AUTO:
THE ETHICS OF MACHINE LEARNING ALGORITHMS & AUTONOMOUS VEHICLES

by

Stephan Furlin

Honours Bachelor of Arts McGill University 2017

A thesis

presented to Ryerson University

in partial fulfilment of the requirements for the degree of

Master of Arts

in the program of

Philosophy

Toronto, Ontario, Canada, 2020

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

Auto: The Ethics Of Machine Learning Algorithms & Autonomous Vehicles, Stephan Furlin, Master of Arts in Philosophy, Ryerson University, 2020

ABSTRACT

This thesis aims to address what I hold to be the most pressing ethical issues in autonomous vehicles. Chapter one focuses on the applicability of moral philosophy in guiding autonomous vehicle regulation. Chapter two highlights the need for "explainability" in machine learning algorithms in order to ensure that autonomous vehicles are fair and rights-respecting, and the need for a change in regards to distribution of liability for driving behavior. Chapter three asserts that laws regarding liability must be altered in order to keep pace with the changes in driver responsibilities which come with less direct control of the vehicle.

ACKNOWLEDGEMENTS

I would like to express my appreciation to Dr. Chris MacDonald for his role as my thesis supervisor. His thoughtful insights, dedicated effort, and numerous constructive contributions were invaluable throughout the process of writing this thesis.

I would also like to thank the following professors for their feedback as second readers:
Dr. Antoine Panaioti
Dr. Michael Milona

TABLE OF CONTENTS

INTRODUCTION

0.0 THESIS STATEMENT

My primary project in this thesis is to address what I hold to be the most pressing ethical issues in autonomous vehicles. These are the applicability of moral philosophy (via the Trolley Problem) to guiding autonomous vehicle regulation, the need for "explainability" in machine learning algorithms in order to ensure that autonomous vehicles are fair and rights-respecting, and the need for a change in regards to distribution of liability for driving behavior. In this introductory chapter I will introduce the factors which bring about these ethical issues and clarify key definition which will be used throughout the remainder of the thesis.

0.1 EXISTING LITERATURE

The literature on autonomous vehicle ethics is in its infancy, as the topic has only recently become a pressing one due to the rise of autonomous operation functionality in vehicles. As a subset of A.I. ethics, which is itself relatively young as well, autonomous vehicle ethics shares several key interests with A.I. ethics pertaining to the ethical development of new technology, how to approach regulation, concerns about safety, and impacts of the technology on society. Autonomous vehicle technology is an area of interdisciplinary research and debate. This includes a debate as to the place of autonomous vehicle ethics, with industry experts often resisting the integration of existing moral frameworks into autonomous vehicles in favor of tenets of other fields such as public policy, legal theory, and engineering standards. My first chapter aims to challenge this exclusion by demonstrating the relevance of moral philosophy in addressing the ethical challenges of autonomous vehicles.

Similarly, the need for explainability in machine learning algorithms is often approached from the perspective of a need to prevent bias, intentional or unintentional, from affecting the system. The purpose of these efforts is often to prevent such systems from compromising the rights and liberties of individuals who have to interact with that system. How autonomous vehicle ethics differs from A.I. ethics is that autonomous vehicles must always directly and physically interact with humans and must do so in complex environments like roadways or cities. This introduces a much stronger need to for precautions against unjust autonomous vehicle behavior resulting from bias or manufacturer programming errors as these systems cannot be isolated in the same way that a closed-environment A.I. system like a data processing algorithm can. The factors which influence whether or not a system is biased can originate in technical limitations, biased or incomplete input data, unintentional programming errors, and many others. The primary barrier to identifying this bias is the opacity of black box machine learning algorithms, which greatly impedes independent scrutiny of the autonomous vehicle's behavior. Chapter two explores these issues as they pertain to autonomous vehicles and suggests that the potential risks are great enough to encourage a transition away from black box algorithms.

Due to the rapidly developing nature of autonomous vehicle technology there are often situations in which legislation lags behind real-world operation of autonomous vehicles. Chapter three explores on such instance of this problem, namely the degree to which the role of a manual vehicle driver differs from that of an autonomous vehicle driver not being adequately recognized in multiple international jurisdictions. The purpose of focusing on this issue is to encourage more evaluation in the area of autonomous vehicle liability in regards to manufacturers who will be shown to have increasing influence over autonomous vehicle behavior.

0.2 OVERVIEW

The theoretical ideal of an autonomous vehicle is a mode of transport in which an individual merely sets a destination and is able to arrive at that destination without any risk and without the need for any additional input needed. This idea is not new. One of the earliest mentions of a mode of transport resembling an autonomous vehicle can be observed in Homer's Odyssey in the form of the mythical Phaeacian ships which finally carry a weary Odysseus back to Ithaca at the end of his ten-year journey.

> *"Tell me also your country, nation, and city, that our ships may shape their purpose accordingly and take you there. For the Phaeacians have no pilots; their vessels have no rudders as those of other nations have, but the ships themselves understand what it is that we are thinking about and want; they know all the cities and countries in the whole world, and can traverse the sea just as well even when it is covered with mist and cloud, so that there is no danger of being wrecked or coming to any harm." (Homer Od. Book VIII)*

These ships are written to be so adept at their tasks that Odysseus is able to sleep through the entire journey home after merely relaying his destination. Theoretically, this is the ultimate end goal of autonomous vehicles in terms of functionality.

0.3 WHAT IS AN AUTONOMOUS VEHICLE?

Broadly speaking there are many modes of transportation which may be considered autonomous vehicle ranging from autopilot on airplanes to even simple location-based transportation systems like elevators. All perform the same task, namely, transporting individuals, animals, and objects from one location to another. What I will be focusing on in my

thesis are autonomous vehicle which operate in uncontrolled environments like roadways and cities, as such environment present a distinct set of ethical challenges which I aim to address. Technically, an autonomous vehicle is any vehicle which contains systems which handle part or all of the "dynamic driving tasks" of a given vehicle operating within a specific "operation design domain". To be specific, what SAE International, an international engineering standards organization which is regarded as the primary source for autonomous vehicle terminology and engineering standards, defines as "dynamic driving tasks" (DDT) are all of the real-time operational and tactical functions required to operate a vehicle in on-road traffic such as lateral and longitudinal vehicle motion control (steering, acceleration, deceleration, and braking), recognizing and monitoring objects and events in the driving environment, and both preparing and executing responses to said objects and events by means of manoeuvering, signalling, or enhancing conspicuity via lighting (SAE 2018, pg. 6-7). The degree to which DDT tasks are automated in a given vehicle are classified by the SAE in six levels of autonomous vehicles. The SAE levels of autonomous vehicle automation are summarized as follows:

Level 0 – No Driving Automation

"The performance by the driver of the entire DDT, even when enhanced by active safety systems." (SAE 2018, pg. 19)

Level 1 – Driver Assistance

"The sustained and ODD-specific execution by a driving automation system of either the lateral or the longitudinal vehicle motion control subtask of the DDT (but not both simultaneously) with the expectation that the driver performs the remainder of the DDT." (SAE 2018, pg. 19)

Level 2 – Partial Driving Automation

"The sustained and ODD-specific execution by a driving automation system of both the lateral

and longitudinal vehicle motion control subtasks of the DDT with the expectation that the driver completes the OEDR subtask and supervises the driving automation system." (SAE 2018, pg. 19)

Level 3 – Conditional Driving Automation

"The sustained and ODD-specific performance by an ADS of the entire DDT with the expectation that the DDT fallback-ready user is receptive to ADS-issued requests to intervene, as well as to DDT performance-relevant system failures in other vehicle systems, and will respond appropriately." (SAE 2018, pg. 19)

Level 4 – High Driving Automation

"The sustained and ODD-specific performance by an ADS of the entire DDT and DDT fallback, without any expectation that a user will respond to a request to intervene." (SAE 2018, pg. 19)

