a method of calculating every possible action, the general notion is that, on the whole, our actions ought to be geared towards utility. This is important because if we wish to use this sort of ethical school for driverless cars, then it would seem that the literature is operating with a naive conception of this method of ethical reasoning. Despite this, we should still act with the consequence of our actions in mind. But what sort of actions, and how should these actions be evaluated? Sinnott-Armstrong points out two distinct means of doing this. The first is to evaluate the act with the actual consequence of the action. The other is to hold it to what was intended, or was seen as the foreseeable or probable consequence of the act.

The reason for this distinction rests upon the idea found within the example of a woman who gives money to a runaway who wishes to get home. In this example, the woman buys the bus ticket and then the bus on its route

crashes and the runaway died. Was the woman buying the ticket morally at fault? In the first method, she would seem to be at fault in as much as the act of buying the ticket enabled the runaway to ride the bus and then die. In the other method, it wouldn't seem to be immoral, in that the woman could not have foreseen the crash. People who focus on the former are called objective consequentialists and people who agree with the latter are subjective consequentialists, with the important caveat that they are not subjectivists in the sense of what the agent wills but rather foresees as a probable consequence. [35]

Questions of “what” are related to previously discussed issues that were addressed in Foot and Thomson's respective works. The central concept that is taken under consideration is what do we owe to people. The example that Sinnott-Armstrong draws upon is the donor example. There we have one doctor who has six patients. Of these patients, one is perfectly fit and the others are in dire need of organ transplants. Much to their fortune, and the healthy patient's misfortune, the fit patient is a match for organ transplant for all five people on death's door. The problem is, should the doctor turn his scalpel on the fit patient and harvest his organs to save five lives [35]?

Some consequentialist, for example Sprigge and Singer, would bite the bullet and agree that the doctor in such an unusual circumstance should not go with our common moral intuitions on the matter, and how common moral intuitions have evolved to handle common moral situations. Others however are not so keen to do so. To overcome this, they need to introduce a modification to the theory. One such modification is to add values, where killing is worse than letting die. Other methods would be to introduce agent-relativity where we discard the agent-neutrality and see what is judged best from the perspective of the doctor in this case. An other option is to adopt a form of rule consequential, as mentioned in Kizza, where the doctor must act in accordance to a set of rules (which in themselves do not have consequences), which

would allow the doctor to follow the rule “do no harm” and escape carving up the fit patient[35].

The final consideration of consequentialism that Sinnott-Armstrong addresses is in relation to delimiting the demands of morality. This issue is related to the demandingness objection against utilitarianism where it would seem that the demands of utilitarianism make things that are merely permissible or supererogatory obligatory and/or we would need to consider far too many people when acting. Here Sinnott-Armstrong gives a survey of the various answers to this issue. The first is to accept that we should, in fact, be maximizing utility in all cases and accept that the vast majority of people are simply not acting morally. The other option is to follow one of the options described above in the “what” question. The first option is to follow Mill and argue that while we ought to maximize utility, it is permissible not to (therefore derogable). The other option is to have a rule which indicates that we ought to maximize utility in specific ways that delimit the scope of our actions. We can also adopt an agent-relative approach where we can incorporate subsidiarity into our maximization of utility, so we maximize it for ourselves and then family, friends, neighbors etc. Lastly, Sinnott-Armstrong raises that we may give up on the goal of maximizing utility and be happy with *enough* utility. So in this case we may be satisfied with making general improvements in the world and need not be overly concerned with creating the best possible improvement with every single action [35].

Moving towards an introduction of deontological ethical thought, Kizza provides only a cursory account of what this school of thought is and what it considers to be ethical. His understanding of this branch of ethics is given, in full, with the following paragraph:

The theory of deontological reasoning does not concern itself with the consequences of the action but rather with the will of the action. An action is good or bad depending on the will inherent in it.

According to deontological theory, an act is considered good if the individual committing it had a good reason to do so. This theory has a duty attached to it. In fact, the word deontology comes from two Greek words: deon meaning duty and logos meaning science [citation omitted]. For example, we know that killing is bad, but if an armed intruder enters your house and you kill him or her, your action is good, according to deontologists. You did it because you had a duty to protect your family and property [34, pp. 34 - 5].

This account is similar to the SEP's account of deontology in that it is written in reference to consequentialism. However, it differs in a significant way from the text given in the SEP article concerning deontology. Noticeably, unlike Kizza's work, Alexander and Moore's account breaks deontology down into its sub-schools.

Alexander and Moore in their article *Deontological Ethics* [36] provide more information on deontological ethics, and break it down into the following schools with more emphasis given to the first school. This first school is an agent-centered account of deontological ethics, the second is patient (or victim) centered account, while the third is a contractarian account.

In the agent-centered school we, unsurprisingly, find that the focus of the form of ethical consideration is placed upon the actions of the agent who acts. Each agent has ascribed to them a set of permissions and obligations that act as an agent-relative justification for them to act in such and such a manner. Alexander and Moore highlight the agent-relative aspects in the justification of certain actions. Notably, I may have an obligation that others may not have, such as caring for my children or my elderly parents. The same also holds for permissions, where a worker may have the right to access a terminal but Joe Smith off the street does not. Not surprisingly, the central feature to this school of thought is the conception of agency, and “[o]ur categorical obligations are not to focus on how our actions cause or enable other agents

to do evil; the focus of our categorical obligations is to keep our own agency free of moral taint” [36].

In addition to the focus on the agent in this school, Moore and Alexander draw to our attention three sub-schools. These sub-schools place moral emphasis on the intention or mental states of the agent, on the action of the agent itself, and on both equally. In the first, the agent is forbidden from taking actions that he believes are forbidden to do, which takes -according to Moore and Alexander- into consideration three interrelated ideas – belief, risk and cause. Belief deals with what we think or predict will happen if we do an act of one sort or the other. Risk is a related concept where we can foresee some evil that may come from an act but act anyways provided that the risk is low, while cause is directly related to the effect of the act that we achieve. This sub-schools of agent centered deontology

are committed to something like the doctrine of double effect, a long-established doctrine of Catholic theology (Woodward 2001). The Doctrine in its most familiar form asserts that we are categorically forbidden to intend evils such as killing the innocent or torturing others, even though doing such acts would minimize the doing of like acts by others (or even ourselves) in the future. By contrast, if we only risk, cause, or predict that our acts will have consequences making them acts of killing or of torture, then we might be able to justify the doing of such acts by the killing/torture-minimizing consequences of such actions. Whether such distinctions are plausible is standardly taken to measure the plausibility of an intention-focused version of the agent-centered version of deontology [36].

Alexander and Moore state that, in the second sub-school, although the focus is upon the action actually taken by the agent, there is still reference made to the mental states of the particular agent we are considering. Yet, the difference rests, as they put it, in the distinction between pulling the trigger

of a gun and any intention to kill a person [36]. Here, they argue that this distinction depends upon our understanding of causation and how a person's willing the death of another person causes the death of the other. A distinction is made between a direct causing and an omission. The example they give is that of drowning a baby, while if we held the baby under the water we would be at fault for the drowning of the baby but if we saw the baby and omitted to do something it is a mere omission. 'Causing' are further distinguished from 'allowings' where allowing is understood in two ways. The first way is that the agent removes a "defence" that the recipient of the action had against some ill (in their example death) or the second way is some sort of action that returns the recipient to some morally acceptable state.

Causation is further distinguished from enabling (or aiding), where the agent somehow brings it about that another agent can do something morally impermissible, leading to situations where "one is not categorically forbidden to drive the terrorists to where they can kill the policeman (if the alternative is death of one's family), even though one would be categorically forbidden to kill the policeman oneself (even where the alternative is death of one's family)" [36].⁹ A fourth distinction given is taken from Thomson's Trolley Problem, where one may not cause a present evil upon an other person or persons by redirecting it from many people to a few. The fifth and final distinction rests on that "agency is said not to be involved in mere accelerations of evils about to happen anyway, as opposed to causing such evils by doing acts necessary for such evils to occur" [36].

The third sort of agent-centered deontological theory combines the two previously discussed sub-schools. Here, Alexander and Moore state that this places the focus upon intended 'causings,' rather than merely intention of the act in and of itself. They describe the difference in the following way:

⁹Which, as an aside, is contrary to the description given in the ethics' chapter in the computer science textbook.

For example, our deontological obligation with respect to human life is neither an obligation not to kill nor an obligation not to intend to kill; rather, it is an obligation not to murder, that is, to kill in execution of an intention to kill[36].

They state that the advantage to this theory is that it allows us to avoid the undesirable consequences of the previous theory. Using the example of killing again, in the action-centered theory, the agent would still be morally liable for the full brunt of killing, even in instances of negligent or accidental killings. In the intention centered theory, “we could not justify forming such an intention when good consequences would be the result, and when we are sure we cannot act so as to fulfill such intention” [36] This third option avoids this by requiring both the act and the intention behind this act to lead to the death of the other person [36].

Patient (or victim) centered deontological theories differ from their agent-centered counterparts in that they are rights-centered rather than duty-centered moral theories. Variations of this sort of deontological theory invariably, according to Alexander and Moore, play on the dynamic between the rights of one person and the corollary duty in another person. For example, my child’s right to care has a corollary duty on me to care for my child. They also provide an alternative method of interpreting duties which are called ‘usings’. Usings are when other people make use of another person’s body, time, talents etc. against the will of the other person, and fall into a “libertarian” account of deontological ethics [36].

Alexander and Moore describe this libertarian version of deontological ethics as falling in two camps. The first is left-libertarian, which is represented by the likes of Michael Otsuka, Hillel Steiner, Peter Vallentyne; the second is right-libertarian, which can be seen in the works of Robert Nozick, Eric Mack [36]. Central to these theories is the prohibition of using others, against their will, for my own benefit. This notion is similar to Isaiah Berlin’s conception

of Negative Liberty, where a normative wall is built around the individual and we cannot breach said wall without violating a norm.¹⁰ Moore and Alexander also underscore the important feature of these libertarian understandings of patients-centered deontological theories of ethics. Namely, these theories are not aimed at “discrete rights, such as the right against being killed, or being killed intentionally. It is a right against being used by another for the user’s or others’ benefit” [36].

Additionally, these patent theories set out to solve the quintessential trolley problem examples examined extensively in the literature, such as Fat-Man¹¹ or the transplant example¹². Alexander and Moore relate how this version of deontological ethics differs from the agent-centered version, namely, in that this version does not concern itself with either the intention of the act or the act itself but rather with the rights of the victims, or in the case of the libertarian strand of this theory, the usings of the agents.

Acknowledging that each variety of deontological ethics will have its own take on these issues, they note the following:

Take the acceleration cases as an example. When all will die in a lifeboat unless one is killed and eaten; when Siamese twins are conjoined such that both will die unless the organs of one are given to the other via an operation that kills the first; when all of a group of soldiers will die unless the body of one is used to hold down the enemy barbed wire, allowing the rest to save themselves; when a group of villagers will all be shot by a blood-thirsty tyrant unless they select one of their numbers to slake the tyrants lust for death—in all such cases, the causing/accelerating-distinguishing

¹⁰see Berlin’s *Two Concepts of Liberty* in [85]

¹¹Where a fat man, who has just enough mass to stop the trolley, may be pushed in front of the trolley to stop it from killing anyone besides the unfortunate fat man.

¹²Which, as a reminder, is where a doctor may make a donor out of one patient to save the lives of five other critical patents.

agent-centered deontologists would permit the killing but the usings-focused patient-centered deontologist would not[36].

They also note that other problems arise within patient-centered deontological ethics in the handling of *prima facie* wrongs such as killings when done not as a means to some end or for no reason at all. Here Moore and Alexander note that a consequentialist moral calculus, especially one where all people are treated equally, is then needed to be added to supplement our ability to do moral reasoning and prohibit these sorts of acts[36]. Other implications of this patient-centered deontological ethics, especially of the libertarian strand, also make clear the difference between 'usings', which is an action, and failing to aid persons, which is a non-action. They note that the advantage of this is that it follows the alleged intuition that we do not have a claim on others for help. Furthermore, these sorts of deontological theories lend themselves to being agent neutral, where "John has a right to the exclusive use of his body, labor, and talents, and such a right gives everyone equal reason to do actions respecting it" [36]. Moore and Alexander, however, are quick to point out that this leads to the so-called, "paradox of deontology".

They present the paradox of deontology, which is when we are faced with the need of balancing the rights of various persons against non-usings. However, there are situations where, in order to respect two or more individuals' rights, we are faced with the question "Why isn't it permissible or even obligatory to violate the rights of one individual to safeguard the rights of others?"

One solution they pose is to take a step back and acknowledge that we should retain some agent-relative aspects of this theory, especially the core right of the principle agent not to be used, even if it would prevent other 'usings' in the future. They then imagine another claim about adding the numbers of wrongs committed against various individuals; for example it would seem worse to do five wrongs against the five people on the track rather than one wrong against either the fat man or the person on the other track. However, to this

point, the deontologist can raise the objection that ills and misdeeds are not addable. To illustrate this, Moore and Alexander provide an example that my misdeed towards person A and towards person B constitute just that: an act towards them individually. If we were to combine these two acts together, then there would need to be some person C against whom I am committing the greater evil [36].

This move however entails problems for the trolley problem in the following way:

In Trolley, for example, where there is neither agency nor using in the relevant senses and thus no bar to switching, one cannot claim that it is better to switch and save the five. For if the deaths of the five cannot be summed, their deaths are not worse than the death of the one worker on the siding. Although there is no deontological bar to switching, neither is the saving of a net four lives a reason to switch. Worse yet, were the trolley heading for the one worker rather than the five, there would be no reason not to switch the trolley, so a net loss of four lives is no reason not to switch the trolley. If the numbers don't count, they seemingly don't count either way [36].

The other version of deontological ethics is contractarian-based. Alexander and Moore view this as being “orthogonal to the distinction between agent-centered versus patient-centered theories” [36]. In this school of thought, what is seen as being morally obligatory, permissible, and forbidden depends upon the social contract within which these actions are committed. An example of this would be Rawls in his *A Theory of Justice* [36].

As we can see, Alexander and Moore provide a more in-depth account of deontological ethics than Kizza in his own work on ethics in computer science. Even more striking is that within Kizza's work there is a conflation

of the first two branches of deontological ethics, namely agent-centered and patient-centered. It is important to note these differences in ethics, and in the approaches to ethics in technology, especially between those working in the field proper and those working on a philosophical level. The following sections concern themselves with various experts of ethics in driverless cars and provides a survey of their work within the field.

4.2.3.2 Lin:

In the section concerning the popular literature, we have seen that Lin is instrumental in the popularization of ethics for driverless cars. In addition to these works, Lin has published several scientific works on the subject. Here we will analyze in depth a previously discussed article written by him called *Why Ethics Matter for Autonomous Cars* [37], published in “Autonomous Driving Technical Legal and Social Aspects” [6]. In this text, Lin repeats many of the themes we have read in his popular works.