Level 5 – Full Driving Automation

"The sustained and unconditional (i.e., not ODD-specific) performance by an ADS of the entire DDT and DDT fallback without any expectation that a user will respond to a request to intervene." (SAE 2018, pg. 19)

0.4 EARLY VEHICLE AUTOMATION

Interestingly, the SAE levels of automation are broad enough to include quite a range of older technology that one wouldn't typically associate with autonomous vehicles. Single-task automations to simplify a vehicle operator's ease of use have been in cars for over 100 years. The first came to cars in the form of automotive niceties such as the automatic transmissions in 1921, the automatic braking system in 1929, and cruise control in 1948. What these systems have in common is that they automate menial tasks in a rudimentary form of driver's assistance, thus meeting the requirement s for level 1 automation. Most of these features, by means of their very

simple, specialized functions, raised few ethical issues. These features were previously considered ethically trivial extensions of manual car operation as the tasks which they focused on automating were minimally relevant when compared to the actual driving decisions a driver made. After all, one hardly thinks that simply because their turn signal blinks automatically instead of having to hold down a button that it should somehow create a need for new ethics.

0.5 OPERATIONAL DESIGN DOMAIN

This changed with modern autonomous vehicles. The foundations of modern algorithmic autonomous vehicle technology were first developed in the 1980's, when Mercedes-Benz produced a vehicle guided by computer vision that was successfully able to navigate empty roads at speeds of 39 miles per hour, thus meeting the bare minimum requirements of the SAE definition of a level 2 vehicle (Delcker 2019). The difference between the aforementioned historical autonomous vehicles and modern autonomous vehicles is the difference in "operational design domain". The SAE defines the operational design domain (ODD) as the operating conditions under which a given automated driving feature or automation system is specifically designed to function, including environmental conditions, geographical location, and time-of-day (SAE 2018, pg. 6-7). For example, the ODD of an elevator is lateral movement within an environment of elevator shaft, up to a certain rated weight threshold, when electricity is present, except during fire evacuations. In comparison to the elevator, modern autonomous vehicles such as autonomous cars have a much wider operational design domain in that they are able to take on a wider variety of tasks and operate in increasingly diverse environments.

0.6 THESIS OUTLINE

With this increase in ODD comes increased variance in the circumstances that autonomous vehicles must face, at least some of which cannot be easily dealt with by classical programming methods. This problem has resulted in the increasing use of machine learning algorithms in autonomous vehicles. The programming values used to generate these algorithms use a complex series of pre-set goal directives set by the programmer alongside vast swathes of data to allow the program to self-generate methods of automatic vehicle operation via a process called “training”. The first chapter of this thesis will aim to demonstrate that the need to prioritize these various pre-set programming directives in cases which they come into conflict necessitates the application of moral philosophy. I argue that as these pre-set directives must be determined in advance this necessitating that manufacturers make pre-set choices as to how their vehicles will operate.

The second chapter will argue that the opaque nature of machine learning algorithms make them ethically problematic. and will argue that the algorithms guiding autonomous vehicles must be at least partially explainable in order to be properly regulated. This chapter will also address controversy as to whether “full explainability” requirements place an undue burden on manufacturers. I will first demonstrate how machine learning algorithms are “black box” systems which are opaque to regulators, industry experts, and even their own programmers. In doing so I will demonstrate that it is impossible to regulate such an opaque system and provide several examples as to how such a system can easily and undetectably infringe on individual rights. The conclusion of this chapter will then propose solutions that could mitigate the “black box” problem. This solution is based on installing operational data recorders in autonomous vehicles and requiring certain key subtasks of the vehicles system to be fully explainable.

The third chapter will assert that laws regarding liability must be altered in order to keep pace with the changes in driver responsibilities which come with less direct control of the vehicle. This chapter will address the widespread issue of regulators in multiple countries improperly treating driving liability in autonomous vehicles as a straightforward extension of how they treat driver liability in manual vehicles. I argue that said practices is unjustifiable. I will demonstrate that a vehicle operator is often not directly controlling the vehicle, nor partaking in shared agency in jointly operating the vehicle. The ultimate conclusion of this chapter will be that autonomous vehicle liability needs to expand to hold manufacturers responsible for potential defects in the autonomous vehicle's programming as such malfunctions directly affect vehicle behavior. My hope is that, through addressing the issues explored in these three chapters, I will have provided a clear means to approach these issues philosophically, thereby encouraging further work on these problems.

CHAPTER 1

THE TROLLEY PROBLEM AND ITS APPLICABILITY TO AUTONOMOUS VEHICLES

1.0 INTRODUCTION

There is significant disagreement as to whether conventional moral philosophy is applicable to the kinds of challenges facing the programmers and manufacturers of autonomous vehicles. This to say, manufacturers and programmers of autonomous vehicles often put forth the position that an entirely novel ethical frameworks must be developed in order to solve problems which existing conventional moral philosophy cannot. The idea behind such claims is that existing moral philosophy cannot properly account for nuanced ethical situations faced by autonomous vehicles. Those skeptical of the applicability of conventional moral philosophy (represented in this chapter by Noah Goodall) even go so far as to take the position that "harm avoidance" algorithms, legislation reactive to issues which occur, and approaches which meld multiple competing public objectives can be relied on as the primary guiding forces in creating ethical autonomous vehicles. Thus, they conclude that to use theoretical models from conventional philosophical sources such as Philippa Foot's trolley problem to think about how autonomous vehicles should be programmed is distracting from issues which occur in real-world scenarios. However, proponents of the applicability of conventional moral philosophy, represented in this chapter by Janet Fleetwood, disagree with this assessment on the basis that there is an unavoidable ethical component to the algorithmic determinations of autonomous vehicles which should be informed by traditional ethical frameworks. The purpose of this chapter is to address this disagreement and clarify the role of conventional models from moral philosophy in guiding autonomous vehicle regulation. My aim for this section is to demonstrate the applicability of conventional models from moral philosophy for the purpose of programming

autonomous vehicles. Using ideas derived from Philippa Foot's "trolley problem", which is a central topic of contention for Fleetwood and Goodall, I aim to demonstrate the relevance of conventional moral philosophy in informing choices made by manufacturers which in turn have a direct impact on autonomous vehicle behavior. First, I will identify two distinct interpretations of the trolley problem at play in the disagreement between Goodall and Fleetwood. Next, I will use Fleetwood's analysis of the pre-emptive algorithmic driving choices made autonomous vehicle manufacturers to demonstrate a need for ethical priorities which goes beyond "harm avoidance algorithms". Finally, I will show that Goodall's criticisms pertaining to the "idealized" and "unrealistic" nature of the trolley problem in real-world scenarios refer to a version of the problem which I identify at the "Programming Trolley problem" but that they do not address the points put forth by Fleetwood's second interpretation of the trolley problem identified as the Objective Prioritization problem.

1.1 THE TROLLEY PROBLEM

Foot's trolley problem depicts a runaway trolley barrelling down a railway track towards five individuals who are tied up and unable to move. There is a lever which can be pulled which will divert the trolley down a side track on which there is one individual who is also tied up and unable to move. In this scenario there are two possible outcomes. In the case that the lever is pulled the trolley will be diverted to the side track and the individual on that track will die. Conversely, if the lever is not pulled, the trolley will remain on its current course and five individuals on the track will die. The problem asks whether or not the lever should be pulled, in order to test an individual's ethical intuitions. There are also numerous versions of the trolley problem which test the convictions of the decision maker by inserting a particular context or slightly altering the problem. For example, a version of the problem known as the "trauma case"

substitutes the railway context for that of a hospital and the train track victims for one healthy patient and five organ donor recipients. Another case asks if someone had to push a fat man over a bridge to block the trolley instead of pulling a lever. In recent years this same ethical dilemma has been posed in the context of autonomous vehicles. There are several relevant contextual versions of this issue which can be summarized as being a subset of the following "Programing Trolley problem" (PT problem):

The programming of an autonomous vehicle is forced to choose between veering into X party/parties or remain on its present course and collide with Y party/parties. Whichever party is collided with will die. How should it decide which party to veer towards?