As the title of the article would suggest, this paper is an argument for why we should be concerned about ethics in autonomous cars. Here Lin argues that there is a need to consider ethics for autonomous vehicles in two respects. The first respect is how these devices should operate in the real world. To foster this, Lin suggests using thought experiments to help us reflect upon “no win” scenarios. Once we have done this reflection, we can then consider how we should create policies that reflect our intuitions upon the matter. The second respect is to consider the outside moral aspects of these new cars. This would include the “value of” autonomous cars. These values include, but are not limited to, safety, increased mobility, environment, extra-productivity.

Further examination of Lin’s first point about ethics reveals that there is more problems than solutions at this point. First and foremost driverless cars are beholden to various sets of rules that may be in conflict with each other. He presents us with a story where a car is faced with the choice of hitting a

grandmother, a young girl or do nothing and hit both. A fairly common line of reasoning would be the following passage:

Striking the grandmother could be the lesser evil, at least to some eyes. The thinking is that the girl still has her entire life in front of her—a first love, a family of her own, a career, and other adventures and happiness—while the grandmother has already had a full life and her fair share of experiences. Further, the little girl is a moral innocent, more so than just about any adult. We might agree that the grandmother has a right to life and as valuable a life as the little girl's; but nevertheless, there are reasons that seem to weigh in favor of saving the little girl over the grandmother, if an accident is unavoidable. Even the grandmother may insist on her own sacrifice, if she were given the chance to choose [37, p. 70].

While this might strike us as being acceptable, and as we have read previously in Cimons' popular work reporting on the research of Pipa, Koenig and Gast, see [83], seems to be the preference that people have when driving in simulations, it however runs into conflict with the law and various industry ethical codes. Lin notes especially:

But either choice is ethically incorrect, at least according to the relevant professional codes of ethics. Among its many pledges, the Institute of Electrical and Electronics Engineers (IEEE), for instance, commits itself and its 430,000+ members “to treat fairly all persons and to not engage in acts of discrimination based on race, religion, gender, disability, **age**, national origin, sexual orientation, gender identity, or gender expression”. Therefore, to treat individuals differently on the basis of their age, when age is not a relevant factor, seems to be exactly the kind of discrimination the IEEE prohibits [37, p. 70]. (emphasis added and citation omitted)

This conflict highlights the difficulty in establishing ethical codes and underscores the importance of hashing out the details of what these devices should do now.

The second aspect of ethics involves moving beyond crash avoidance. Here he begins by addressing two common workarounds for ethics. The first is that the car should simply brake. To this, he argues that it is not always the best solution to a problem. For example, the road is slick due to weather or there is a tailgater. The other solution is to hand control back to the human driver. To this, Lin cites that humans need, on average, up to 40 seconds to regain situation awareness depending upon the other activity that they are doing, and in an accident situation that is far too long [37, p. 71]. From here, he moves to the previously discussed topic of crash optimization.

The issue of crash optimization subsides primarily in the underlying preferences that either the owner, manufacturer, or society has when programming the car. These preferences are seen in how we optimize crashes to attain our desired outcome. To this, Lin provides various examples that run contrary to our moral/ethical intuitions on who or what to select in the event of an unavoidable accident. For example:

[t]here may be reasons, by the way, to prefer choosing to run over the eight-year old girl that I have not yet mentioned. If the autonomous car were most interested in protecting its own occupants, then it would make sense to choose a collision with the lightest object possible (the girl). If the choice were between two vehicles, then the car should be programmed to prefer striking a lighter vehicle (such as a Mini Cooper or motorcycle) than a heavier one (such as a sports utility vehicle (SUV) or truck) in an adjacent lane [original citations omitted] [37, p. 72].

However, if we have a preference for prioritizing the safety of other drivers and pedestrians, then we would select the opposite. Hit the (presumably) heavier

grandmother, the heaviest object possible whether it be a truck, a wall, a tree, etc. Either of these preferences, however, are not without problems. In choosing either option, we are targeting classes of persons in accident situations for reasons not entirely within their control. Age, weight, and even the size of family - in terms of the size of car needed to transport them - can be deciding factors that count against, or for, them in these situations [37, p. 72].

In addition to targeting, Lin also brings up other ethical aspects that need to be considered in evaluating ethics for autonomous cars. One issue is related to animals, self-sacrifice, and avoiding harm entirely.

In regards to animals, Lin argues that not all animals can be treated the same. Without needing to go into the value of pets or being able to identify Fido the dog from Nuts the squirrel, and Bambi the deer – where each of these three animals can be treated differently– discrimination amongst these animals can be found in other sources. The simplest distinction can be the amount of harm a collision will cause to the vehicle if struck. Nuts would likely cause little harm, Fido more, and Bambi could total it. Additionally, avoidance needs to be weighed against other objects in the environment. Lin raises questions concerning whether it is permissible to hit some animals rather than others, i.e. Fido is more permissible than Bambi and Nuts is more so than the others, and how do these weigh against unknowns or other vehicles [37, pp. 71-4].

The notion of self-sacrifice is interesting in relation to questions of when is self-sacrifice an appropriate solution to dilemmas. Here, Lin rises some difficulties to the solutions posed by “your good” by the “standard-issue consequentialist.” Here the typically desired outcome is to minimize harm and maximize good, so in trolley-like problems, it would be to maximize the number of people saved, or alternatively put, target the minimal number of people harmed. The example he proposes is the following:

In this thought-experiment, your future autonomous car is driving

you on a narrow road, alongside a cliff. No one and no technology could foresee that a school bus with 28 children would appear around the corner, partially in your lane [citations omitted]. Your car calculates that crash is imminent; given the velocities and distance, there is no possible action that can avoid harming you. What should your robot car do [37, p. 76]?

Lin proposes two possible solutions, provided that all agents are seen as being relatively equal from a utility point of view. We can either slam on the breaks and risk the lives of everyone or the car can simply drive off the edge as we have seen previously in the utilitarian meme 4.2. If the likelihood of death is such that the accident will result in a 1-in-10 deaths of people in their respective vehicle, we could expect to see three deaths in the bus (the 28 children plus their teacher and the driver). Given that, your death would only cause one death and so is the preferred option [37, p. 76].

Other situations may turn out differently, if the odds of harm change and the numbers of people involved changes as well. Difficulties are introduced in the next section where we begin to consider situations where a person may be morally obliged to duck harm. In Lin, ducking harm is the basic principle that we should avoid harm to ourselves in situations where we can easily do so. This is compounded when we consider situations where the lives at stake have intrinsic value either in themselves or value to others. The example that he provides is to consider a situation where the victim is the sole provider for a dependent family [37, p. 77].

How should we deal with this issue? Lin suggests that there is no right answer for humans, as these choices are made on the spot and are simply but reactions to these life and death situations. The same however may not be said of driverless cars. These devices need to be programmed to react, and by making a choice, Lin argue that it starts to look a bit like premeditated

homicide on the part of the programmer, and if left out of the program it seems to be negligence on the programmer's part.

Moving to the second, non-crash related aspect for ethics in autonomous cars, Lin raises new questions about the broader ethical considerations for driverless cars. The first consideration deals with questions of ownership. Who owns the vehicle and does it matter? Should the car "owe allegiance" to its owner or should it consider others as well? The second is concerned with the economic impact. Do these devices pose a problem for the insurance industry? Will there be no more accident? Or perhaps will there be mega-accidents resulting from vulnerabilities introduced by hacking? A third consideration is related to privacy, as the technology relies on GPS data, companies can track our movements and tailor ads for us in the car's apps. Finally, he even raises questions of city revenues, which often depend upon fines from traffic and parking violations, which will be drastically reduced by these cars following the rules [37, pp. 80-81].

4.2.3.3 McBride:

One of McBride's contributions to the topic of ethics for autonomous vehicles is found in his work entitled *Ethics of Driverless Cars* [38]. In this work, McBride provides a critique of truly autonomous driverless cars, by which he means a car that doesn't depend upon any external inputs including GPS (one project working on this sort of driverless car is the Oxford Car). This is contrary to the majority of developments discussed previously in chapter 2 of this work, especially in terms of V2I and V2V. McBride's critique is focused upon a truly autonomous car, using the Oxford Car as an example.

To begin, McBride describes the stated goals of the Oxford Car. Here we can find the familiar citations of safety, freeing humans from being "slaves to the car", enabling the self reliance of the vehicle, and saved time [38, p. 181]. McBride, however, questions the philosophy behind these motivations

and questions whether a solution that completely takes the human out of the loop and it being totally autonomous is preferred or even desirable.

To each of these points, he provides the following counters. In terms of safety, he points to how Google's driverless car often, unbeknownst to its user, drives above the speed limit to match the expected behavior of other drivers. However if this is unknown to the occupant, it is deceptive. In terms of human being enslaved to cars, he raises that far from freeing us, we are only changing masters and turning over our autonomy in the operation of the device to the programs controlling it. The autonomy of the car faces problems in its lack of concern with other users or society outside of it or even its user, which results in a loss of autonomy for the human. When it comes to saving time, he first asks if this time even has value and, even if it does, it will be used well rather than on video games; moreover he notes that "Saving time as a quantitative good may not mean we are gaining anything of human and moral value. And may not reference relationships. For example, driving with my son creates a non-threatening environment where he may open up and we have useful discussions" [38, p. 181].

Underlying all of the supposed goods, McBride draws out the implicit "technological utopia" found in these claims. He surmises this idea with the following quote:

Technology is seen as invincible, provable, permanent, materially-grounded, and reliant only on the solidity of physical laws and mathematics. It is clean, amoral, invulnerable, repeatable, unstained. The only threat of compromise and failure comes from humans included in the loop. Therefore our ultimate goal is the complete exclusion of the humans and the full autonomy of the technology [38, p. 181].

The general goal is to totally remove the human from the loop to safeguard humanity from those troublesome humans. The problem, however, is that

this mentality subordinates humans to technology rather than the other way around as it is promoted. To counter this undercurrent, McBride proposes the *A.C.T.I.V.E.* (Autonomy, Community, Transparency, Identity, Value, Empathy) ethics framework.

Autonomy: As previously discussed, in McBride’s work, total autonomy is seen as being undesirable. To support this claim he provides the following arguments. First, he sees a potential goal of total autonomy to be “deontological disconnection, and ethics” [38, p. 182] where the individual is sovereign and is the ultimate arbiter. It seeks to remove humans from the loop as they are unreliable, especially when compared to the mathematical certainty of the car’s algorithms. The difficulty, however, rests in the fact that the car must travel in the world and interact with humans, and learn from them so that it may interact with them in the driving environment. Once the car starts to do this, then the certainty of its choices becomes tainted. Thus the car would need to limit its interactions with the world, which defeats the purpose of the car. Instead of this model, McBride suggests a conception of autonomy that focuses on human-machine interface which balances human control with the cars ability to control itself [38, pp. 181-182].

Community: This part of McBride’s ethical framework concerns itself with the car’s interactions with society. By this, he means the “cars are created out of the interactions of a community, supported by a community of workers and serve a community. They are elements of a community, both as a participants in a relationship between humans and technology and as a technological mediators in social relationships” [38, p. 182]. These relationships are marked by the growing need for a more connected society as this technology develops and becomes more widespread. So there would be an expansion of workers, i.e. repairmen, energy workers etc., in addition to public servants to regulate them. To streamline this, and to avoid a “private transportation hell”, there is a need to facilitate communication, negotiation, and compromise

in three areas, human–human, human-machine, and finally machine–machine [38, p. 182].

Transparency: This serves as the foundation for ethical engagement in McBride’s work. The key concept here is that there needs to be a clear understanding of what the car is capable of and is not capable of, and that there is no deception involved. Furthermore, transparency also is important in the understanding of the learned algorithm used, and they should be made public so that society can understand them and pit them against each other in competitions¹³ and further the technology and our understanding of its limits. Lastly, transparency should require the testing of these driverless cars in different environments much like how we test human drivers [38, pp. 182-183].

Identity: This concept is tied to how technology shapes people and becomes part of their identity. Reflecting upon the Pixar film *Wall-E*, McBride notes how in the film people have become passive, obese, ignorant and unquestioning about the world and about themselves. For McBride, we need to do a careful reflection on the importance of driverless cars to avoid the tendency of humans to fall into a trend where “the rule of technology renders humans passive, incompetent and hedonistic” [38, p. 183].

Value: Here he draws our attention to the need to evaluate our values when considering this new technology. Values, in his usage, are neither merely a cost-benefit evaluation nor are they solely about our moral values. Rather is is “an analysis of what we value will point to the values underneath. Freedom might be valued above safety, pleasure above health” [38, p. 183]. Other values could include, human flourishing, or economic growth, the advancement of technology, quality of life, individuality, or the good of a homogenized market [38, p. 183].

¹³much like the ImageNet competition discussed in chapter 2 of this work

Empathy: This final section is about the need to cross the “empathy gap”. This gap is the difference between the manufactures of these devices and the consumer / user of these vehicles. In McBride’s words:

A brief poll of family and friends will reveal a wide range of reactions to a driverless car. Some regard it with fear and revulsion. Wary enough of driven cars and the danger of the roads, the prospect of a driverless car is completely unacceptable. Others may view driverless as a novelty, and want to know ‘how it works’ out of interest or a need for assurance about the reliability of the technology. The latter point relates to a need for transparency and a reluctance to treat a driverless car as a black box initially and get into it without the sufficient knowledge as to its technology and its reliability.

For some males, the prospect of being driven around by a driverless car may bring about a primitive sense of emasculation [38, p. 183].

Here, we need to consider how people feel about these devices and make them approachable, rather than act rashly towards people. An example of this harsh attitude can be seen in a quote given by the head of the Oxford Car project Paul Newman, “If you don’t believe this you need to leave ... this has to be a true thing (original citation omitted)” [38, p. 183]. Rather than being imperialistic about the future of the technology, McBride suggests that we meet the future consumers of these devices and address the underlying fears that they have. This would have the benefit of alleviating them of their fears but requires a constant stream of communication [38, pp. 183-184].

McBride concludes his reflection of the qualities needed for an ethics for autonomous vehicles by arguing that autonomy is not in fact the goal of these technology. Rather, the technology should be seen as serving a support roll in society, promoting connectivity and is, more importantly, a reflection on how

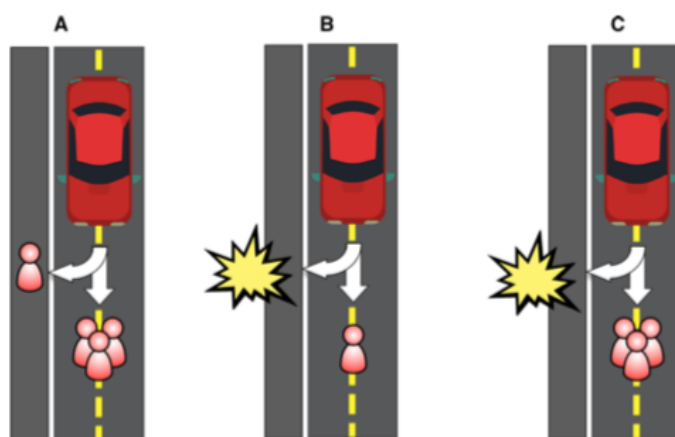


FIG. 4.3. Three scenarios involving inevitable collisions.

FIGURE 4.3: The Ethical Knob's Trolley Problem

we should conduct ourselves in the three new relationships, human–human, human–machine, and machine–machine.