In the latter sections of this paper we will examine a second interpretation of this problem, inspired by Fleetwood, which re-imagines the trolley problem as a means to assess the prioritization of a myriad of objectives such as lawfulness, harm reduction, and driver safety. In this section I will formalize this interpretation by demonstrating how the specifics of the trolley problem are replaced with a dilemma involving a "forced choice" that pre-empts driver choice in determining which of various objectives is prioritized over the others. The reason why I reinterpret the trolley problem in this way is that the order in which these objectives are prioritized will drastically affect a vehicle's behavior. For example, a car which prioritizes driver safety above lawfulness will break the law in favour of driver safety. I will refer to this re-imagined version of the trolley problem as the Objective Prioritization problem (OP problem):

The programmer of an autonomous vehicle who is designing algorithmic response system must choose between prioritizing A objective/objectives and B objective/objectives. When these objectives are in conflict whichever objective is de-prioritized will be disregarded or adhered to less strongly by the vehicle. If objective A is prioritized will change the vehicle's behavior in X way, if objective B is prioritized it will change vehicle behavior in Y way. What prioritization should be programmed into the vehicle?

To illustrate how the variables in the OP problem interact, consider a case in which a programmer decides to prioritize pedestrian safety (objective A) over lawfulness (objective B). In this case whenever the objectives of pedestrian safety and lawfulness are in conflict the vehicle will break the law in order to maximize pedestrian safety (behavior X). If this priority order were reversed (lawfulness prioritized over pedestrian safety) whenever the objectives of pedestrian safety and lawfulness are in conflict the vehicle would adhere to the law even in situation where it compromised pedestrian safety.

The OP problem is related to the PT problem in that, in any given scenario, the choice made in the PT problem is dependant upon the choice made in the OP problem. To provide a very simplified example of this relationship, if one were to prioritize "lawfulness" over "harm minimization" in the OP problem that same autonomous vehicle would pursue whichever solution to the PT problem caused it to break the fewest laws, regardless of how many people are injured or die as a result. In other words, the autonomous vehicle's behavior is the dependent variable while the prioritized objective is the independent variable. I hold that if the OP problem, a re-interpretation of the trolley problem, can be shown by Fleetwood to be applicable to autonomous vehicle behavior, then conventional philosophical models, the value of which

variants of the trolley problem are designed to illustrate, are relevant to programming autonomous vehicle behavior.

1.2 FLEETWOOD ON THE APPLICABILITY OF THE TROLLEY PROBLEM

Fleetwood's thesis regarding the programming of autonomous vehicles specifically targets autonomous cars meeting the criteria for levels 3, 4, or 5 of automation as defined by the SAE standards (Fleetwood 2017, pg. 533). In doing so, Fleetwood focuses on a more advanced range of autonomous vehicles which are classified as automatic driving systems (ADS systems). ADS systems differ from other forms of autonomous vehicle technology in that rather than being "assistive" to a human operator in a particular task such as in features like cruise control, they operate in the range in which the vast majority of the dynamic driving tasks (DDT) such as signalling, changing lanes, observing road conditions, and others are fully automated and human intervention in their operation is rare (SAE 2018, pg. 6-7). In ADS cases, the DDT "choices" are not made by the operator of the vehicle but are effectively made pre-emptively by programmers during the design phase of manufacturing through pre-programmed objectives, hence Fleetwood's description of such choices as "forced choices". Fleetwood asserts that if there are pre-emptive "forced choices" being made as to objectives which affect the vehicle's behavior, there is also a clear need establish a priority structure for said choices should they come into conflict with one another (Fleetwood 2017, pg. 535). Such a choice prioritization structure can be directly modeled by the OP problem given that the order in which these pre-emptive choices are prioritized will affect an autonomous vehicles behavior. As a result, if this claim for this objective prioritization structure can be defended, it provides a direct application of conventional philosophical models of ethics.

Fleetwood's position is based on the claim that autonomous cars have to make rapid, time-constrained decisions using incomplete information, often in situations that programmers will not have considered (Fleetwood 2017 pg. 534). As noted in the previous chapter, there are a multitude of potential operational design domains ranging from simple domains with minimal variance, such as elevators and monorails, to complex, highly variable domains such as cities and airports. Within these domains autonomous vehicle technology is already prevalent in multiple modes of transport from elevators to airplanes. Even when one limits the type of vehicle to a specific kind, such as cars, these vehicles have the potential to participate in multiple operational design domains. Consider how an autonomous car may find itself on a highway, in a parking garage, at an airport drop-off zone, or off-road at a remote campsite. Each of these environments will have differing variables which influence the car's algorithmic decision-making. Manufacturers try to address this problem by limiting the domains in which vehicles operate to specific "design" zones which they have directly tested and account for. This is usually accomplished either through direct written guidance to consumers or by programming a means to detect conditions outside of the ODD approved range and subsequently disable autonomous features if such conditions are detected. For example, an autonomous vehicle not is not certified to operate in snow can be programmed to refuse to engage autonomous driving features when snow is detected. However, that does not change the fact that some of these environments are so variable that even with these safety limitations there will be situations which the system has never encountered nor been programmed to deal with. In such situations, in order for operation in these environments to be fully automated, the software that makes DDT decisions in level 3, 4, and 5 vehicles must be pre-programmed with specific priorities to fall-back on, such as the safety of the driver, the minimization of risk, obedience to the law, and many others. There are

clearly situations in which autonomous vehicles will need to make decisions which their programming has not accounted for. Therefore, Fleetwood argues that there is a need to consider the OP problem as a way to help establish priority order for these objectives so that these decisions can be made in a reasonable and predictable way. The OP problem, in other words, constitutes an important simple model for the prioritization task faced by autonomous vehicles.

1.3 OBJECTIONS TO FLEETWOOD

Karl Iagnemma, president of Aptiv Automated Mobility and cofounder of the autonomous vehicle company nuTonomy has attempted to rebut positions similar to Fleetwood's by arguing that autonomous vehicles can be programmed in such a way that they will avoid scenarios which they are not programmed to deal with (Marshall 2018). What Iagnemma implies by this statement is that "objective conflict" interactions will occur so rarely that they will be of minimal concern. This argument is questionable considering that several well-publicized autonomous car crashes have already occurred. Even if one were to disregard these past instances on the basis of technological advancements which improve safety, the argument that redundancy features for objective conflict scenarios should not warrant concern because they are rare is not defensible. There should always be a redundancy present in the case that the vehicle encounters a situation beyond the scope of circumstances programmers have accounted for.

Along similar lines, potential critics of the idea of applying the OP problem to autonomous vehicles could also argue that the human operator of the vehicle is present to be a "fall-back" substitution for the ADS system for such situations. They could thus, deny the necessity of decisions based on extrapolations of vehicle programming objectives. However, this argument seems to ignore a number of reasonably foreseeable autonomous vehicle use cases in

which the human operator is unable to perform a "DDT fallback". This would be the case, for example, if the operator of the vehicle is a disabled person, a child with no knowledge of how to operate a vehicle, an operator otherwise incapacitated/unable to perform a fallback due to distress/fright, or if there is no human operator at all. There are also foreseeable driving situations in which actions would need to be taken so quickly that there is not enough time to perform a DDT fallback. Such situations may be rare, but given the stakes they still cannot be ignored.