4.2.3.4 Contissa, Lagioia, Sartor:

In their paper entitled the *The Ethical Knob: ethically-customisable automated vehicles and the law* published in 2017 in “Artificial Intelligence and Law” Contissa et al. present a way of customizing ethics for driverless cars that can take into consideration user preference.

They begin by presenting their reader with the now familiar trolley problem. They present the following situation with three outcomes as seen in 4.3. Here we have three scenarios that present the driverless car with three unavoidable situations. In scenario A, the car has the choice of hitting one or three people. In situation B, the car has the choice of slaying a pedestrian or its own passenger. In scenario C, the car has the choice of killing three pedestrians or its own passenger [39, p. 2].

The authors present two alternative methods of handling these sorts of situations. The first method is to introduce a mandatory ethics setting (MES)

where all vehicles handle these situations in the exact same way. The alternative is a personal ethics setting (PES) where each car is equipped with its users preferred ethics.

There are advantages to both the PES and MES when implemented within driverless cars. The chief advantage of PES rests in that it allows autonomy for the user / owner of the vehicle in selecting the ethics that the car they are driving to follow, and that may include a more selfish preference, which they note—citing Bonnefon et al. [51]—is the preference of people when in the car. The advantage, however, of MES equipped cars sits in that society knows how each vehicle will act and will limit socially undesirable outcomes, minimize the risk that vehicles may pose to society, and finally it allows for the sacrifice of the owner of the vehicle when a greater number of lives may be saved [39, pp. 2-3].

Tensions, understandably, exist between the PES and MES conceptions of ethics in vehicles. Contissa et al. note exactly the tension between consumers wanting cars that save the most amount of lives, yet when they are in the target scope, they prefer not to be the one slain. To resolve this issue, they introduce the idea of an “ethical knob” which allows three preferences that the car may be set with. The first is an altruistic mode, which prefers third parties. The second is an impartial mode that gives equal importance to both the passenger and third parties. The third mode prefers the passenger and is called the egoistic mode [86, p. 5].

Assuming that the driverless cars are installed with these knobs, then the users would be able to set the PES within the vehicle they are riding in. This should help restore consumer confidence in the implementation and use of these devices in a mixed environment. The three modes present within the ethical knob are controlled by an algorithm that takes the user defined preferences for the weights of people into account. The affects that these settings have on accident situations should also be taken into consideration when considering

issues of liability and how responsibility should be meted out. In their example, this is most important in situations B and C [39, p. 6].

The algorithm designed by Contissa et al. extends the basic three settings of egoism, impartiality, and altruism and allows for continuous preferences and employs a probabilistic approach. They created these extensions to allow for more realistic, that is to say non-deterministic, situations on the road. The continuous setting specifies

the weight of the life of the passengers relative to that of third parties. For this purpose, relative preferences can be determined according to the following linear function:

$$y = 1 - x \tag{4.1}$$

where x , namely, the knob's position from left to right, indicates the importance for the passenger's life and y is the importance for the lives of third parties [39, p. 7].

This allows for the knob to give relative values to both passengers and third parties on a spectrum, where the driver can value their life more (or less) than third parties and the vehicle will take that into consideration.

This continuous setting is then parred with a probability of specific outcomes for negative outcomes of an accident and follows the following method. We take the expected utility and disutility of an accident and multiply it by the PES values given by the user in their setting of the car's ethical knob. The example that Contissa et al. give is the following. A user has defined their life as being valued at .6 and third parties at .4, resulting in the user valuing their life 1.5 times more than other people's. In an accident situation, the car predicts that swerving has a .5 probability of killing the passenger and going straight will have a .9 probability of slaying the third party. In their example, we then arrive at two different disutilities where the "expected disutility

of swerving will be $0.6 * 0.5 = 0.3$, while the expected disutility of keeping a straight will be $0.4 * 0.9 = 0.36$ " [39, p. 8]. From here the car should choose the option which has the lowest expected disutility.

The expected disutilities of the different parties can then be compared to each other in situations where it is assumed that death is the only outcome and there is only one passenger and third party involved in the following way.

$$Dis(c_i, a_i) = R(a_i) * Pr(Death(c_i, a_i)) \quad (4.2)$$

Where $Dis(c_i, a_i)$ is the disutility resulting from the driverless car's behavior c_i is the affecting an agent of type a_i and $Pr(Death(c_i, a_i))$ is the probability of death that choice c_i has for a_i [39, p. 9].

This then can be extended to incorporate more persons either in terms of passenger, third parties, or both according to the situation with their following formula:

$$TDis(c_i, a_i) = Dis(c_i, a_i) * n(a_i) \quad (4.3)$$

Where $TDis(c_i, a_i)$ is the total disunity expected to a number of agents a_i [39, p. 9].

Finally Contissa et al. introduce a normalized total disunity $NTDis(c_i, a_i)$, which is the effect of choice c_i on agents of type a_i where "we have to divide the total disutility of choice c_i (where i can be 1 or 2) by the sum of the disutilities of the two alternative choices c_1 and c_2 affecting a_1 and a_2 , respectively. Thus, we obtain the following formula:"

$$NTDis(c_i, a_i) = \frac{TDis(c_i, a_i)}{TDis(c_1, a_1) + TDis(c_2, a_2)} \quad (4.4)$$

[39, p. 10]

After calculating the expected disunity of the situation at hand, the question of how to interpret the outcome and act accordingly remains. The authors

explore two possible means of interpreting the output. These two means are utilitarian and Rawlsian. The standard utilitarian approach is to choose the option with the least expected disutility. The Rawlsian approach, however, offers a bit more nuance. Here, they note that it would seem that under the veil of ignorance, where we do not know if we are the third party or the passenger, we should opt for the decision with the lowest disunity. Nevertheless, if we can take into consideration other aspects such as personal injury, we can add more nuance to the situation. The example that they give is the following: “Assume that by proceeding the AV would cause with certainty the pedestrian to become paraplegic, while by swerving it would cause with certainty each one of three passers-by to lose one leg each. Then it might be argued –though this conclusion is very debatable– that swerving is preferable to proceeding, on grounds of equity/equality” [39, p. 12]

4.3 Concluding Remarks

In this chapter, we have laid out the current state of affairs for ethics as it is understood within the literature. We began with a recounting of the “Trolley Problem”, which typically takes the center place of ethical considerations for driverless cars. This we can see both in the popular and scientific literature on the topic. From here, we then outlined the broad understandings of ethics as used both in the computer science and within the philosophical communities. We then took a more detailed look at three examples of specific literature concerning ethics and driverless cars. These accounts, however, are not without difficulties and we will address this in the next chapter.

Chapter 5

Difficulties in Standard Accounts and a Solution

5.1 Introductory Remarks

As we have seen in the previous chapter, there is a wide spectrum of beliefs concerning ethics in autonomous cars, ranging from Lin's broad considerations of ethics both in terms of ethics within autonomous vehicles and outside of them and McBride's general proposal with his ACTIVE ethical framework to more specific methods of implementing ethics with Contissa et al.'s ethical knob. Within each of these approaches we can find either a version of consequentialist or some version of deontological ethics built within it, and each are designed to address specific issues relative to the strengths of each. While these underlying ethics take various forms, the general trend is to pit hard and fast rules, as seen in deontological ethics, against a minimization of harm done, or, alternatively, the maximization of lives saved when we are presented with either a dilemma or trolley problem-like situations. While these efforts are all worthwhile, there is a problem that haunts each one. This problem is that we are still not left with a clear understanding of what sort of ethics is used

within each of these proposals and how ethics is compatible when considering its application within autonomous cars.

But in what respect is this the case? When considering ethics in general – and recalling the works of Kizza, Sinnott-Armstrong, and Alexander and Moore – we are left with two different (although not incompatible) understandings of ethics as such: the simplified versions, as found within Kizza, and a more complicated accounts within Sinnott-Armstrong and Alexander and Moore respectively. To illustrate the difficulties we find in the more specific accounts, as previously described, let us consider an example which is based upon the now familiar trolley problem originally created by Foot and Thomson.

Example “Turning”: You are riding in your driverless car and when your car turns the corner there appears five workmen in front of your car and one workman having his lunch in the adjacent lane. Your car has three principle options. First it may swerve out of the way into the other lane but in doing so it will hit another workman having his lunch. Second, it can swerve the other direction and hit a wall and risk injury to you its occupant and itself. Thirdly it may simply go straight and hit the five workmen. What should the car decide to do? As previously discussed there is a wide variety of ethical solutions to choose from in the ethical toolbox, but as we have noted they seem to fall into two main camps, either consequentialist or deontological. In this basic example, each has its own problems that we will presently address.

5.2 Difficulties in applying consequentialist ethics:

In regards to the consequentialist approach, we will need to address both the simplistic and more complex accounts in turn. We will begin with the Kizza’s simplistic account and address issues of egoistic consequentialism in terms of ought and psychological. Then we will address the utilitarian account in terms

of act-based and rule-based. And finally we will address the issues concerning the altruistic account.

To recount, Kizza's egoistic account states that we ought to act in such a way to maximize our own individual happiness. This can be understood in two ways either ethical - as a description of how the agent ought to act - and psychological - as a description of the agents thoughts and how the agent actually acted. In our "Turning" example, a car that is fitted with this preference is still faced with its three options: 1) swerve and hit the workman having lunch, 2) go straight and hit the five workmen ahead of it and finally 3) swerve and hit a wall. Within the ethical egoistic account, it would seem that the agent ought to act in such a way that it would maximize its own happiness. And to do this, it would seem that the car should target the lone workman who is having his lunch as he poses the least amount of risk to the vehicle and its occupant.

Two interrelated difficulties arise with adopting this process though. The first aspect is proposed by Lin who was apt in pointing out that the long awaited and sought after promise of safety comes at the price of us needing to determine the best means of targeting the outcomes of crashes. And pre-determining the selection of people based upon their relative risk to ourselves, such as their weight and size, or other features such as age and sex, poses problems for our moral intuitions. The BMVI's ethics commission report on autonomous driving concurs with this assessment, arguing in the report that there is a strict prohibition on the offsetting of people against one another both from legal and moral grounds [4, p. 18]. The second related difficulty is drawn from Contissa et al.'s Ethical Knob, where we are confronted with situations where there is only a risk of harm. Strictly speaking, in the ethical egoist account, the happiness of the agent is what counts most, regardless of other agents within the playing field. In our "Turning" example, while the first option may conceivably produce the highest utility for ourselves, it would all but guarantee the death of that lone workman, where if we were to hit the wall

or even the larger group, the brunt of the impact would be lessened minimizing harm overall provided that, as the mass of the target increases, our relative risk increases but so too does the benefit of the object or objects hit. This would follow the example found in Contissa et al.'s work where "in a situation in which there is a 100% probability of killing the pedestrian by proceeding and 10% probability of killing the passenger by swerving, the AV would choose to proceed. We may doubt that such a behaviour would be legally acceptable. It would be up to every legal system to determine the threshold for acceptable selfishness" [39, p. 10].

In regards to the psychological version of egoistic consequentialism, we will set it aside as the agent we have in mind, that is the driverless car, has neither a psyche nor exists in the world so that we can describe its behavior in such a way that we may adopt this method of consequentialist thought.

In Kizza's account of the utilitarian model of ethics, we are confronted with both the act-based and rules-based versions. In the act-based model of utilitarian ethics, we need to measure the utility / happiness that would be created by each of these three acts in relationship to the group as a whole. Act-based utilitarianism is susceptible to the same sort of criticism provided by Lin that we find within the egoistic account, where the selection of who to hit poses a problem for our ethical considerations, where sitting down to calculate who to strike down is a far cry from the reaction-based response that a typical driver would have in this situation and, to paraphrase Lin's own words, seems to move from the realm of an unfortunate accident into the realm of premeditated murder [33].

Both Kizza and Sinnott-Armstrong describe a rule-based utilitarian line of ethical reasoning which may dodge this problem, namely by changing the object of ethical consideration. That is to say that the agent is acting against some norm rather than some other agent, or some sort of state of affairs.

In our example, a norm that could be reasonable to take under consideration is the ninth rule proposed by the BMVI's Ethic Commission, where:

In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical or mental constitution) is strictly prohibited. It is also prohibited to offset victims against one another. General programming to reduce the number of personal injuries may be justifiable. Those parties involved in the generation of mobility risks must not sacrifice non-involved parties [4, p. 11].

In the application of this rule to our three example outcomes, we run into two intertwined problems. The first problem is how can we balance the prohibition of not offsetting victims against one another with the permission to program a reduction in personal injuries? Secondly, how does the prohibition of not involving non-involved parties affect the outcome? Either options seem to violate some part of the norm in force.

For our purposes, we need not solve these two problems, as we could simply sidestep the issue by choosing not to adopt the proposed norm. So we will temporarily set it aside. Nevertheless, even if we do not want to adopt the proposed norm above, we may still consider more general rules that coincide with our general moral intuitions such as “thou shalt not kill” taken from the Decalogue or any other variations of that norm found within a multitude of ethical and moral codes both throughout history and around the world. Given the near universality of this norm, we would be hard pressed to simply set it aside.

In our example, our agent, that is the driverless car, has three options, all of which violate this basic rule of “thou shalt not kill”. To break down this problem, we will first recall the Hohfeldian corollaries that we discussed in the previous chapter; we note that this prohibition is a duty placed upon an

agent, namely not to kill, and is related to a claim in some other agent for the duty-bearer not to kill them. This reflection is often brought up in discussions about the Trolley Problem, especially in the work of Thomson herself and other authors' critiques.

The "Turning" example that we have been examining in this section is based upon the so-called "bystander" example, which we have examined previously in the trolley problem section, where, to jog our memories, a bystander is near the switch that may change the oncoming trolley onto a new track, thereby diverting it from hitting five persons to only one. When considering our example, the car has the choice of switching lanes into either the other lane, towards the wall or continuing down its current lane. In Thomson's original example, "bystander" is contrasted to another example entitled "Big Man" where a thin man is able to push a big man onto the track, where once the big man is struck he will die, but will also cause the trolley to come to a stop.

To understand the difficulties here, Thomson draws our attention to two different principles provided by Foot.

- (i) *Letting Five Die Vs. Killing One Principle*: A must let five die if saving them requires killing B [87, p. 360].
- (ii) *Killing Five Vs. Killing One Principle*: A must not kill five if he can instead kill one [87, p. 360].

For an agent to make a decision in such a case, Foot applies the notions of negative and positive rights where a negative right is "markedly weightier" so that the first principle of letting five die is superior to killing one, as letting five die entails a violation of the positive rights of the five for aid of some sort, whereas killing one is a violation of the one's negative right of not to be killed [31, p. 4] [32, p. 206]. Foot maintains that, in the second principle, the agent must not kill five when he can kill one, in as much as such an action entails only a violation of one person's claim against the agent rather than a violation of

the claims of five persons. In her 1976 paper, Thomson agrees in part, though rather than imposing an obligation, she argues that the driver has permission to turn the trolley [32, p. 207].

While the second option seems to sit well with us as it is a regrettable act that minimizes harm in a “no-win” situation, the first option of letting five die rather than killing one seems worse. This is especially so in some variations of the consequentialist calculus where we need to maximize the utility of states of affairs.