Goodall takes a different approach from Iagnamma's by denying the applicability of the PT problem, rather than the OP problem. Despite directly responding to Fleetwood's paper, Goodall fails to realize that the PT problem is just a single potential derivative of the OP problem, one that can be substituted for any scenario which involves algorithmic objective prioritization. This misidentification of Fleetwood's position with the PT problem is the primary failing of Goodall's critique. Goodall's first objection to Fleetwood's view is that Fleetwood does not take into account the real-world variability which is present in a feature of vehicle collision scenarios. Goodall's primary objection reflects this issue when he points out that the trolley problem, "…represents a clear choice with only two distinct alternatives, and assumes completely certain outcomes with obvious moral consequences…real driving dilemmas have many subtle choices, uncertain outcomes, and often an obviously superior course of action…" (Goodall 2017, pg. 496) Consider the fact that a human driver may slam on the brakes, decide to hit a tree or lamppost instead of a human, turn off the road, or even swerve around the pedestrians entirely; an autonomous vehicle has a similar range of options. While this is a substantial critique of the PT problem, Fleetwood specifically recognizes this same potential variability when describing forced choice algorithms. Fleetwood's position is not based on the

PT problem but instead argues for the applicability of moral philosophy through the OP problem. The OP problem, as mentioned previously, recognizes that the methods of prioritizing affect all potential driving determinations made by autonomous vehicles. Thus, Goodall is not actually directly critiquing Fleetwood's position.

Goodall goes on to argue that the PT problem places unjustifiable emphasis on inevitable crash scenarios. In particular, Goodall argues that "trolley problems" are a vast oversimplification of the ethics pertaining to algorithmic decision making. Instead of critiquing Fleetwood's position directly, he instead critiques the popular fixation with the trolley problem when he states,

> *"All driving, not just pre-crash driving, requires assigning values to different objects. How much space to give a cyclist as it passes, how much to slow down in a residential neighborhood—these decisions require the vehicle to balance the safety of its own passengers and road users, and to balance safety and time. These subtle decisions will affect safety." (Goodall 2017, pg. 496)*

In making this critique Goodall points out that even philosophically mundane circumstances will need to be accounted for when designing ethical frameworks for autonomous vehicles. Given the large number of circumstances a vehicle may find itself in, Goodall implies that the possibility of using a traditional ethical framework designed around narrow circumstances such as the trolley problem incorrectly direct vehicle behavior to compensate for uncommon PT problem scenarios instead of common scenarios. Therefore, due to the complexity of such a project, Goodall argues for moving away from conventional moral philosophy in

favour of a practical methodology which mixes attributes of multiple traditional frameworks in a manner similar to other public health issues such as in healthcare organ donation systems (Goodall 2017 pg. 496). But this argument falls flat on account of the fact that the systems which govern organ donation processes use consistent prioritization standards, the same kind of standards which inspired Fleetwood's likening of autonomous vehicle priorities to public health issues, in order to be ethically consistent for the sake of fairness to patients (Bickenbach 2016).

1.4 CONCLUSION

When viewed in the context of real-world traffic behavior as the PT problem, the trolley problem's abstract and highly theoretical nature might makes it seem superficially ineffective and inapplicable for the accurate assessment and direct operation of autonomous car behavior. However, while one might be justified in arguing that this literal application of Foot's trolley problem in the form of the PT problem is unrealistic, such a position neglects the actual purpose of integrating conventional moral philosophy into autonomous vehicles in the form of the OP problem. What is being tested by the OP problem is not the ethics of select situations nor thousands of micro-decisions which occur over the course of the average drive, but rather the outcomes of prioritizing certain ethical values over others. In other words, the PT problem is only one of a vast set of potential scenarios both collision-specific and mundane that can be tested by the OP problem making it both highly applicable and able to be integrated with real-world data. Without knowing what values or "objectives" are being prioritized one cannot guarantee that an autonomous vehicle will behave in a predictable and consistent way; thus, potentially risking public safety. The OP problem, and by extension conventional models of moral philosophy, have therefore been shown to be directly applicable in guiding the development of objective prioritization in autonomous vehicle behavior.

CHAPTER 2

AUTOMATION & THE NEED FOR EXPLAINABILITY

2.0 INTRODUCTION

This chapter will argue for regulators to impose legal obligations on autonomous vehicle manufacturers requiring them to do away with the opaque calculations typical of "black box" systems, at least in regards to a specific subtasks of the automatic driving system (ADS system) known as object and event detection and response tasks (OEDR) (SAE 2018). My primary reasoning for this is that in order for a vehicle to be considered "rights-respecting", it must be able to be provably compliant with legal protections afforded to individual rights. My position differs from hardline position of "full explainability" in that it allows for all other non-crucial subtasks to be "black box" systems. Such a system, with both opaque and clear calculations is referred to as a "grey box" system and represents a meaningful compromise between the risks of complete opacity and the prohibitive costliness of implementing absolute transparency. I hold that the OEDR subtask, by nature of its role in object recognition, environmental interpretation, and response determination functions, is the most significant factor in regards to developing rights-respecting autonomous vehicles and as such must be directly explainable and assessable by independent experts. Additionally, I will demonstrate that if allowed to persist as a "black box" system, the OEDR subtask will directly interfere with regulatory oversight, human rights, and determinations of legal liability.

2.1 TRADITIONAL PROGRAMMING VS. ALGORITHMIC PROGRAMMING

In simplified terms, traditional programming methods rely on a human software engineer to manually code the logic of a system. For example, if a programmer were to change how often

a GPS system polls a satellite for location data that change would be “explainable” to other software engineers or regulators. That is, someone with suitable training would be able to read the software and know how often the GPS was programmed to do. In this way, every action a system takes can be explained and traced to a human action. The difference in the case of learning algorithms is that while the initial version of the algorithm and the parameters are set by humans, subsequent “adaptations” are determined by the algorithm itself. In order to program such systems, software engineers feed the system vast quantities of data and “train” it by tuning its responses to create a set of desired results. For example, a software engineer may use a learning algorithm to sort a set of images based on their content and then “tune” the algorithm by feeding more data into it which indicates which images were identified and sorted into the correct categories and which images were sorted incorrectly. This “feedback loop” is often automated with more complex systems due to the high number of potential results which need to be narrowed down in the analysis portion. For example, chess-playing machine learning algorithms like Google’s Alpha-Zero are automated to pursue a win in chess. Such an algorithm achieves its effectiveness by playing millions upon millions of games against itself and selecting winning situational assessments and strategies without additional human input (Silver et Al. 2018).

2.2 WHY USE MACHINE LEARNING ALGORITHMS?

The limitation of using traditional coding methods to program a system (an autonomous car, in particular) is that the work required to make a reliable system grows exponentially as the complexity of the operational design domain increases. The complexity of and sheer scale of variable interactions within certain operation design domains such as cities or airports are so great that to conceive of a programming framework which accounts for every possible

interaction that could occur would require a supercomputer if not multiple supercomputers. As a result, not only would the traditional calculation take so long as to effectively prevent a vehicle from making real-time decisions, it would also be impossible for most vehicles to house the necessary processing resources. Even if radical developments pertaining to the miniaturization of this computer technology occurred, the amount of specificity programmed into the vehicle would be so intensive that the production of autonomous vehicles above SAE level two (true ADS systems) would not be commercially viable. Therefore, manufacturers of autonomous vehicles use machine learning algorithms in order to make their products practically and economically feasible.

2.3 THE BLACK BOX PROBLEM

In this section I will describe what a "black box" system is in relation to other types of systems in order to describe the explainability problems such a system causes. An "open system" is any system which has both input and output interactions with an external environment (Ehmer & Farmeena 2012). In this context, an "interaction" is defined as any exchange of material, energy, or information between a system and its environment. When defined in this manner, a "clear box" system (also commonly referred to as a white box, transparent box, or glass box system) is able to have the manner of operation behind its determinations observed directly (Ehmer & Farmeena 2012). For example, a standard "office suite" computer program has a source code which can be directly observed by someone with suitable technical expertise, thereby allowing the expert to determine how a keystroke input is interpreted by the code to select a particular ascii character and display it on a printed document as an output. By contrast, the calculations within a "black box" system are opaque to any observer, pretty much regardless of expertise. In other words, the only means by which to attempt to test such a system is to

provide a series of inputs and observe the outputs. A modern example of such a system is readily present in complex learning algorithms. Programmers often do not know how a learning algorithm arrives at its solutions, merely that it outputs the response the programmers are aiming for. If the chess program produces winning strategies, for example, that is all that matters and all that can be determined. Currently, simple learning algorithms are relatively explainable, because they are relatively noncomplex, but they become increasingly opaque as they become more complex. While this "opacity" is a non-issue for harmless, contained applications such as chess algorithms which only operate in controlled virtual environments, it becomes ethically problematic when these "opaque" calculations involve the welfare of living beings, most notably in regards to issues of safety.

2.4 ELECTION EXAMPLE

To better contextualize this issue, consider an electronic voting booth for a small-town election between two candidates. Call this the "system". The "inputs" are all of the completed ballots of the town's residents and the "output" must be one of the two candidates. If the electronic voting booth were to operate like a "clear box" system it would operate according to a clear set of rules. For example, it would be programmed to count all the votes and record the results, then select the candidate with the most votes and declare them the winner of the election. In this way, its methods, reasoning, and variables for determining the winning candidate are directly observable. What is programmed in by one programmer can be read and verified by another. In this scenario if a regulator were asked if the election had been handled fairly, and in accordance with existing laws all the information needed to make such a determination is regularly available. The regulator could have an expert look at the electronic voting booth's code, and even test each stage of its operation.

By contrast, if this electronic voting booth were to operate like a "black box" system it would first be given a directive such as "select the most appropriate candidate based on the provided data". The townspeople would then give some parameters to the election official such as needing to use a method which not random, restricting the possible outcomes to the two candidates, and needing to use the completed ballots in its decision making. It would then be fed votes and results from past elections to "train" it to develop a method of determining a winner. Once the program could consistently arrive at the same results as past election (presumably stumbling upon some variant of the 'majority wins' rule) it could then be fed "new" data from the current election and use it to select a winning candidate according to the method it developed from the training data. If the same regulator were asked if the election had been handled fairly and in accordance with existing laws there would be no means for them to do so as while the inputs (ballots) and outputs (candidates) could be observed the exact method implemented by the "black box" used by the election official to run the election would be unknown, and perhaps indecipherable with modern technology.

2.5 ALGORITHMIC BIAS

But why should it matter that the inner workings of the voting machine are unknown? What if the results are always viewed as "good enough" by its user's standards that there is no question as to if the machine should ever be doubted? Why should one care that they cannot directly observe the method of a black box system if the outcomes are always acceptable once it is sufficiently trained? The answer to this lies in the fact that algorithmic black box methodology could be anything that produces results that fit a programmer's/client's subjective goal parameters. As the program is being tuned based on its "end goal" there may be multiple ways to

achieve the same election results, it could be sound logic, random chance, or even a mixture of rational and irrational variables.

To explain why this could be ethically problematic, suppose a resident named "Unlucky Anne" lives in the same small town described in the previous example. Anne is very elderly and as a result has voted in every town election since the town was founded. As a result, her vote is present in all iterations of the past data that was used to "train" the "black box" electronic voting machine. Anne has different political views from the majority of the townspeople and as a result has voted for the losing candidate every single time an election has been run. This data could be interpreted by the learning algorithm to be a very strong determiner of the unsuccessful candidate as Anne's vote is, historically, an accurate predictor that the other candidate is going to win. Given this 100% accuracy, Anne's vote could have the effect of biasing the system. The system could accidentally hit upon the rule, "If Anne votes for X, then X loses." As a result, the electronic voting machine ignores all other inputs except for Anne's vote and simply designates the candidate that Anne did not vote for to be the winner. In this instance, not only is the election being run in a questionable way but Anne is actively being discriminated against by the algorithm as her chosen candidate can functionally never be elected. This discrimination will occur for as long as the electronic voting box is in use due to its opaque methodology.

Given that the possibility of this unfairness and discrimination inherently exists within a black box system by means of its opaque methodology, there is a significant need to move away from such methods and promote explainability algorithms instead. To provide a practical example of how such a problem may manifest in autonomous vehicles, consider that current autonomous vehicles rely heavily on algorithmic object recognition to identify pedestrians, other vehicles, and obstructions. This kind of software has been demonstrated to often contain an

unintentional bias, namely being less skilled at recognizing and identifying individuals of a number of identifiable groups (Bushwick 2019). There is a very real risk that this inability to identify individuals of certain identifiable groups could manifest through autonomous vehicles in lower road safety for said identifiable groups.

2.6 THE NEED FOR GOVERNMENT INTERVENTION

There is minimal financial incentive for automotive companies to slow their production in order to develop "clear box" systems. Companies might plausibly adopt the attitude that, if a problem (e.g., a pattern of crashes) occurs, there is no need for the kind of direct insight that a clear box system provides: the behavior of vehicles can simply be tuned through additional data and algorithmic "training". In other words, there is little incentive for companies to be proactive rather than reactive in regards to autonomous vehicle safety. This, paired with the fact that opaque "black box" technology make it significantly more difficult for regulatory agencies to prove non-compliance to government standards and more difficult for courts to assign clear liability, means that there may be a tendency for the technology necessary to bring about "clear box" systems to be neglected, barring government intervention. This unfortunate reality, paired with the standard financial conflict of interest that manufacturers face in being trusted to impartially evaluate or report on the safety of their own products, suggests the possibility of a substantial risk to public safety.

Numerous cases of regulatory non-compliance and of financial interests overshadowing public safety have been widely publicized. An egregious example of a known mechanical failure being intentionally overlooked is the case of the Ford Pinto. When informed of a vehicle defect which caused engine explosions in the Pinto, the auto manufacturer Ford infamously decided to risk the safety of its customers for financial savings rather than to fix a known mechanical issue

which resulted in engine fires resulting in an estimated 180 deaths (Leggett 1999). Similarly, attempts to use software techniques to circumvent government regulation in this manner have already been attempted. For example, Volkswagen deliberately installed software in its vehicles main computer to recognize test environments and then cheat on environmental emissions standard testing (Hotten 2015). This raises an obvious worry with regard to autonomous vehicles in that a similar "test environment" exploit program could be deployed in autonomous vehicles. With the added opacity of black-box system such a program would be extremely difficult to detect.

2.7 MAIN CRITICISM

Autonomous vehicles largely employ black box learning algorithms to determine their behavior. As a result, one presently cannot tell why an autonomous vehicle acted in a certain way, merely that it did act in a certain way when provided with a set of inputs. This issue would persist even if one had a complete list of all inputs and outputs. The key issue of the black box problem is the inability of the exact methods of the system to be scrutinized. If the programmers of the system themselves have only a minimal understanding of how a system works, there is little means by which to diagnose the cause of unintended behavior whether it be due to mechanical malfunction, software error, or even if the system has been compromised in some way. In addition, if a black box algorithm is compromised by a malicious party in some way it is extremely difficult to detect. In the case of a clear box system, the person who designed it is qualified to detect problems. But the methods used by the black box system to achieve its ends are designed by the algorithm meaning that there is no designer. Not only could this compromise road safety and endanger lives, but it could also completely lock down transit systems. By the same reasoning it is nearly impossible for third parties or regulatory bodies to adequately assess

black box systems, meaning that there is no possibility of meaningful oversight. In other words, it is not possible to determine if an autonomous vehicle is actually "rights respecting" or "legally compliant" or if it merely appears to be in most circumstances. While autonomous cars represent a vast improvement in convenience it is very clearly not possible to condone the widespread use of this technology until meaningful safeguards can be put in place.