Later, in a paper from 2008, Thomson further refines the second option by adding a third principle:

- (iii) A must not kill B to save five if he can instead kill himself to save the five. [87, p. 365]

This principle, aims at capturing the intuition that “the altruistic bystander is not entitled to assume that the one workman is equally altruistic, and would therefore consent to the bystander’s choosing option (ii) (that is killing the one workman rather than five). Altruism is by hypothesis not morally required of us” [87, pg. 367]. Additionally she furthers the first principle of letting die with a fourth addition that:

- (iv) A may let five die if the only permissible means he has of saving them is killing himself.

This aims at capturing that self-sacrifice is a supererogatory, or in her own words altruistic, act and by its very nature may not be required of anyone.

In sum, our agent now has the basic norm not to kill others, and this norm takes the form of a prohibition of killing others, or alternatively an obligation not to kill others. This duty, as Krammer pointed to and as we discussed in the second chapter of this work, simultaneously creates a claim, or put otherwise

a right, in the other agent which is intrinsically imposed upon our duty-bearer not to kill them.

In our “Turning” example, our agent is faced with slaying the occupant of the car, the five workmen, and the workman having his lunch, but all can make this claim upon the self-driving car. Which principle applies? Is the situation more akin to letting five people die and killing one, or is it closer to killing one rather than five? Any act the autonomous car takes leads it to a situation where our agent is forced to violate the basic norm of not killing, as its actions will lead to either of one or five of the workmen’s deaths, or to the demise of the occupant of the car. To elucidate this, we turn to the Foot and Thomson’s principles.

Within the original “Bystander” example, a bystander (or trolley operator) has the option of continuing forward or changing tracks and killing five or one respectively. Here both Thomson and Foot agree that it is the second principle of killing one vs. killing five that applies as the general duty to not kill is seen as a negative right, and the bystander / operator have a general duty not to kill others. Likewise, in “Turning” we see that the same duty applies to the driverless car. A further application of Thomson’s third principle does not seem to apply as the option of self destruction leads to the death of the car’s occupant.

But that is not all that can be said on the matter. When we consider these first two principles – or rules – and compare them to a rule-based consequentialist calculus, different problems emerge. In regards to the letting five die vs. killing one, the application of this rule leads to a counter-intuitive state of affairs where five people are dead whilst one remains unharmed due to our inaction. In the second rule, our action seems permissible, yet violates a more basic rule of not intentionally killing another human being, and furthermore it is the rule this whole dilemma rests upon. The third rule, that is that we may not kill some other person when we can kill ourselves doesn’t apply as the act of the car killing itself causes the death of its occupant. Additionally

as this rule is supererogatory, in that we are never required to kill ourselves, it is indeed possible that our agent may choose not to do so even if there was no occupant. The final rule makes it permissible for there to be a state of affairs (both in reality and normatively) that five should die.

What is more, the general application of this sort of rule-based consequentialist account runs into a problem of how we can weigh violations of rights and duties against each other. In these examples either with “bystander” or “turning”, what exists is the claim and duty to not kill an individual nor an aggregate of individuals. Every act leads to the unfortunate demise of some individual or individuals but only to the violation of the norm not to kill the individual in question. So in “turning” hitting the workman on his lunch break violates “thou shalt not kill”; similarly hitting the group of five doesn’t violate a third-manesque rule of “thou shalt not kill five people” but rather it violates “thou shalt not kill” in regards to the first, second, third, fourth, and fifth workman; lastly the car turning into the wall and slaying the occupant violates the rule against the occupant. As these rules are all the same, it is hard to say that any option is preferable or worse than the other.

Moving towards the altruistic account, that is where the vehicle will only take the utility of other users into account, even to the detriment of itself or its user(s), also poses problems. The issues here are similar in many ways to the issues found when considering the egoistic account. The issue however rears its head when the passenger has a higher chance of injury or death than any other of the other potential victims. So in our example, if there is a 15% chance of the driver being injured while the group of men in the lane only have a 10% chance of injury (say due to some barricades and other protective measures they have wisely installed), the vehicle will still choose the higher risk and thereby the lower utility option of running into the wall and needlessly endanger its user’s life.

In addition to the aforementioned difficulties in implementing a consequentialist ethics, other difficulties found within the SEP article remain and in

particular, they are related to hedonistic act consequentialism, hedonistic and pluralistic accounts, and finally actual and expected outcomes.¹

When considering hedonistic and pluralistic accounts of consequentialism, and their application within driverless cars, there are problems that emerge from the start. To recall, hedonistic accounts measure both pleasure and pain within their calculus when deciding what the best state of affairs should be, whereas a pluralistic account is pluralistic with respect to the number of values that exist to temper or modify the value of pleasure and pain in the calculation.

The most fundamental difficulty that a driverless car has when confronting the hedonistic account is how it can calculate pleasure and pain, and for whom it should calculate it. As we have discussed in the previous chapter, the ethical agent in question is none other than the car itself, which is not capable of experiencing either pleasure or pain. If the car is acting as an agent on behalf of its owner, or current occupant, how does it know what sort of actions would bring them pleasure or pain? In short it cannot, as each of these are subjective experiences relative to the subject and are inaccessible to the car.

In terms of the pluralistic account of consequentialism, similar problems arises when we consider how the vehicle can know these various norms and how they would be perceived among other agents within the system, if it is even a factor. Even if the car can perceive these values, there is a difficulty in which values are relevant and how they should be weighed against each other. While what values the calculus may take under consideration in itself is not an issue, the weighing of said values is. Considering our “Turning” example, let us assume that every person has the value of preserving their life and minimizing harm done to their property. Even though they have these values, it is reasonable to assume that they don’t hold them to the same degree. The workman having his lunch may value his life less than any of the other five workmen or the user of the driverless car, or even the driverless car itself (assuming it has a self-preservation directive). The self-driving car may have

¹Issues of to whom do we owe what are captured in the egoist and altruist accounts above.

the lowest preference for self-preservation, but that preference may be stronger than the value of the owner of the wall for their property not to be damaged. In any event, all of these would need to be taken into consideration in order for the vehicle to make a decision, which it simply does not have access to.

5.3 Difficulties in applying deontological ethics:

Perhaps the first difficulty in addressing deontological ethics for driverless cars is the lack of its proponents, and well argued claims for it. Kieth Abney in his chapter entitled “Robotics, Ethical Theory and Metaethics: A Guide for the Perplexed” [88], describes the typical approaches to ethics as applied to driverless cars in the following way:

The usual divide within rule-based approaches is between those who say one must intend to obey the rules, no matter what – even if the consequences will be bad (deontologists, associated with Kant), versus those who say the main or only rule is always to make the future consequences as good as possible – *ends justify the means* (consequentialists, most commonly represented by utilitarians, who tend to measure the ends or results in terms of happiness gained or lost).

Within the previous chapter’s survey of the leading works of Lin [48], McBride [38], and Contissa et al. [39] they themselves don’t make reference to deontological ethics in driverless cars, or if they do it, it is in passing.

Other works, however, such as the Ethic Commission Report for the BMVI [4], Kizza [84], for ethics in computer science, do make more references to this ethical theory, but specifically to the rights and will of the device’s human user. In the other way of looking at how deontological ethics are applied within the topic of driverless cars, heavy emphasis is given to the duty, or alternatively

put hard-coded rules implemented within the device. This approach is seen in popular sources such as [89] where the authors propose utilitarian and deontological solutions to various polemical scenarios for a driverless car, or as in Abney's text, as simply following hard-set rules.

Turning our attention to the deontological side of these "rule-based approaches" we have also noted that there are two different, though related, understandings which focus upon various duties as having a central role in understanding this ethical theory. Underpinning all of this is the conception of will of the ethical agent in the actualization of his or her duty towards the bearer of some adjoining right. As discussed previously, deontological ethics breaks into two broad sub-categories: either patient-centered or agent-centered, both of which depend upon some set of rules and the will of the agent, and recalling further our previous chapters work on will and interests theories of rights, this ethical school rests solely and exclusively on will theory.

My objection to the use of deontological ethics is two-fold, and exists on the foundations level of this ethics, and so, unlike the previous section, I will not address each sub-school that was addressed by Alexander and Moore in their article [36]. Both of these objections are made in reference to the sort of moral agent we have in mind and relates to the duties and will of the autonomous car.

The first objection is pragmatic and relates to the nature of AI and in particular its implementation in driverless cars. As we discussed about driverless cars in the world, these devices are not hard programmed with a large set of if-then rules. Instead they are trained using some variety of machine learning that is chosen by its developer. The choice of using machine learning over hard programming the device is a result of the fact that first, it provides better results, and second, you don't have to program in every conceivable situation that the car may face, which is itself an impossible task. This has direct bearing on the impossibility for the implementation of some variant of

either patient-centered or agent-centred deontological within these vehicles, for they simply cannot have these sets of rules to act upon.

The second objection relates to the importance of will in this school of ethical thought. In a previous chapter of this dissertation, I presented an argument for how driverless cars can be considered normative agents so that the foundations of applying norms within these artificial agent can be discussed. There I presented two theories of norms that are incompatible with each other, which are will theory and interest theory. In order to render the legal horn of our considerations intelligible, and therefore make driverless cars that have duties towards other entities, and even have the right to participate in the norm-governed activity of driving, which itself defines the function of this artificial agent, we needed to adopt interest theory. This again was the result of two considerations: first the lack of will these entities (and individuals in general) have in the enforcement of duties and rights within criminal law, and second interest theory can account for that in addition to civil law. This adoption readily allows for the realm of the norms to co-exist with the realm of facts. As a result, deontological ethics is precluded on the foundational level.

5.4 A Virtue Ethical Approach:

If a consequentialist and deontological ethics have faults when applied into driverless cars, we are left wondering what, then, is a proper ethics for these new devices. My proposal is that rather than using an ethics that is either designed to calculate the utility, in terms of pleasure or pain or some pluralistic account, and either in states of affairs actual or expected or perhaps some rule, or a deontological ethics which relies upon the will of the agent and the application of rules, we should instead prefer an ethics that is agent-centered such as virtue ethics. Broadly speaking, within virtue ethics, each agent contains a set of virtues that enable the agent to act in such a way that it would be considered virtuous.

Classically this is presented by Aristotle in his *Nicomachean Ethics* [40], where the use of virtue is related to the excellence of some human. Roger Crisp in his introduction to the *Nicomachean Ethics* notes that virtue, within the Greek context, can also be conceived as being broader than applying only to humans. “a horse that ran fast could be said to have a ‘virtue’ or excellence, in so far as it performed well its characteristic activity” [40, p. xiv].

This conception of creating a specific set of virtues has existed on the periphery of discussions related to robot ethics for quite some time, but has recently been picked up in greater length by Nicolas Berberich and Klaus Diepold in their article “the Virtuous Machine - Old Ethics for New Technology?” [41] where the conception of a virtue ethics for robots has recently been expanded upon. In this text, Berberich and Diepold set out to answer the question “How can we build a machine that, owing to its constitution, acts appropriately in arbitrary situations [41, pg. 4]?” The answer to which they find in virtue ethics.

I am in agreement with the authors that this is indeed the best place to start looking if we want to build a virtuous machine. If we consider my first objection to the use of a deontological ethics for driverless cars and its incompatibility with the sort of agents that these devices are, as we discussed in the chapter concerning the technical aspects of these machines, they learn specific tasks through trial and error and the building up of vast data sets, with the application of some variety of machine learning tools. As a result of this, we end up with an agent who learns to do something and builds up habits on how to react when it perceives a certain sort of situation. At the beginning of its training, driverless cars often fail, yet after millions of miles of practice, it is able to perform its task often far better than a human.

Berberich and Diepold also pick up on this feature of what they call autonomous moral agents (AMA), of which a driverless car is one example. Here they also root the sense of using virtue ethics in AI for runner discipline cybernetics, which returned a teleological understanding of things from its exile in

the last century and in particular of teleological understanding of living being and machines [41, p. 5]. This reliance on teleology introduces their reliance upon Aristotle for the formation of ethics for AMA.

A key feature of Aristotle's ethics is that eudaimonia, which is happiness or excellence, of man is acquired through the practice of virtues. These virtues vary and are both physical and intellectual, and importantly are gained through habituation. The parallel here is that while we are not considering the acts of a human agent, we are considering the acts of an artificial agent that learns through habituation. If this artificial agent, that is to say the driverless car, can be "raised in good habits", then it will act in a virtuous way, and furthermore will be endowed with an ethics by virtue of its acting in the world.

Berbrich and Diepold begin by addressing two objections to the use of virtue ethics in regards to AMAs. The first is that ethics is anthropocentric by nature, citing that rationality is the basis of ethical virtues and actions. While admitting that it was indeed the case in the past, they argue that machines could also reason and have the same dispositions for virtues as humans. The second objection they address is the use of eudaimonia, which is often understood in a utilitarian sense, which machines are incapable of having. They rightly argue that understanding is wrong but rather it should be understood as:

The Greek term eudaimonia has a much broader meaning and refers mainly to a successful conduct of life (according to one's ergon²). A virtuous machine programmed to pursue eudaimonia would therefore not be prone to wireheading, which is the artificial stimulation of the brain's reward center to experience pleasure [41, p. 8].

²Here they explain their meanings in a previous passage: "Aristotle begins his *Nicomachean Ethics* with the teleological thesis that everything pursues a goal (telos) and that every species has a specific function or purpose (ergon). This function and thus the goal of all pursuit is not provided externally, but lies within the nature of each living being. A good life means for Aristotle to fulfill one's ergon through the species-specific way of life and to thus exhibit virtue (arete)[41, pg. 4]."

It is upon this conception of eudaimonia that they base their virtue ethics upon, and as a result follow Aristotle and St Thomas Aquinas and argue for certain virtues that an AMA needs to be trained with to become ethical. These virtues include prudence, courage, temperance, justice, gentleness, and friendship to humans.

While I agree that virtue ethics has the advantage of constructing an ethics for AMAs, and specifically to our topic at hand for driverless cars, that corresponds to the bottom up approach according to which these devices are built and learn their tasks with, as opposed to the top down approaches of utilitarian and deontological theories, I disagree that a eudaimonic-centered virtue ethics is the best starting place to build a virtue ethics for AMAs and specifically driverless cars.

My objection is that their conception of eudaimonia falls short of Aristotle's understanding, especially as they argue that AMAs may reason. To frame this objection, let us consider the Ethics where Aristotle begins his consideration of what the good is that man ought to orient his life towards. In chapter 7 of the first book (Butler page 1097a), we find his considerations about the good that he is looking for. He notes that:

... it appears to vary between different actions and skills: it is one thing in medicine, another in military science, and so on in all other cases. What then is the good in each case? Surely it is that for the sake of which other things are done? In medicine it is health, in military science, victory, in housebuilding, a house, and in other cases something else; in every action and rational choice the end is the good, since it is for the sake of the end that everyone does everything else. So if everything that is done has some end, this will be the good among things done, and if there are several ends, these will be the goods [40, 1097a].

He continues his consideration and concludes that “Happiness in particular is believed to be complete without qualification, since we always choose it for itself and never for the sake of anything else [40, 1097a - 1097b].” All of the virtues, while goods in themselves, are also goods for the end goal of happiness and to live the happy life in accordance to the person’s nature.