2.8 RISK TO PUBLIC

One might object to my assessment on the basis that human drivers themselves are arguably "black box" systems wrought with a range of biases that affects their judgement in ways which may endanger other individuals in any operational design domain in which they participate. However, this objection does not take into account that there is variability in bias from person to person. The subtraction of the human variance behind vehicle decision-making means that autonomous vehicles will likely make mechanistically similar decisions to one another in most scenarios. This unfortunately has the consequence of potentially amplifying a minor programming error, bias, or hardware limitation to the potential scale of a public health crisis due to the sheer number of times the same error will be repeated. Take for example the Boeing 737 Max which had all planes of that model grounded due to a defect in its autopilot system (Slotnik 2020). The mere presence of an observable algorithmic error in one autonomous vehicle will mean that it is present in thousands if not millions of other vehicles, all of which will need to be serviced. In a recent study it was shown that if the majority of road vehicles were autonomous vehicles, only 20% of those vehicles would need to be compromised (either via software defect or malicious hackers) in order to completely gridlock an entire city (Vivek et al. 2019). This translates to weeks if not months of being forced to endure the threat of unsafe road conditions and a potential major disruption transportation. Likewise, due to potential algorithmic

bias, whether intentional or unintentional in nature, black box systems have the potential to disproportionately affect identifiable groups on a mass scale. The fact that individuals do not have the option to "opt out" of sharing spaces with autonomous vehicles means that they could be forced to endure unjustifiable disadvantages merely to travel in public.

2.9 DOES REGULATION STIFLE INNOVATION?

The common response to calls for regulation of black box programming pertains to how such regulation will stifle innovation. While prevalent, this is a rather weak argument that, in the case of learning algorithms, transparency requirements do not actually stop development of the technology. Instead, transparency requirements mean insisting that the technology be developed in a way which is provably law-abiding and rights-respecting. It does not matter how beneficial a technology is if it risks compromising the rights and safety of identifiable groups. While in the short term there are additional developmental barrier to overcome to clear box autonomous vehicles to market, the long-term advantage of regulations requiring clear box systems is the development of explainable learning algorithms, which is an ultimately superior version of the technology by means of avoiding the problems mentioned in previous section.

To provide an analogy to an existing product where such an outcome has already occurred, one can look to the field of medicine. Medications which have a well understood mechanism of action (the way which they interact with the body) are generally regarded as "safer" due to the fact that the additional understanding afforded to medical professionals allows for better predictability, more accurate dosages, better understanding of drug interactions, safer combined drug regiments, and research on novel differential uses for the drug. For example, to relieve a headache one could take Aspirin, a provably safe over the counter medication with a well-understand mechanism of action, or eat willow bark. Prior to the identification of salicylic

acid as the active ingredient which causes the analgesic effect in aspirin that quells headaches and fevers a similar but more less predictable effect was historically achieved by means of consuming willow bark which contains the same compound (Vlachojannis et. Al 2011). This method was considerably more risky as one could improperly dose oneself by consuming too much or too little bark and also risked consuming contamination from the bark. To say that we should forgo the development of explainable algorithms in favour of black box algorithms simply because they can both do a certain task reasonably well and black box systems are easier to produce is like saying that the development of Aspirin was unnecessary because willow bark does the same job of treating headaches and fevers reasonably well and is easier to produce. In doing this one would forgo ever knowing that salicylic acid is the compound which produces the analgesic effect; thus, condemning anyone who has an allergy to this compound to an unexplainable anaphylactic reaction in the same way that a black box autonomous vehicle might condemn an individual to an otherwise avoidable injury via an unexplainable algorithmic bias or design defect. I hold that developing explainable algorithm is as important for machine learning as understanding the mechanisms of action of substances is for medicine. By virtue of the overcoming the difficulties of black box programming, explainable algorithms are ultimately likely to be safer, right-respecting, and more secure. As well, there will be a considerable market advantage for projects which use explainable algorithms, considering that a number of nations have already expressed a preference for explainability (Madiega 2019).

2.10 ARE “CLEAR BOX” SYSTEMS POSSIBLE?

While there is presently no means definitively to interpret the interactions generated by learning algorithms, the claim that there is no possibility of there ever being such a method is technically fallacious. Modern autonomous vehicle technology using learning algorithms began

development in the early 1980's meaning manufacturers have already had four decades to pursue explainable A.I. initiatives. Pilotnet, an explainable end-to-end A.I. autonomous driving system, is already well into development by Nvidia (Bojarski et al. 2018). Likewise, DARPA (the U.S. Defense Advanced Research Projects Agency) has been funding explainable A.I projects since the early 1990's through the XAI program (Defense Advanced Research Projects Agency 2020). IBM is also presently spearheading methods of interpretable and explainable A.I. through its "ai explainability 360" initiatives which focuses on translating all methods of algorithmic learning into interpretable systems (IBM Research Blog 2018). Explainable A.I. is not impossible. The real question is whether the extensive time and resources needed to develop fully explainable A.I. is a reasonable burden to require autonomous car manufacturers to handle upfront, and whether or not the nature of that delay inhibits technological development.

2.11 SUGGESTIONS FOR AUTONOMOUS VEHICLE REGULATIONS

In summary, this chapter has demonstrated a clear need to move away from black box systems and towards explainable algorithms (clear box system) for autonomous vehicles. This section will attempt to provide policy suggestions as to how this transition could be approached. There is the potential to streamline the process of making autonomous vehicle rights respecting and regulatable, while also minimizing undue commercial restrictions, by committing neither to a black box nor to a clear box model, but instead to a "grey box" model. A grey box model merges the development benefits of black box algorithms with the reliability and safety of clear box systems by having compartmentalized versions of both systems in the same system. To explain how this would work in autonomous vehicles, certain tasks deemed to be of minimal safety relevance such as a subtask which maximizes vehicle fuel efficiency would be permitted to be fully black box systems, while others which are deemed relevant to safety, such as object

recognition and response, would be required to be clear box systems. This has the effect of making the goal of explainability much more feasible by limiting excessive development costs in the short term. This need only be a short-term solution given the aforementioned evidence provided in previous sections that explainable A.I. will one day be commonplace.

Which subtasks should be considered ethically relevant subtasks? I suggest that any algorithm which involves the assessment of a human-relevant variable (such as an object recognition system which differentiates pedestrians from vehicles) should be considered ethically relevant. For example, the OEDR is a subtask which is responsible for monitoring the driving environment and autonomously executing appropriate responses to objects and events. In other words, this is the subtask of the car which makes evaluative determinations based on input and then executing said decisions as output. Object recognition, maneuvering plans of action based on environmental data, and tactical decisions such as when to initiate fallback to manual operation are all handled by the OEDR. As a result of having to recognize pedestrians and other vehicles and prepare responses to their actions the vast majority of the OEDR subtasks involve human variables. To supplement the "gaps" which will be present in this model of explainability "data recorders" similar to those commonly used in the aviation industry should be installed on all autonomous vehicles. The reason for this is that if a record of all the sensor data, DDT data, user inputs, and driving decision outputs is available it will allow for the possibility of simulation and diagnostic procedures to determine "what went wrong". The result of these measures will give investigators, courts, and policymakers a meaningful view into the operation of the vehicle. This view will be similar to that kind of transparency currently demanded by aviation authorities investigating plane crashes.