Aristotle furthers this line of reasoning by noting that different practitioners of crafts have their own characteristic activity, the flute player, the sculptor, the tanner etc. and furthermore that every part of man seems to have a function like the eye, ear etc. Yet, it seems off that man himself should not have his own characteristic activity. To answer this, he looks at the specific difference of man from other living creatures, where it seems, as Berbrich and Diepold noted, that reason differentiates us from both living beings with animate souls (horses, cats, dogs) and vegetative souls (plants). [40, 1098a] Man’s rational nature (and by extension rational soul) differentiates him from other creatures. The perusal of goods for the rational should set the mark for the sort of characteristic activities that man should pursue. It is here were Berbrich and Diepold make subtle mistakes which we will examine in further detail.

Their use of eudaimonia for AMAs relies upon a parallel between the rationality of man and the supposed rationality of these artificial moral agent. They state the following:

This [anthropocentric ethics] might have been correct in the past, but only because humans have been the only species capable of higher-level cognition, which, according to Aristotle, is a requirement for ethical virtues and thus moral action. If there was another species, for example a machine, with the same capacity for reason and dispositions of character, then it appears probable that its arete would also lie in excellent use and improvement of those [41, p. 8].

But this assumption misses the mark of the sort of entities we have under consideration. AMAs that currently exist are trained to do a very specific task, and they do that task very well. It is mistaken, however, to draw the conclusion that they may reason in the same sense that humans do.

This becomes all the more clear when we pause to consider what Aristotle understands reason to be. Within the *Ethics*, we find in 1139a that the rational soul breaks down into two components. The first part Aristotle calls the scientific part, which contemplates the unchangeable things³ and the calculating part which regards the changeable things perceived by the senses, and the contingent things of everyday life [40, 1139a].

For Aristotle, the distinction between these two aspects of the rational soul rests in their relationship to truth. For the scientific part, the goal is truth and falsehood as such. It examines first principles and objects that are necessary and eternal (such as axioms, mathematical truths or universals), or in other words, theoretical reasoning. This is in contrast to the calculating part: it examines truth and falsity in relationship to action and the desire for correct action, or put otherwise, practical reasoning.

From these two aspects of the rational soul, there are different virtues that relate to each and build upon virtues of character such as temperance, courage, justice, etc. In 1139b, we find a list with intellectual virtues (or put otherwise virtues of thought). These virtues are *techne* (skill or craft), *episteme* (scientific knowledge), *phronesis* (practical wisdom), *sophia* (wisdom), and *nous* (knowledge or intuitive understanding) [40, 1134b] [90]. Each of these relies upon different aspects of the soul depending on their use, and serves as a guide to determine the best goal of our actions.

This leads us back to our objection. Can AMAs really participate in all of these rational activities to the same degree? The answer to this question is provided by Berberich and Diepold themselves, and has been discussed in

³thus granting certain and scientific (in his understanding of the term) knowledge and would include things like mathematical truths

the second chapter of this work. Let us consider the following passage of their article:

Out of the three subcategories of machine learning, supervised learning, unsupervised learning and reinforcement learning (RL), the latter is the lifeworldly approach. In contrast to the other two, RL is based on dynamic interaction with the environment, of which the agent typically has only imperfect knowledge. Reinforcement learning divides the world into two parts, the agent and the environment. The agent takes actions based on its knowledge of the environment's current state. This action changes (updates) the environment's state and elicits a reward feedback which the agent receives to improve the policy with which it chooses actions based on environment states. This improvement is called learning.

Reinforcement learning could also be connected with the moral theory of utilitarianism, a variety of consequentialism, since both have the aim of maximizing estimated utility. However this combination is not intuitive since utilitarianism does not provide for a learning process, which is central to reinforcement learning.

Virtue ethics incorporates both, the value-maximizing policy and the focus on learning.

“The man who is without qualification good at deliberating is the man who is capable of aiming in accordance with calculation at the best for man of things attainable by action.” (NE 1141b 10)

This quote from Aristotle shows that a virtuous person is capable of deliberately choosing the action which promises the best outcome for man [41, pp. 8-9].

Here I agree that AMAs, and for our consideration driverless cars, operate using some form of machine learning, and in particular reinforcement learning

has been successful in the implementation of these devices. But this functionality of these machines does not mean that they have the capacity to reason in the full sense as meant in Aristotle and is integral to his conception of eudemonic virtue ethics. As a result of their programming, driverless cars operate (i.e. with probabilistic reasoning and through constant training) only with the aforementioned calculating soul, as they act upon sensible phenomena and changeable things with no further considerations of first principles, universals and the like.

As a result, they are incapable of possessing the scientific soul, and therefore do not have the virtues that relate to it. Rather, they only function and act within the sphere within which they have been trained. Following this line of thought, a driverless car that has been trained in the task of driving will have the intellectual and moral virtues related to that specific sort of activity, such as *techne* (the skill of driving), *phronesis* (practical wisdom of driving), which neatly fit into the sense-plan-act model upon which they operate. Their design, however, entails that they will not have the intellectual virtues of *episteme* (scientific knowledge). Provided that they only have a limited set of virtues, which is constrained by the sort of normative agent they are, we turn to a version of virtue ethics that focuses only upon specific virtues that some agent makes. This has been called a “target-centered virtue ethics” and will be the sort of ethics that we ascribe to driverless cars.

5.4.1 The good driverless car?

Considerations about the good life make little sense in regards to artificial moral agents, such as driverless cars, due to their limited faculties to reason and thereby their set of virtues and general incapacity to live a contemplative life. We therefore consider adopting an approach that focuses upon the acts of our normative agent, which may nevertheless arise for virtues that they do pose. This approach is called the “target-centered” virtue ethics, and has

been developed by Christine Swanton in her 2003 book entitled *Virtue Ethics a Pluralistic View* [42], and was further elaborated upon by Liezl van Zyl in her chapter *Right action and the targets of virtue* in “The Handbook of Virtue Ethics” [43]. This version of virtue ethics arose from difficulties that other non-eudemonic versions of virtue ethics contain. It is in the category of non-eudemonic virtue ethics that have cropped up over the last few decades and that were inspired by Elizabeth Anscombe’s article “Modern Moral Philosophy” [91], which revitalized contemporary interest in virtue ethics. Swanton frames her own theory as a response to a deficiency found within two prominent conceptions of the right action in agent-centered virtue ethics. The hallmark of these agent-centered virtue ethics is an evaluation of the agent taking the action in regards to various virtues. Swanton points to two dominant conceptions of right action that are found in modern literature concerning this agent-centered virtue ethics and are the qualified-agent view and motive-centered view.

The qualified-agent view has its exemplary expression found in the works of Rosalind Hursthouse, which she advocated for in her work, *Virtue Theory and Abortion*. Hursthouse’s conception begins by describing the rightness of an act in the following way: “an act is right if and only if it is what a virtuous agent would do in the circumstances” [42, p. 227], with a later modification to: “An act is right if and only if it is what a virtuous agent would characteristically (i.e. acting in character) do in the circumstances” [42, p. 228].

The second prominent view is found in the work of Michael Slote and focuses upon the motives of the agent. In this conception, the rightness of an action is agent-based and “according to which an action is right if and only if it exhibits or expresses a virtuous (admirable) motive, or at least does not exhibit or express a vicious (deplorable) motive” [42, p. 228].

Both of these notions are compelling, especially when we consider driverless cars. Using Hursthouse’s modified account, we can easily imagine that the car acts rightly in driving if and only if its actions are the sort of actions that a virtuous driver would characteristically do in the same circumstances. Or,

using Slote, the car's actions are right if and only if it exhibits or expresses a virtuous motive instilled in its programming. But despite this appeal for our purposes, Swanton finds problems with each.

In regards to Hursthouse's account, she points to problems that arise when we consider how rightness is determined by this qualified agent with a threshold view of virtue, which will be discussed below. Here there are both vertical and horizontal difficulties that Hursthouse's view faces. The horizontal problems arise when we consider that an agent may be virtuous in general, yet in particular fields of activity where they have no experience, they lack the practical wisdom needed to be virtuous in that activity. The vertical problem is that while an agent may be virtuous, there is always a more virtuous agent that our otherwise virtuous agent should differ to.

When considering driverless cars, Swanton's horizontal problem is rendered moot, as the agent's fields of competence are limited and it is not expected to operate outside of its field. The vertical problem, however, does challenge the application of Hursthouse's idea of rightness for these vehicles. Recalling the often repeated boons of these devices time and time again, we hear how they are supposed to surpass human drivers who are more prone to error and are less responsive than these machines. If a right action is determined by emulating other more virtuous drivers, and yet all other drivers have less "natural" capabilities than you, you run into a situation where the car cannot learn to behave virtuously as there are no teachers that are themselves competent to teach the driver. A solution to this is simply to emulate those drivers that are good enough, and so acquire a sub-adequate set of virtues related to driving, but such a solution would undermine their abilities to be better drivers than their natural counterpart, thereby undermining their appeal and in itself hindering a virtue of excellence.

Let us move on to Swanton's difficulty with Slote's account of right actions. Here she points to a problem he faces, known as the "bungling do-gooder" objection, where the agent has all the best motivations yet fails to act well.

While Swanton acknowledges that he tries to answer this objection with the following:

The well-motivated agent is concerned to determine facts: an agent genuinely desirous of being helpful is concerned that her help reaches its target, in a suitable way. To a reply that such an agent may not be aware of her ignorance, Slote would claim that a motive to help contaminated with intellectual arrogance is not an admirable motive. However, not all ignorance about one's expertise need be so contaminated [42, p. 230].

She furthers her critique by recalling W.D. Ross' distinction between the rightness and the goodness of an action.

This distinction rests upon the role that motivation plays in the accomplishment of some act. Drawing upon Ross, she gives an example of a man who needs to pay off his debt. There can be various reasons for paying off his debt; for example, he may wish to pay it off out of a sense of duty. If he pays it off because of that reason, then it can be said that he has done a good action as the action has a good motivation. However, it is entirely possible that the man only pays his debts because if he fails to do so, he will face legal consequences for that act. Nevertheless, as he has paid his debt, it can still be said that he did the right act [42, pp 230-231]. In this case, the motivation behind the action affects the moral tenor of the act, though the right act, that is paying off one's debts, is still obtained. Bearing this in mind, in Slote we see that the right action and morally good action are combined into the same act. This, however, doesn't allow for us to address situations where the act may be right, although the agent's internal dispositions don't reflect that. Swanton finishes her critique of Slote by relying upon Aristotle, who argues (along with her) that motivations are not the only internal state that affects the rightness or wrongness of acts, though they do have the capacity to change the tenor of the act, or in her words the deontic status, of the act, from right to wrong.

Having presented arguments against the other non-eudemonic virtue ethics, Swanton proposes her own account. This account begins with the distinction, found within Aristotle, between a virtuous act and an act from virtue, and is found within Aristotle thus:

A difficulty, however, may be raised as to how we can say that people must perform just actions if they are to become just, and temperate ones if they are to become temperate; because if they do what is just and temperate, they are just and temperate already, in the same way that if they use words or play music correctly they are already literate or musical. But surely this is not true even of the arts. It is possible to put a few words together correctly by accident, or at the prompting of another person; so the agent will only be literate if he does a literate act in a literate way, viz. in virtue of his own literacy. Nor, again, is there an analogy between the arts and the virtues. Works of art have their merit in themselves; so it is enough for them to be turned out with a certain quality of their own. But virtuous acts are not done in a just or temperate way merely because they have a certain quality, but only if the agent also acts in a certain state, viz. (1) if he knows what he is doing, (2) if he chooses it, and chooses it for its own sake, and (3) if he does it from a fixed and permanent disposition [42, pp. 231 - 232]. (Citing *Nicomachean Ethics* 1105a9-b2)

In this text, she teases out the question about how an act can be just or temperate if it does not exhibit a just or temperate state. She answers that the act can have those virtues if the act hits the target of justice or temperance within a given context⁴.

⁴Furthermore, it allows for the possibility of “moral luck” where an agent may happen to achieve a virtuous act despite his intentions or conversely allows for the agent to be absolved of fault if, for reasons beyond his control, fails to act virtuously despite his intentions.

How should we understand the target of a virtue? Swanton provides an explanation, which is succinctly captured by van Zyl in her chapter *Right action and the targets of virtue* in the “Handbook of Virtue Ethics” [43], and hinges upon her view of right action, which contains two central theses:

- P1: An action is virtuous in respect V if and only if it hits the target of virtue V.
- P2: An action is right if and only if it is overall virtuous [43, pg.119].

Van Zyl breaks these two features down into the following components *virtue*, *virtuous action*, *right action*, and *virtue from action*. She relates that Swanton defines *virtue* as “a good quality of character, more specifically a disposition to respond to, or acknowledge, items within its field in an excellent or good enough way” [43, p. 119]. The fields of some virtue “are things those ‘items that are the sphere(s) of concern of that virtue’ [43, p. 119].”

A *virtuous action* is a successful response in a particular situation. Van Zyl gives an example of a liberal person. Such a person embodies the virtue of generosity, and within the field of the virtue of generosity, we find things like gift giving, liberalness etc. And so this liberal person, when doing a virtuous act with respect to the virtue of generosity, will give the right amount of money to the right people at the right times. Provided that they do this, then it can be said that their acts are in fact virtuous.

Swanton leans upon Aristotle to establish how an act may be seen as virtuous. As a result, these sorts of acts are constituted by the fairly common notion of the mean: the action must not fall into either of its relevant virtues’ extreme (and thereby become vicious instance of the act) but fall in between these vicious extremes. So in regards to the liberal person, they will not give too much money, and by that means become a spendthrift, nor will they spend too little, and thereby become a miser. In addition to how they themselves spend money, they also must not spend too little, or too much, or just the

right amount on the wrong people; and moreover, they must not spend too little, or too much, or just the right amount at the wrong times. Or, for that matter, any other combination of actions that fall outside of the mean of that virtue of generosity [43, p. 120].

Given this example, van Zyl relates how a *virtuous action* is such an action that, in respect to some virtue V, hits the target of V. But that is not all. As Swanton herself explains, “[w]hat counts as hitting the target of a virtue is relatively easy to grasp when the aim of a virtue is simply to promote the good of individuals, and hitting that target is successfully promoting that good [42, p. 233].” But it can be complicated fairly easily as the particular response may require at least five different features, Which are the following:

1. There are several modes of moral response or acknowledgement appropriate to one kind of item in a virtue’s field, so hitting the target of a virtue may involve several modes of moral response.
2. The target of a virtue may be internal to the agent.
3. The target of a virtue may be plural.
4. What counts as the target of a virtue may depend on context.
5. The target of a virtue may be to avoid things [42, pp. 233 - 234].

She then addresses each one of these in turn.

In regards to the first feature, different virtuous actions may, and often do, require thinking about the various responses needed. So that when we consider virtues such as generosity, we need to be sensitive to the recipient of our actions. We may, for example, need to be generous to the extent that is needed due to our own relationship with the recipient of the targeted virtue – so I may be more generous with family and friends than to acquaintances or strangers. Even still, generosity towards people you do not know may garnish feelings of friendliness, or may be responsive to some need that the stranger

has that we ought to give due consideration. Other related modes of response may involve not expecting reciprocity in terms of gifts given, or respecting certain boundaries of gift giving or even not putting people off.