CHAPTER 3

AUTOMATION, AUTONOMY, AND ACCOUNTABILITY

3.0 CURRENT LEGISLATION

This chapter focuses on problems of autonomous vehicle liability, namely the improper distribution of autonomous vehicle liability. A trend in legislation across multiple international jurisdictions has seen autonomous vehicles treated in a similar manner to traditional manual vehicles. For example, California, the American state where several major autonomous vehicle manufacturers including Tesla and Google are headquartered, has no specific laws regarding autonomous vehicle owner liability (Baker et Al. 2020, pg. 14). Similarly, under German law, if the driverless car causes death, personal injury or property damage, the owner of the vehicle will be liable as if they had been driving the vehicle themselves (Baker et Al. 2020, pg. 14). In effort to promotes autonomous vehicle development and adoption Ontario has taken a perspective similar to that of California and Germany by actively pushed forth legislation which states:

> *"A human driver is required at all times to take back the driving task when alerted to do so by the vehicle. Drivers will need to be in full care and control of vehicles with SAE Level 3 technology and all existing laws (such as distracted, careless and impaired driving laws) will continue to apply to drivers of these vehicles. Drivers are responsible for the safe operation of these vehicles at all times." (Ontario Ministry of Transportation 2013)*

But is the human occupant of an autonomous vehicle really "driving" when the majority of the vehicle tasks are fully automated? The UK Law Commission suggests that this question warrant a new category of vehicle operator referred to as a 'user of a highly operated vehicle'

("user-in-charge"). It also recommends that the 'user-in-charge', "should not be considered a 'driver' while the vehicle is driving itself and legislation must develop to clarify the role of a 'user-in-charge'" (Baker et Al. 2020, pg. 14). I hold that the issue of autonomous vehicle accountability is primarily an issue of ambiguity as to who the operator of the vehicle is. This problem is further complicated by differing levels of autonomous capability which alter the obligations of the operator. For example, a level two autonomous vehicle requires the human operator of the vehicle to be alert and actively participate in operating the vehicle at all times, whereas a level four autonomous vehicle functions largely independently, only requiring operator involvement in extremely select circumstances which are outside of the capabilities of autonomous technology in the vehicle. This ambiguity is particularly apparent when talking about level three autonomous vehicles as they include driving features which are fully autonomous with one's that require human operator intervention, and those which require driver intervention may actually vary from manufacturer to manufacturer. As laws regarding autonomous technology lacks specificity to a point of often not explicitly denoting what kinds of actions warrant the attribution of responsibility to the operator of the vehicle, the manufacturer of the vehicle, and external actors such as pedestrians or other vehicles.

This chapter aims to outline a way of thinking about how to determine when an operator of an autonomous car is able to make "free choice" in the form of a driving decision and when that operator is at the mercy of pre-programmed algorithmic "forced choices" made for them by autonomous vehicle manufacturers. This distinction is a particularly important as it has significant implications for accountability. For example, instances of operator "fall-back" (situations in which an automated vehicle hands control back to a human driver) require an individual to take over the majority of dynamic driving tasks of an autonomous vehicle on very

short notice. If an autonomous vehicle's programming caused it to become involved in a dangerous situation should the driver be responsible for the consequences simply because a "fall-back" protocol was engaged at the last second? What if the driver is not given adequate time to respond and the sudden nature of the fall-back causes an otherwise manageable situation to become dangerous? In this chapter, I will first focus on identifying instances of individual agency and shared agency in order to identify a clear means to identify when actors are making decisions independently and in tandem. This argument will then be supplemented by a discussion regarding when a choice made by an agent may be considered a voluntary or free choice, against when a choice may be considered involuntary or forced. I aim to demonstrate that whether or not an individual may be justifiably held accountable for the consequences of an autonomous vehicle collision, accident, or illegal behavior is determined by whether or not they are freely making choices and acting collaboratively with the algorithm of the vehicle.

For the sake of clarity, in the following arguments I will be using terms to refer to specific realms of responsibility typically associated with autonomous vehicle operation. Recall that, the SAE defines "dynamic driving tasks" as all of the real-time operational and tactical functions required to operate a vehicle in on-road traffic such as lateral and longitudinal vehicle motion control (steering, acceleration, deceleration, and braking), recognizing and monitoring objects and events in the driving environment, and both preparing and executing responses to said objects and events by means of maneuvering, signalling, or enhancing conspicuity via lighting (SAE 2018, pg. 6-7). I will define the "operator/operators" of the vehicle as the party responsible for DDT decisions not made by the autonomous vehicle system such as setting a destination. The operator/operators are considered the party responsible for the outcomes of those DDT decisions. By contrast, "passengers" are merely other parties which are present in the

vehicle. They do not make decisions regarding dynamic driving tasks and are not responsible for any outcomes which occur as a result of said decisions.

3.1 INDIVIDUAL AGENCY & SHARED AGENCY

The primary project for this section is to discern a morally useful description of shared agency for the operation of autonomous vehicles. In order to determine when an individual is responsible for an act, one must first assess what constitutes the capacity to make a voluntary choice. In order to be able to make a choice one must possess agency, namely the capacity to act intentionally (Schlosser 2019). The average adult human, for example, is functionally capable of acting intentionally and is therefore capable of exercising agency. According to the event-causal theory of agency, when agency is exercised it consists of the instantiation of the right causal relations between agent-involving states and events, namely an agent is causally linked to a state of affairs (Schlosser 2019). Such agency is instantiated by, for instance, a human vehicle operator in a manual vehicle seeing an obstruction and intentionally avoiding it by manipulating the vehicle controls. In such a situation the human operator's actions are in line with their intentions, making said actions a manifestation of individual agency.

The conditions of shared agency differ from individual agency in that despite both individuals having a hand in bringing about a state of affairs the intentions of individuals are not always entirely aligned. Bratman argues that, for an act to be considered a manifestation of shared agency the sub-plans which bring about the main objective must also be shared (Bratman 1993, pg. 106). A sub-plan is most accurately described as the steps or method to bring about a certain goal. This pertains to shared agency for Bratman in the sense that if two agents have the same goal but different sub-plans, they are not partaking in shared agency but merely acting independently towards the same objective. One could argue that two movers, moving a couch up

a flight of stairs, both working towards bringing about a certain state of affairs (the couch being placed in a second story room) can each act intentionally towards a common end goal and are therefore can potentially partake in shared agency if they agree to a shared sub-plan as to how to accomplish this task. For example, suppose one of the movers, upon seeing an upcoming curve in the staircase, begins to pivot the couch without informing the other mover. What allows this to continue being an instance of shared agency is that in cases of conflict the movers need to consent to a shared solution to re-align their sub-plans. In this case, one mover might communicate to the other might begin to pivot as well. But suppose the second mover believes that pivoting the couch will cause it to get stuck and therefore refuses to do so. If the first mover threatens to drop the couch and crush the second mover if he does not pivot the two movers are no longer sharing sub-plans, and thus no longer partaking in shared agency. If the couch then does become stuck as a result of pivoting, we should hold the first mover responsible but not the second. Similarly, this sub-plan distinction identifies circumstances relevant to the operation of autonomous vehicles in which an actor is unable to make a free choice. For example, in "fallback" cases an autonomous vehicle, upon encountering a situation it cannot properly process, transfers full control of the vehicle back to the human operator of the vehicle. If this fallback occurs too suddenly or in an unsafe manner, the human operator may not have the free choice to properly choose a safe course of action. In such situations the operator does not have a shared sub-plan with the autonomous vehicle and therefore cannot be considered to be participating in a form of shared agency.

My definition of shared agency is summarized as involving a combination of sharing intentions, acting with other in a collaborative way, having shared plans and method, and working towards a common goal. According to this definition, a human operator of an

autonomous vehicle is no longer partaking in "shared agency" if the vehicle behaves in a way clearly contrary to the human operator's goals and intentions. For example, if autonomous assistive technology malfunctions and forces a car to veer into a telephone pole, the human in the operator is not responsible for the negative outcomes of the vehicles behavior. But this does not imply that the vehicle itself is somehow "responsible" for committing a wrongful act or making a bad decision as a vehicle does not meet the criteria for agency. What I will argue in the next section is that the manufacturer, by means of its causal link to the algorithmic determinations of the vehicle is partaking in shared agency with consumers of its products. This shared agency makes the manufacturer morally liable for improperly tuning the algorithm to adequately respond to road conditions. Much in the same way that a structural engineering firm is held responsible if a building collapses due to foreseeable structural issues I hold than an autonomous car manufacturer should be held responsible for foreseeable algorithmic design flaws.