The second point recognizes that the targets of some virtues may be internal (either exclusively or in a mixed way). These internal targets are difficult to evaluate, and require a great degree of flexibility. In terms of generosity, part of the target is not only the act of giving, but may also include the motivations, mannerism, etc. so that a person who gives in a rude way may not be hitting the target of generosity. Yet they may be if such a rude mannerism is somehow within the field of the virtues of social mores.

The plurality of targets acknowledges that some virtues may have many targets that they aim at. This is a result of them having more than one field. Virtues such as courage aim at controlling fear, but it may also aim at managing dangerous situations in a successful way. These targets may also be plural in regards to them being both internal and external.

The contextual dependency of the target of a virtue helps us determine the target(s) of a virtue in a concrete circumstance. The example Swanton gives of this is the following:

What counts as a virtuous act is more heavily contextual than what counts as an action from virtue. In some contexts, for example, where there is considerable need, one may be said to have performed a generous act if one donates a large amount of money, say, even if that donation is made with bad grace. However, in other contexts, we may deny that an act of giving is generous on the grounds that it was not made in a generous spirit. Here the target of generosity is to alleviate need, in the right way, where ‘in the right way’ makes reference to manner of giving, and even motivation [42, p. 236].

Similarly a parallel to this can be seen in the well-known parable of the Widow's Offering as found in Mark 12:41-44:

Jesus sat down opposite the place where the offerings were put and watched the crowd putting their money into the temple treasury. Many rich people threw in large amounts. But a poor widow came and put in two very small copper coins, worth only a few cents. Calling his disciples to him, Jesus said, "Truly I tell you, this poor widow has put more into the treasury than all the others. They all gave out of their wealth; but she, out of her poverty, put in everything—all she had to live on."⁵

As we can see, being generous depends not only upon giving itself but also upon a host of other factors, including how one makes the donation and the amount of wealth one has in making the donation (so that a few coins given from a billionaire is seen as being miserly whereas the same amount from a person who can hardly put food on the table is considered to be laudable). But these need not be the only factors, for as Swanton rightly points out, the generosity of the rich may still be generous and help many people in times of great need even if they toot their own horn while making the donation.

In Swanton's view, there are certain virtues whose target is to simply avoid certain things. The example she gives is modesty. In this virtue, being modest involves the agent in question to avoid drawing attention to themselves. This can be done by not excessively boasting, or talking about themselves ad nauseam, and this sort of behavior constitutes the target of modesty [42, pp. 237-238]. Other examples may include the target of prudence, where someone with a gambling addiction would hit the target of prudence by not going to a casino, or a person battling their concupiscence hits their target by abstaining from various nefarious nightclubs.

⁵Taken from the New International Version.

The above features allow Swanton to distinguish between an action from virtue and a virtuous act. The key distinction is that actions from a virtuous state require far more to be true than a virtuous action. To begin, an act from virtue may fail to hit its target due to some lack of knowledge that the agent possesses. Secondly, because for some act to be from a state of virtue, all modes within the virtue's field must be exhibited, while such stringent requirements are not needed for an act that hits the target of a virtue. Additionally, actions from virtue require that they be "displayed in an excellent way, in a way which expresses fine inner states [42, p. 238]" of the agent, in addition to expressing all modes of moral acknowledgement, resulting in cases where an act may be virtuous but it fails to be an act from virtue as it falls short in that regard. Finally, an action from virtue is less contextual than a virtuous act, for a virtuous act relies on the particular circumstances in question and the successful hitting of the target of the virtue in question, whereas an act from virtue requires the agent to be merely (or perhaps not so merely) a paragon of the virtue in question[42, pp. 238-239].

The second component of Swanton's theory for right action concerns itself with the overall virtuousness of the act. The account begins with an admittedly standard (by both her own and van Zyl's accounts) distinction between right actions and good actions, which itself is based upon the works of W.D. Ross. The view is surmised in the following passage from Ross: "[m]oral goodness is quite distinct from and independent rightness, which ... belongs to act not in virtue of the motives the proceed from, but in virtue of the nature of what is done[43, pg 121]". As we can see, the distinction between these good acts and right acts fits nicely with virtuous acts and acts from virtue, where a virtuous action may be said to be right, while an act from virtue is good. Ross admits that this distinction may seem artificial, especially with the common usage of good and right being synonymous; however, it allows us to capture the common sense view that there is a difference between an act in itself and the motives behind it.

Van Zyl explains a crucial aspect where Swanton differs from this classic distinction between right acts and good ones as found within Ross. The agent's motives may, in some circumstances, affect the rightness of the act. These circumstances are instances where (some of) the targets of some virtue *V* are internal, and the actions themselves are internal acts. The provided examples of this are the following:

Consider the case of a politician who is in charge of a public health campaign which can have a significant impact on many people. In this context the targets of the relevant virtues – beneficence and efficiency – are external, so that a selfish motive (such as a desire for status or money) does not affect the rightness of the action. But contrast this to a more intimate context, such as the role of a private nurse in charge of caring for a terminally ill patient. Consider the case of Nurse John, who takes excellent care of his patient, but secretly cannot stand her; were it not for the fact that she is rich and pays well, he would not try so hard to please her. In so far as he successfully promotes human welfare and displays the appropriate kind of behaviour and demeanour, John hits the targets of caring. However, he misses an important (internal) target of the virtue of care, which is to have genuine concern for another. Thus, it is not merely that John does what is right from an inferior motive. Rather, he fails to act rightly because he misses an important target of caring. Compare this to the case of Nurse Tessa, who takes excellent care of her patient, genuinely cares about her, but who makes a mistake that causes the patient great pain and discomfort (perhaps she accidentally administers the wrong dosage of a certain medication). Tessa fails to act rightly, despite her good motives, for she misses one of the targets of caring (namely, promoting human welfare) [43, p. 123].

As we can see, the target of the virtue has a direct bearing upon whether or not one's motivations affect the rightness or goodness of an action, in addition to the success of the act in hitting its target virtue.

In the case of the politician, we can see that Ross' traditional distinction between right and good holds, as his motivations for fame and wealth do not affect the outcomes as such. So a well-formed health program that still benefits many people can be said to be right, though his work on the program cannot be said to be a good act given his motivations. In the case of the nurses, where the relevant virtue of their profession is caring, their motivation do, in fact, seem to have bearing upon the rightness of the act. These situations are rightly noted to have various targets, some of which are external, while others are internal. The nurses' internal dispositions do seem to have a bearing upon the nurses' duty to care for their respective patients. Swanton's extension of this distinction does seem to be helpful in these sorts of situations.

The multitude of targets lead Swanton and van Zyl to the consideration of how to best handle these mixed situations. Swanton proposes three methods that one may adopt in order to build a theory of right action that can hit the targets of virtue. These methods are maximalistic, permissivistic and minimalistic, and rest upon her threshold concept of virtue, which we alluded to previously in our discussion of her critiques of Hursthouse and Slote.

To begin, her threshold concept of virtue relates to the previously discussed notion of right act. Here a right act can be evaluated in one of two way. The first way is that the action is either right or wrong, and that evaluation is binary, with no degrees of rightness or wrongness between. The second means of evaluating an action still aims at trying to determine whether or not the act is/was right or wrong, but rightness is a threshold concept with vague boundaries in between it and wrongness.

In this vague account, the spectrum of rightness ranges from "perfect" to "good enough", where "good enough" is contextually dependent. So that "in a

world characterized by considerable evil, neediness, and frequent catastrophe, less than ideal states may count as virtuous” [42, p. 25], although it should be noted that all “admirable, useful, and praiseworthy states of the agent are to count as virtuous, for one may wish to distinguish virtue from other states which in certain contexts are useful and praiseworthy, such as self-control” [42, p. 25]. Van Zyl unpacks this for us and describes Swanton’s theory of a threshold account of virtues and right acts as having three categories of “right actions, “all right” actions, and wrong actions – but argues that these categories do not have sharp boundaries [43, p. 124].”

This vague account of virtue maps nicely to Swanton’s methods of hitting the targets of virtues. The first option (1) is the maximalist approach and uses the following understanding of determining a right act:

(1) An act is right if and only if it is overall virtuous, and that entails that it is the, or a, best action possible in the circumstances. *Assuming that no other virtues or vices are involved*, we could say that a given act is right insofar as it is the most generous possible. The target of generosity on this view is very stringent: there is no large penumbra such that any act which falls within it is deemed right [42, p. 239].

The second (2) option is the permissivist approach and uses the following understanding of determining a right act:

(2) An act is right if and only if it is overall virtuous, and that entails that it is good enough even if not the (or a) best action. Here it is assumed that there is much latitude in hitting the target of virtues such as generosity. Right acts range from the truly splendid and admirable to acts which are ‘all right’ [42, p. 240].

The final option (3) is the minimalist approach and uses the following understanding of determining a right act:

(3) An act is right if and only if it is not overall vicious. Here it is assumed that not being overall vicious does not entail being overall virtuous. An act may avoid the vices of meanness or stinginess, for example, without hitting the target of generosity, which demands more than mere avoidance of stingy, mean acts. This may be true even if the target of generosity is interpreted as in (2), rather than (1) [42, p. 240].

Swanton herself excludes the third option and prefers the first option to the second, where “[p]rovided a distinction is made between rightness and praiseworthiness, and between wrongness and blameworthiness, it seems natural to think of the targets of a virtue as best acts (relative to the virtue), though it does not follow that a rational agent should always aim at such a target directly, or should necessarily deliberate about reaching that target [42, p. 240].”

From this we are able to handle mixed situations where a tension is present between various virtues and their targets. Van Zyl takes up this discussion in regards to the famous case of Jim and the Indians, told by Bernard Williams in his 1973 article, *A Critique of Utilitarianism*. In this story:

Jim finds himself in the central square of a small South American town. Tied up against the wall are a row of twenty Indians, most terrified, a few defiant, in front of them several armed men in uniform. A heavy man in a sweat-stained khaki shirt turns out to be the captain in charge and, after a good deal of questioning of Jim which establishes that he got there by accident while on a botanical expedition, explains that the Indians are a random group of the inhabitants who, after recent acts of protest against the government, are just about to be killed to remind other possible protesters of the advantages of not protesting. However, since Jim is an honoured visitor from another land, the captain is happy to offer him

a guest's privilege of killing one of the Indians himself. If Jim accepts, then as a special mark of the occasion, the other Indians will be let off. Of course, if Jim refuses, then there is no special occasion, and Pedro here will do what he was about to do when Jim arrived, and kill them all. Jim, with some desperate recollection of schoolboy fiction, wonders whether if he got hold of a gun, he could hold the captain, Pedro and the rest of the soldiers to threat, but it is quite clear from the set-up that nothing of that kind is going to work: any attempt at that sort of thing will mean that all the Indians will be killed, and himself. The men against the wall, and the other villagers, understand the situation, and are obviously begging him to accept. What should he do [92, pp.98-99]?

In this story, van Zyl finds that there are multiple virtues that come into play which includes courage, justice, and wisdom, but for her considerations, she chooses to focus upon benevolence and non-malevolence as being central to Williams' story.

How can Jim hit the targets of these two virtues? Van Zyl argues that Swanton's theory allows for this in both cases despite being in a dilemma situation. If we assume that Jim gives in to Pedro, or listens to the pleas of the villagers and slays one of their number, it would seem hard to say that he is being benevolent and non-malevolent, and in general the killing of an innocent human being would fail to hit the targets of these virtues. Yet, as noted above, Swanton's theory states that what counts as hitting the target of virtues is heavily contextually dependent. Van Zyl offers the following explanations for how, even in a situation such as Jim's, an agent may act virtuously. In terms of hitting the target of non-malevolence, we should consider the internal targets of this virtue. Here, provided that Jim's "demeanour, motivation and thought processes are not cruel or malicious then he does not act wrongly, even if he causes the death of an innocent person" [43, p. 125].

To hit the target of benevolence, we must consider the particulars of the case. Here, for Jim to be benevolent, he would want to save all twenty of the villagers, but he wisely surmises that such an attempt would lead not only to all of the villagers' deaths but to his as well. So given that saving all of the lives is not contextually possible, the target of benevolence shifts to saving nineteen lives. Provided Jim hits the targets of non-malevolence and can hit the target of benevolence by saving the lives of the nineteen villagers, despite slaying one villager himself as the beneficiary of this special guest's privilege, we can still say that Jim acted rightly, although he need not necessarily take this course of action.

In our considerations of Jim's situation, we note that we are able to determine that the option of slaying an individual is the right action given the more stringent definition that Swanton prefers. That is to say that it, the slaying of one person rather than twenty or twenty-one, is the action that hits the targets of the relevant and is the best possible action in the situation at hand. What is notable, however, is that Swanton and her commentator van Zyl acknowledge that despite it being a (uniquely) right action, it is not necessary that the agent act in this way in the sense that deontologists and consequentialists understand it. This is because one hallmark distinction between virtue ethics and the other prominent schools of ethical thought is the latter's tendency to view right action as that which ought to be done or may be done while the former makes a distinction between right action in both an action assessing sense and an action guiding sense [43, p. 127].

Van Zyl explains that Swanton's use of this distinction hinges upon her understanding of right action, and the acknowledgement that real agents are not ideal agents with perfect knowledge and may not have sole agency in particular situations. Furthermore, they acknowledge that while some right actions are obligatory, others, such as self sacrifice, while laudable, are too demanding to be obligatory, but are rather supererogatory [43, p. 127] [42, pp. 240 - 241].

An agent posed with a problem takes the relevant data that they have and tries to navigate their way through it. The example given is that of a politician who is drafting a policy on the use of genetically modified foods where different people may have different preferences on which virtue trumps other virtues. One take is that it is reasonable that “people in the face of ignorance should guard against such possible dangers” [43, p. 128], and so a politician promoting a policy of tight regulation hits the targets of prudence. However, while that certainly is a reasonable, right action, it may not be necessarily the only right action. It is also possible for that very same politician to attempt to hit the target of benevolence, by means of allowing easy production of more plentiful and cheaper food, which is resistant to disease. While this may prove to be unreasonable, given the lack of knowledge of the potential risks, lax regulation may be right in some situations such as hyperinflation, famine or pestilence, where the politician may recall the words attributed to Cpt. David Glasgow Farragut during the battle of Mobile Bay and say “Damn the torpedoes, full speed ahead!” and allow the use of these new food sources.

Swanton’s distinction rests upon the difference between a right action and a reasonable action, as explained by van Zyl, where a reasonable action helps us determine the “degrees of rightness and wrongness, and there can be reasonable and pervasive disagreement about the rightness or wrongness of a particular action” [43, p. 128] and the reasonable action – such as killing one Indian to save twenty or exposing your citizens to unknown risks– is “that [which] ought to be done in the sense of what is commended rather than obligatory” [43, p. 128].

5.4.2 The Target-Centered Virtues of a Driverless Cars:

What does this mean for driverless cars? Let us recall my “Turning” (5.1) example, where a driverless car makes a turn and is presented with the three options of turning left and hitting a workman on his lunch break, going straight

and hitting the five workmen working on their shift or turning right and hitting the wall risking injury to itself and its occupant. This turning example is not so dissimilar from the story of Jim and the Indians nor is it too different from the politician considering the use of genetically modified food in his country. Here the car finds itself in a dilemma situation where it is not possible for it to act perfectly, so instead it must act in the best possible way by choosing the “best possible action” given the circumstances. But what would be the best possible action here?

To answer this, I will first propose a list of principle virtues that a virtuous autonomous car should have. This list is by no means exhaustive but serves as a springboard for future works and is changeable as the agent itself becomes more technologically advanced and thereby capable of doing more and more things. Second, I will apply the relevant virtues from this list to our present example, to decide the right and reasonable actions that the driverless car may have recourse to given the situation at hand.

Candace Upton, in her chapter *What virtues are there?* in the “Handbook of Virtue Ethics” [43] provides the following list of uncontroversial moral virtues, which includes honesty, generosity, courage, and justice. Additionally she provides other virtues, taken from Aristotle’s *Ethics*, that often count as moral virtues as well, and this includes virtues such as “pride, good temper, truthfulness, temperance, benevolence, generosity, friendliness and ready wit” [43, p. 165]. Of these virtues, the virtues that a driverless car may and should exhibit are the following, justice and benevolence, both of which we will draw upon Aristotle and Swanton to help clarify.

Book V of Aristotle’s *Ethics* is dedicated to the discussion of the topic of the virtue of justice, and at its core deals with the just relations. The targets of justice are both internal and external and the bearer of this virtue will exhibit both when needed. Aristotle recognizes that there is a universal justice that chiefly roots this virtue in acting lawfully, and what is lawful is rooted in the

promotion of the common good; and its converse, injustice, is understood as acting contrary to the law. The example that Aristotle gives is the following:

Again, someone who commits adultery for gain and makes money out of it would seem unjust, but not intemperate, while another who does so through appetite, though it costs him and he loses money for it, would seem to be intemperate rather than greedy. Obviously, this is because the first acts for gain.

Again, all other unjust acts are always attributed to some form of wickedness, such as adultery to intemperance, desertion of a comrade in battle to cowardice, physical assault to anger. But if the person gains by what he does, it is attributed to no other form of wickedness than injustice [40, 1130a].

In addition to this universal justice, which concerns itself with the law, there is also a particular justice that is aimed at the fair distribution of divisible goods such as honor, money, and other things that can be shared by the political community [40, 1131a].

In regards to driverless cars, we need to address the field, mode and target of justice. The field of this virtue primarily (though not necessarily) relates to following the universal law and the various norms that relate to the car itself as a normative agent. The patients of the car's duties of justice include itself, its passengers, other road occupants, and the state (or some other normative authority) within which it is driving. The modes of response to this universal justice depend upon the particular patient in question and so will depend upon the particular context of some virtuous act. The hitting of the targets of this virtues also is contextually dependent upon the laws related to the situation at hand. At any given time the driverless car, being the bearer of the various norms related to driving, owes something to the state – such as the obligation to obey the speed limit or to stop for pedestrians and other laws inscribed in the criminal code and statutes as described in the previous chapters– or

to private individuals – in matters related to civil law. Successfully hitting these targets depends extensively upon the particular circumstances of the situation and balancing it with other virtues that may also be at play. Finally, an act is deemed to be right when it makes the best possible action given the circumstances, although, as discussed above, it need not be obligatory to take.

Benevolence aims at the promotion of what is good. What counts as a benevolent act, much like what is a just act, depends upon the context within which the act is undertaken. The field of this virtue relates to the various goods that exist. Swanton acknowledges that at its base utilitarian-esque form, it is simply the promotion of the good of others [42, p. 23], so that their state of affairs are better off after the action than they were before. This basic understanding, however can be more fine-grained and allow for more nuance. Here Swanton links benevolence to the notion of love for other, and the examples that she provides include special considerations for the particular love that a parent has for their child or children to even a general concern for the benefit of humanity or for others in general, which reflect a narrow and broader fields respectively [42, p. 103]. Crucial to this understanding of love and benevolence towards others is the Kantian notion of not treating others as mere means to one's own ends, conjoined with a recognition of the moral value of other humans as entities that are ends in themselves [42, p. 107]. To this universal understanding of benevolence, particular considerations can be added due to particular circumstances. So a nurse has a special duty of care for their patient, a guardian to their ward, a parent to their child.

The proper modes of response, and likewise hitting the targets of the virtue of benevolence, depend upon the situation at hand and may be internal or external and relevant to the agent and the patient of the act. Furthermore, successfully hitting the target of the act and the ability to deem the act as being right also depend upon the situation that presents itself. Considering our principle agent of a driverless car, the targets of benevolence would be related to its sphere of activity, i.e. driving, and the patients of these acts can

include is occupants, other road users, pedestrians, etc. where the autonomous vehicle has the appropriate duty of care relevant to its relationship with the patient in question.

Benevolence in this form has been discussed in the 2010 work of Hidi Li Feldman, *Prudence, Benevolence, and Negligence: Virtue Ethics and Tort Law* [93], where a particular duty to care, resulting from an agent's particular normative status, can be seen elsewhere in the law. For example in the fiduciary relationship discussed in chapter 3, the agent has a particular duty of care towards their principle in making sound financial decisions on their behalf. Or alternatively this can be seen in family law, where in custody cases, the good of the child has chief precedence rather than the good of the interested parents [93, pp. 1439 - 1440]. I would also add that in a similar way an employee has a duty to promote the good of their employer and vice versa, which results from the special relationship that they have formed between themselves, that may go beyond the duties adumbrated within their work contract. By the same token, a driverless car would have a special duty to promote the good of its keeper/owner in addition to the good of its users and other entities it interacts with to the best of its capabilities and relevant to its function as driver.

The use of these target-centered virtues of benevolence and justice can help in the training of virtuous autonomous moral agents that were discussed in our presentation of the work of Berbrich and Diepold [41]. Driverless cars being presently trained are done so by means of machine learning techniques, where wrong acts are “punished” and “right”⁶ acts are rewarded, allowing for the machine to learn through habituation, and through trial and error. Moral virtues are likewise gained through trial and error and habituation. The more a normative agent acts and successfully hits the targets of virtues, the easier it is for them to successfully hit similar targets of that virtue in the future.

A mundane example of driverless cars hitting the targets of the virtue of justice includes its learning to drive in a legal way. These devices have come a

⁶or correct acts

long way since the Stanford Cart, and have logged millions of miles of driving experience, to the point where it has now been able to outperform human beings in an expanding list of typical driving situations. This topic has been previously discussed in length in the second chapter of this work. The very training of driverless cars imbues them with various virtues that makes them moral machines. Or put otherwise, the training of a car to drive legally trains the car to be virtuous relative to the virtue whose target is hit by driving legally, which is justice in its universal sense.

The evaluation of the car's acts may be deemed to be right provided that the action in question "if and only if it is overall virtuous, and that entails that it is the, or a, best action possible in the circumstances [42, p. 239]." This evaluation of the driverless car's action as being right serves both as a tick of approval for previous actions and a guide for future acts having similar factual content, being obligatory in the sense that van Zyl and Swanton describe. Furthermore, the lack of necessity of the right action needing to be taken allows for the permissibility of the machine to act otherwise, which it may choose to do so as a result of its probabilistic reasoning.

This "do and become virtuous" approach is beneficial especially considering the nature of the moral agents we have in question that learn extensively through training. However in situations like the turning example, it would seem best to not learn by intentionally placing the car in such a situation, and would arguably be a failure on the trainers to hit the target of benevolence on their part. To remedy this, we may use virtual simulations like MIT's Moral Machine among other similar thought experiments to build a data base from which the machine may draw upon to make future right actions if it is presented with a similar situation.

To conclude, we can now address the "Turning" example I have been using throughout this chapter. In our situation the driverless car is presented with three alternatives, none of which can be said to be good, as no option is morally perfect. In such a situation, we must settle for that which is the right action.

Given the virtues that the car has of justice and benevolence, we must see how it can best hit the targets of these virtues. In regards to justice, it must act legally. To do so, it must not hit and injure a person, much less five persons, nor should it injure its occupant. Each of these claims upon the car arise from each individual having an interest in not being hit by the car. In this particular situation, it is impossible for it to avoid all three prohibitions and satisfy the claims made upon the driverless car, and so it must do the next best action that it can. From this we exclude hitting five, and as a result we are left with either hitting the workman on his lunch break or the occupant in the car. Either of these options would minimize the hitting done and allow the car to hit the target of justice given the situation. As it stands now, either option may be right, and other factors can certainly come into play. Some brute facts may be taken into consideration, such as the safety rating of the vehicle or safety barriers around the workman may affect the expected harm to be done, and may hedge us one way or the other as being the right action in the particular circumstances. Other normative facts, such as a workmen signing away rights to safety or trumps to other persons safety over the occupants to the car, may also change the balance of things. But assuming the car doesn't have access to that information, we can conclude that either option is permissible, and is right in this unique case.

The inclusion of benevolence may also help us determine who the car should select but need not necessarily do so. In much the same way as Jim was benevolent in slaying one Indian for the sake of the other nineteen (provided his internal dispositions were correct), the injury to the workman or occupant of the car may be benevolent inasmuch as the hitting of the target of benevolence is not causing harm to the other six people and not harming all seven is impossible. One way of tipping the scales in this situation could be the option of the occupant pre-selecting their preference on how the car may hit the target of benevolence. The car has a special relationship to its owner/keeper and occupants due to its bearing the role of driver and so it has a greater duty of

care towards them and so it may default to promoting their good to another's good, all other things being equal. However this can be turned the other way around as the owner/keeper may pre-set a preference towards others in a way discussed previously in the work of Contissa et al.'s Ethical Knob [39].

As a result, we are able to evaluate the options and exclude some outright, such as hitting the five workmen, but account for different acceptable outcomes that reasonable people may have, using the virtues we have suggested that driverless cars should have.

5.5 Concluding Remarks:

In this chapter, we examined the shortcoming of the application of the schools of thought discussed in the previous chapter and sought out an alternative theory. This theory was virtue ethics, and specifically we address a target-centered theory. Here we applied this theory to our agent which doesn't have the full breadth of virtues that typical agents (i.e. humans) have but only a limited set relative to their capacities.

Chapter 6

Conclusion

In this dissertation, we have set out to answer the question of how norms can be applied to driverless cars. It began with a challenge to the traditional notion that technology is value neutral. This arose from the reflection of how technological systems appear to no longer be mere tools but are moving towards acquiring an agency of their own. Combined with their increasing prevalence in our world – by taking on the day to day tasks of care giving, surgery and driving – the question of how these new devices ought to be controlled becomes more and more apparent. Yet to recognize the need for this is one thing, to justify artificial agents as being the bearers of norms is quite another. In order to establish this, we raised three topics that needed to be addressed, to which we dedicated a chapter each.

6.1 Outcomes

The first topic was about the current state of affairs of the technology which we address in Chapter 2. There we examined both the development of this technology and its current status. We also took a survey of the legal landscape to see what, if any, rules applied to these devices. By doing this, we were able to establish the sort of entities that we have under consideration.

The second topic was about whether or not these entities are agents, and if so, whether they are normative agents. This was examined in Chapter 3, where we concluded that by standard accounts of agency within computer science, they are, in fact, agents. From here we then took up the task of seeing if they are capable of bearing normative agency. To argue this point, we looked at prevalent theories of rights, such as Hohfeld's theory of right relations and its intersection with will and interest theory as presented by Mathew Kramer, and saw if they could be applied to our agent, the justification for which went hand in hand with an examination of the conception of legal personality, especially as a concept being able to accommodate for non-human normative agency. From here we also took up the application of legal norms within driverless cars as a natural extension of this line of reasoning.

Our consideration of ethics, the third task undertaken here, was addressed in Chapter 4. In that chapter we examined ethics, especially in terms of morality, both broadly and how it is typically applied specifically to the entities we have in question. This began with a discussion on the "Trolley Problem", which is a central feature of this research topic. From there we looked at how ethics is understood in popular and scientific literature. In regards to the scientific literature, we gave a synopsis of both deontological and consequentialist ethics, as they are the two main ethical schools that are considered within the literature, and then their application to topic-specific works. From there, we continued our discussion in chapter 5 where we addressed how both of these schools are inadequate to properly address ethics in driverless cars. This led us to consider a third option rarely taken up in robot ethics, that is the use of virtue ethics, and we argued for its application in driverless cars.

The argumentation for this stemmed from both considering the sort of agents that driverless cars are and how an ethics can be implemented within these new devices. The way that these devices are programmed to operate relies extensively upon machine learning tools, rather than hard coded "if then" rules. Because of their sense-act-plan structure, an ethics that allows for a "do

and become ethical”, such as some strand of virtue ethics, is appealing. Despite this, however, a one-to-one mapping of a virtue ethics for humans to one for a driverless car is not merited. Rather, it needs to be specifically tailored to these devices and limited in accordance with their limitations. Such limitations include only having ethical virtues, thereby excluding the intellectual virtues until warranted by the device’s capabilities. Moreover, even within the set of ethical virtues, we only took under consideration two virtues – justice and benevolence. In regards to the application of these virtues, we argued for the use of Christine Swanton’s target-centered virtue ethics as being well suited for use within driverless cars.

6.2 Further research

While we have argued for the normative agency of driverless cars and have proposed the use of a target-centered virtue ethics to help deliberate and evaluate actions, and thereby determine if these actions were good or right or even wrong, further research needs to be done to directly implement this ethical framework within these devices. For instance, this could involve writing a program to train the car to hit the targets of virtues that it is presented with. This can already be seen in training these cars to drive in a legal fashion, where driving legally hits the target of (universal) justice. Similarly mixed situations could be introduced like the “Turning” example I provided in 5.1, where the agent is unable to choose a good act, but rather only a right act. Further research should also be done in flushing out more specific virtues for driverless cars. This list of virtues needs to be adaptable to the ever changing nature of these devices, and could include virtues such as loyalty, wit, and courage to name some potential candidates.

Bibliography

- [1] National Highway Traffic Safety Administration U.S. Department of Transportation. *Federal Automated Vehicle Policy Accelerating the Next Revolution in Road Safety*. 2016.
- [2] European Parliament Press Room. Robots: Legal affairs committee calls for eu-wide rules. Press Release, Jan 2017. URL <http://www.europarl.europa.eu/news/en/press-room/20170110IPR57613/robots-legal-affairs-committee-calls-for-eu-wide-rules>.
- [3] European Commission. The report of the high level group on the competitiveness and sustainable growth of the automotive industry in the european union final report - 2017, October 2017.
- [4] Federal Ministry of Transport and Digital Infrastructure. Ethics commission: Automated and connected driving, June 2017.
- [5] Karlyn D. Stanley Paul Sorensen Constantine Samaras Oluwatobi A. Oluwatola James M. Anderson, Nidhi Kalra. *Autonomous Vehicle Technology A Guide for Policy Makers*. RAND Corporation, 2016.
- [6] J. Christian Gerdes Hermann Winner Markus Maurer, Babara Lenz, editor. *Autonomous Driving: Technical, Legal, and Social Aspects*. Springer, 2015.
- [7] Hod Lipson and Melba Kurman. *Driverless: Intelligent Cars and the Road Ahead (MIT Press)*. The MIT Press, 2016.

- [8] Javier Ibañez-Guzmán, Christian Laugier, John-David Yoder, and Sebastian Thrun. Autonomous driving: Context and state-of-the-art. In *Handbook of Intelligent Vehicles*. Springer London, 2012. doi: 10.1007/978-0-85729-085-4_50.
- [9] Declaration of amsterdam cooperation in the field of connected and automated driving, April 2016. URL <https://www.regjeringen.no/contentassets/ba7ab6e2a0e14e39baa77f5b76f59d14/2016-04-08-declaration-of-amsterdam---final1400661.pdf>.
- [10] Tatjana Evas. A common eu approach to liability rules and insurance for connected and autonomous vehicles, european added value assessment accompanying the european parliament’s legislative own-initiative report (rapporteur: Mady delvaux), February 2018. URL <http://www.europarl.europa.eu/thinktank/en/home.html>.
- [11] Council of the European Union. Council directive 85/374/eec of 25 july 1985 on the approximation of the laws, regulations and administrative provisions of the member states concerning liability for defective products, July 1985. URL <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:31985L0374>.
- [12] Council of the European Union European Parliament. Directive 2009/103/ec of the european parliament and of the council of 16 september 2009 relating to insurance against civil liability in respect of the use of motor vehicles, and the enforcement of the obligation to insure against such liability (text with eea relevance), September 2009. URL <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32009L0103>.
- [13] John Thune. American vision for safer transportation through advancement of revolutionary technologies act, 2018. URL <https://www.congress.gov/bill/115th-congress/senate-bill/1885/text>.
- [14] Singapore Government. Road traffic act, May 2018. URL <https://sso.agc.gov.sg/Act/RTA1961#pr6C->.

- [15] Convention on road traffic, done at vienna 8 november 1968. Convention, nov 1968. URL http://www.unece.org/fileadmin/DAM/trans/conventn/Conv_road_traffic_EN.pdf.
- [16] Convention on road traffic, signed at geneva 19 september 19 1949, September 1949. URL http://www.unece.org/fileadmin/DAM/trans/conventn/Convention_on_Road_Traffic_of_1949.pdf.
- [17] Laurence F. White Samar Chopra. *A legal Theory for Autonomous Artificial Agents*. University of Michigan Press, 2011.
- [18] Robert Trypuz. *Formal Ontology of Action*. Elpil, 2008.
- [19] Luciano Floridi. *The Ethics of Information*. PAPERBACKSHOP UK IMPORT, 2013.
- [20] Ota Weinberger (auth.). *Law, Institution and Legal Politics: Fundamental Problems of Legal Theory and Social Philosophy*. Law and Philosophy Library 14. Springer Netherlands, 1 edition, 1991.
- [21] Hans Kelsen. *Pure Theory of Law*. The Lawbook Exchange, 2005.
- [22] Wesley Hohfeld. Some fundamental legal conceptions as applied in judicial reasoning. *Yale Law Journal*, 23, 1913. URL <http://digitalcommons.law.yale.edu/ylj/vol23/iss1/4>.
- [23] Matthew Kramer, N. E. Simmonds, and Hillel Steiner. *A Debate Over Rights: Philosophical Enquiries*. Oxford University Press, 2000. ISBN 0-19-829899-4.
- [24] H. L. A. Hart. Are there any natural rights? *The Philosophical Review*, 64(2):175, apr 1955. doi: 10.2307/2182586.
- [25] Leif Wenar. The nature of claim-rights. *Ethics*, 123(2):202–229, 2013. ISSN 00141704, 1539297X.

- [26] Neil MacCormick. Norms, institutions, and institutional facts. *Law and Philosophy*, 17, 05 1998. doi: 10.1023/a:1006034203352.
- [27] Aleardo Zanghellini. Raz on rights: Human rights, fundamental rights, and balancing. *Ratio Juris*, 30(1):25–40, 2017. doi: 10.1111/raju.12156.
- [28] Gordon Campbell. *A Compendium of Roman Law Founded on the Institutes of Justinian*. The Lawbook Exchange, Ltd., 2008.
- [29] Ugo Pagallo. *The Laws of Robots Crimes, Contracts, and Torts*, volume 10 of *Law, Governance, and Technology Series*. Springer, 2013.
- [30] C. van Dam. *European Tort Law*. Oxford University Press, 2013.
- [31] Philippa Foot. The problem of abortion and the doctrine of double effect. *Oxford Review*, 5:5–15, 1967.
- [32] Judith Jarvis Thomson and. Killing, letting die, and the trolley problem. *Monist*, 59(2):204–217, 1976. doi: 10.5840/monist197659224.
- [33] Patrick Lin. The ethics of autonomous cars: Sometimes good judgment can compel us to act illegally. should a self-driving vehicle get to make that same decision? *The Atlantic*, 2013. URL <https://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/>.
- [34] Joseph Migga Kizza. *Ethical and Social Issues in the Information Age*. Springer International Publishing, 2017. doi: 10.1007/978-3-319-70712-9.
- [35] Walter Sinnott-Armstrong. Consequentialism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2015 edition, 2015.
- [36] Larry Alexander and Michael Moore. Deontological ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

- [37] Patrick Lin. Why ethics matters for autonomous cars. In *Autonomous Fahren*, pages 69–85. Springer Berlin Heidelberg, 2015. doi: 10.1007/978-3-662-45854-9_4.
- [38] Neil McBride. The ethics of driverless cars. *SIGCAS Computers & Society*, 2015.
- [39] Giuseppe Contissa, Francesca Lagioia, and Giovanni Sartor. The ethical knob: ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law*, 25(3):365–378, sep 2017. doi: 10.1007/s10506-017-9211-z.
- [40] Roger Crisp Aristotle. *Nicomachean Ethics*. Cambridge University Press, 2004.
- [41] Klaus Diepold Nicolas Berberich. The virtuous machine - old ethics for new technology? 2018. URL <https://arxiv.org/pdf/1806.10322.pdf>.
- [42] C. Swanton and Oxford University Press. *Virtue Ethics: A Pluralistic View*. Oxford University Press, 2003.
- [43] Stan van Hooft. *The Handbook of Virtue Ethics*.
- [44] Us pattent 2,519,859 for speed control device for resisting operation of the accelerator.
- [45] The State of California. Title 13 of the california code of regulations, art 227.00-227.52, feb 2018.
- [46] Bryant Walker Smith. Sae levels of driving automation, December 2013. URL <http://cyberlaw.stanford.edu/blog/2013/12/sae-levels-driving-automation>.
- [47] *Deontic Logic and Normative Systems*. In Deo [47], 2018. ISBN 1848902786.

- [48] Patrick Lin. *Autonomous Driving: Technical, Legal, and Social Aspects*, chapter Why Ethics Matters for Autonomous Cars, pages 69 – 86. In Markus Maurer [6], 2015. Chapter focuses on the ethics of the topic.
- [49] Michael Taylor. Self-driving mercedes-benzes will prioritize occupant safety over pedestrians. Blog. URL <https://blog.caranddriver.com/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians/>.
- [50] Rapjael Orlove. Now mercedes says its driverless cars won't run over pedestrians, that would be illegal. Internet, oct, .
- [51] Iyad Rahwan Jean-François Bonnefon, Azim Shariff. The social dilemma of autonomous vehicles. *Science*, 352:1573–1576, June 2016.
- [52] National Highway Traffic Safety Administration. Automated vehicles for safety. URL <https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety>.
- [53] Robert Latta. Safely ensuring lives future deployment and research in vehicle evolution act. URL <https://www.congress.gov/bill/115th-congress/house-bill/3388/text>.
- [54] Eno. Section-by-section comparison of house and senate autonomous vehicle bills. URL <https://www.enotrans.org/wp-content/uploads/2017/10/AV-Bill-SBS-Senate-Reported.pdf?x43122>.
- [55] Economic Commission for Europe Inland Transport Committee Working Party on Road Traffic Safety. Report of the sixty-eighth session of the working party on road traffic safety, 2014. URL <https://www.unece.org/fileadmin/DAM/trans/doc/2014/wp1/ECE-TRANS-WP1-145e.pdf>.
- [56] Economic Commission for Europe Inland Transport Committee Working Party on Road Traffic Safety. Report of the seventieth session of the working party on road traffic safety, 2015. URL

- <http://www.unece.org/fileadmin/DAM/trans/doc/2015/wp1/ECE-TRANS-WP1-149-Aadd-1e.pdf>.
- [57] United Nations. Convention on road traffic geneva 19 1949 proposal of amendments to articles 8 and 22. URL <https://treaties.un.org/doc/Publication/CN/2016/CN.91.2016-Eng.pdf>. Reference: C.N.91.2016.TREATIES-XI.B.1.
- [58] Council of the European Union European Parliament. Directive 2007/46/ec of the european parliament and of the council of 5 september 2007 establishing a framework for the approval of motor vehicles and their trailers, and of systems, components and separate technical units intended for such vehicles (framework directive) (text with eea relevance), September 2007. URL <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32007L0046>.
- [59] European Parliament. Civil law rules on robotics, european parliament resolution of 16 february 2017 with recommendations to the commission on civil law rules on robotics (2015/2103(inl)), February 2017. URL <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2017-0051+0+DOC+PDF+V0//EN>.
- [60] Government of the Netherlands. Self-driving vehicles. web. URL <https://www.government.nl/topics/mobility-public-transport-and-road-safety/self-driving-vehicles>.
- [61] Government Offices of Sweden. Government paves the way for self-driving vehicles, July 2017. URL <http://www.government.se/articles/2017/05/government-paves-the-way-for-self-driving-vehicles/>.
- [62] Bundesministerium der Justiz und für Verbraucherschutz. Straßenverkehrsgesetz (stvg) § 1a kraftfahrzeuge mit hoch- oder vollautomatisierter fahrfunktion. URL https://www.gesetze-im-internet.de/stvg/_1a.html.

- [63] Department for Transport. The pathway to driverless cars: A code of practice for testing, 2015. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/446316/pathway-driverless-cars.pdf.
- [64] Government to review driving laws in preparation for self-driving vehicles, March 2018. URL <https://www.gov.uk/government/news/government-to-review-driving-laws-in-preparation-for-self-driving-vehicles>.
- [65] Austria, Hungary, Slovenia set up middle Europe driverless region, 2018. URL <http://www.worldhighways.com/sections/technology/news/austria-hungary-slovenia-set-up-middle-europe-driverless-region/>.
- [66] Ordonnance n° 2016-1057 du 3 août 2016 relative à l'expérimentation de véhicules à délégation de conduite sur les voies publiques, August 2016. URL <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000032966695&dateTexte=20180419>.
- [67] Projet de loi ratifiant l'ordonnance n° 2016-1057 du 3 août 2016 relative à l'expérimentation de véhicules à délégation de conduite sur les voies publiques, May 2018. URL <http://www.senat.fr/dossier-legislatif/pjl16-355.html>.
- [68] Danson Cheong. New rules for autonomous vehicles, February 2017. URL <http://www.straitstimes.com/singapore/new-rules-for-autonomous-vehicles>.
- [69] Dubai Future Foundation. Mohammed bin Rashid approves Dubai autonomous transportation strategy, April 2016. URL <http://www.dubaifuture.gov.ae/mohammed-bin-rashid-approves-dubai-autonomous-transportation-strategy/>.

- [70] Video: Now ride a tesla taxi right here in dubai, September 2017. URL <https://www.khaleejtimes.com/news/transport/50-tesla-vehicles-join-dubai-taxi-fleet>.
- [71] Robert Trypuz Michael Musielewicz, Piotr Kulicki. *Towards a Formal Ethics for Autonomous Cars*, pages 193 – 210. In Deo [47], 2018. ISBN 1848902786.
- [72] State of Minnesota. Minnesota statutes pedestrian, 2018. URL <https://www.revisor.mn.gov/statutes/cite/169.21>.
- [73] Nathalie Nevejans. European civil law rules in robotics, 2016. URL http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU%282016%29571379_EN.pdf. European Union.
- [74] Robert Dundobald Melville. *A manual of the principles of Roman law relating to persons, property, and obligations*. W. Green & Son Ltd, 1915.
- [75] Bartosz Brozek. *Legal Personhood: Animals, Artificial Intelligence and the Unborn*, chapter The Troublesome 'Person', pages 3 – 14. Springer, 2017.
- [76] Boethius. H.F. Stewart. William Heinemann. *Theological Tractates and the Consolation of Philosophy*. Harvard University Press, 1918.
- [77] Otto von Gierke. *Political Theories of the Middle Age*. Cambridge University Press, 1922.
- [78] Ernst H. Kantorowicz. *The King's Two Bodies: A Study in Mediaeval Political Theology*. Princeton University Press, 1917.
- [79] Visa AJ Kurki. *Legal Personhood: Animals, Artificial Intelligence and the Unborn*, chapter Why Things Can Hold Rights: Reconceptualizing the Legal Person, pages 69–89. Springer, 2017.

- [80] Virginia Dignum. Responsible autonomy. *CoRR*, abs/1706.02513, 2017. URL <http://arxiv.org/abs/1706.02513>.
- [81] Lauren Cassi Davis. Would you pull the trolley switch? does it matter?: The lifespan of a thought experiment. 2015. URL <https://www.theatlantic.com/technology/archive/2015/10/trolley-problem-history-psychology-morality-driverless-cars/409732/>.
- [82] Olivia Goldhill. Philosophers are building ethical algorithms to help control self-driving cars. *Quartz*, 2018. URL <https://qz.com/1204395/self-driving-cars-trolley-problem-philosophers-are-building-ethical-algorithms-to-solve-the-problem/>.
- [83] Marlene Cmons. What moral code should your self-driving car follow?teaching a robot to make ethical decisions is pretty complicated. *Popular Science*, 2017. URL <https://www.popsci.com/conscience-self-driving-car>.
- [84] Joseph Migga Kizza. *Ethical and Social Issues in the Information Age*. Springer-Verlag GmbH, 2018.
- [85] Isaiah Berlin. *Liberty*. Oxford University Press, 2002.
- [86] Gerald Conway. *The Limits of Legal Reasoning and the European Court of Justice*. Studies in European Law and Policy. Cambridge University Press, 2012.
- [87] Judith Jarvis Thomson. Turning the trolley. *Philosophy and Public Affairs*, 36(4):359–374, 2008.
- [88] Patrick Lin, Keith Abney, and George A. Bekey. *Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed*. 2012.
- [89] Utilitarian and deontological frameworks. Web. URL <https://sdcars.weebly.com/utilitarian-and-deontological-frameworks.html>.

- [90] Richard Kraut. Aristotle's ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2018 edition, 2018.
- [91] G. E. M. Anscombe. Modern moral philosophy. *Philosophy*, 33(124):1–19, 1958.
- [92] Bernard Smart, J. J. C.; Williams. *Utilitarianism (For and Against)* —— *A critique of utilitarianism*, volume 10.1017/CBO9780511840852. 1973. doi: 10.1017/CBO9780511840852.002.
- [93] Heidi Li Feldman. Prudence, benevolence, and negligence: Virtue ethics and tort law. *Georgetown Law Faculty Publications and Other Works*, 78, 2000. URL <https://scholarship.law.georgetown.edu/facpub/78>.