3.2 RESPONSIBILITY

To demonstrate why, for example, Ontario's hardline stance on autonomous vehicle accountability is morally questionable, consider that a level 3 vehicle operating within its operational design domain (ODD) is the lowest level of autonomous vehicle that could be thought of as a fully automated autonomous driving system (ADS). The only real difference separating a level 5 vehicle ADS system from a level 3 vehicle ADS is an expanded operational design domain and better reliability; in particular, when driving within the level 3 vehicle's operational design domain the driver will functionally have as little input as the level 5 vehicle. During normal operation, despite the formal requirement of human driver vigilance, such a system should rarely need any intervention from the "driver" beyond the inputting of a destination. Other than that singular act of destination selection there is no further contribution of

the human “driver” to the operation of the vehicle. In light of the legislation discussed above, if such a level 3 autonomous vehicle were to get into a collision or disobey the law the “driver” would be held responsible despite having no part in the actions which caused it. In a way, holding a driver responsible is as questionable as charging a passenger of a limousine for the poor driving of their chauffer, merely because they used the services of the limousine to travel to a certain destination. This means that the human operators of autonomous vehicles are presently at risk of being unjustifiably punished for acts over which they had no control and could not have reasonably predicted. Meanwhile, the manufacturers who designed the vehicles and who are the direct cause of the vehicle’s programming, which make them more qualified to reasonably predict and plan for negative vehicle behavior, are not likely to be adequately held to account.

This section will focus primarily on defining a particular kind of responsibility referred to as “liability responsibility” which determines when a party may be considered “morally responsible for an act”. The means by which to begin approaching this problem is first to note, as the UK law commission did previously, that with regard to autonomous vehicles the term “driver” is vague. When the term “driver” is used colloquially to refer to a person controlling a conventional, fully manual vehicle, it is used in the sense that said “driver”, barring mechanical malfunction, is the causal source of the behavior of the machine and is therefore the agent of said actions. As the agent of the vehicle’s actions, the “driver” is held to be responsible for these actions. The problem is that this model becomes questionable as soon as any portion of the “driving” tasks is automated, as the “driver” is potentially no longer the sole source of actions in this model. Consider even single-task vehicle automations such as an automatic transmission. If an automatic transmission were, due to a design defect, to start shifting gears improperly while going uphill due to a design defect this could cause the vehicle to stall, roll backwards, and cause

a collision with the car immediately behind it. In such a case no action the human "driver" took contributed to the collision; it is the result of design choices (or defect in manufacturing) made by the vehicle manufacturer.

To illustrate my point, consider the case of unintended acceleration issues which affected multiple Toyota vehicles from 2000-2010. There were multiple design issues, ranging from "sticky" accelerator pedals to bizarre engine surges when the brake pedals were pressed, which caused unintentional accelerating behavior in the vehicles and led to multiple collisions (Safety Research & Strategies 2015). To argue for consumer strict liability for merely choosing to own a Toyota vehicle and driving it on the road is not defensible. The design defects which caused unintentional behavior were the direct result of faults in the manufacturer's design process, thus, making the manufacturer the liable party. If instead of these physical design flaws this unintended accelerating behavior was caused by an autonomous system improperly reacting to traffic conditions, why should the liability of the human operator change? I hold that flaws in autonomous driving algorithms are no different than flaws in physical design features.

A critic of my position could argue that an autonomous vehicle differs from a chauffer-driven limo in the sense that technically the owner of the autonomous vehicle made the decision to buy the vehicle, to assume the risk of operating the vehicle, and to introduce the vehicle to risky environments like roadways, and that this does indeed make them responsible for any consequences which arise in relation to the vehicle. This form of strict liability is similar to what applies to horseback riders who assume full responsibly for any reasonably foreseeable issues caused by their animal when they travel on public roadways (Ross 2019). But an autonomous vehicle is not an animal which is reliant on its owner for training and guidance to behave safely on a public road. An autonomous vehicle is designed to operate in a certain way by a

manufacturer meaning that failure to operate in a safe manner in foreseeable road conditions is more accurately defined as a design flaw. Assuming that a vehicle owner carries out regularly-scheduled maintenance of their autonomous vehicle and follows service notifications and guidelines, there is no reason why they should be held liable for design defects and issues of manufacturing. As owners of non-autonomous vehicles are not held liable for outcomes caused by manufacturer defects, this strict-liability argument would not apply. It would be bizarre to attribute liability to a vehicle owner who gets into an accident because a design defect in the vehicle causes their brakes to malfunction.

3.3 CONCLUSION

The ultimate conclusion of this chapter is that the human operator of the vehicle should only legally be held "responsible" for driving decisions which they are actually controlling and are intentionally partaking in. While there may be specific situations in which a human operator's action or inaction may lead to negative outcomes, and hence in which the human operator might bear some portion of moral responsibility, my position is that the standard of liability from such cases must be higher than simply being present in an autonomous vehicle. And simply being present seems to be enough to warrant legal liability, under for example Ontario's legislation. As the duties of autonomous vehicle operators become more clearly defined, a means to assess this contributory negligence will have to develop and will most likely be based around responsibilities like adhering to regular maintenance of the vehicle, safely assuming manual control of select DDT functions when necessary, and others.

In order for one to be considered to be partaking in shared agency I have argued that one must share a common goal and common subplans pertaining to how to accomplish that goal. When an autonomous vehicle violates these conditions by behaving erratically or unpredictably

as a result of a manufacturer defect there is no possibility of shared agency and therefore no possibility of human operator responsibility. Building upon this conclusion, I argue that from a policy perspective, for all functions in which the autonomous vehicle offers automation of dynamic driving tasks and takes on the role of an "operator", the burden of legal responsibility should be attributed to the parties which manufacture the autonomous vehicle and its software.

CONCLUSION

4.0 OVERVIEW

The primary goal of this thesis was to address what I hold to be the most pressing ethical issues in autonomous vehicles. These include, the applicability of moral philosophy (via the Trolley Problem) to guiding autonomous vehicle regulation, the need for "explainability" in the machine learning algorithms that guide autonomous vehicles, in order to ensure that vehicles are fair and rights-respecting, and the need for a change in regards to distribution of liability for driving behavior.

4.1 CHAPTER SUMMARIES

In chapter one I demonstrated how the incorporation of machine learning algorithms into autonomous vehicles continues to present an avenue for the application of moral philosophy. Using the work of Janet Fleetwood, I explained how the operational design domain of autonomous vehicles had increased to such a degree that there was no possible means for a programmer to anticipate every single possible interaction between the car, the driver, and the environment. This necessitates the integration of models for the prioritization of objectives. These models were shown to draw directly on conventional moral philosophy to build highly applicable objective prioritization structures, solidifying my assertion that moral philosophy is directly applicable to the programming of autonomous vehicles.

In chapter two I demonstrated that black box algorithms operate in a way that is opaque to regulators, industry experts, and even the programmers of the algorithm themselves. I then provided several examples as to how such a system can be biased to easily and undetectably infringe on individual rights. In doing this, I also demonstrated how this issue, paired with the

widespread adoption of autonomous vehicles, could cause widespread rights violations. I then presented a potential compromise in the form of a "grey box" system paired with operational data recorders. This solution attempted to address the aforementioned ethical issues of black box systems while also substantially mitigating the potential ethical shortcomings associated with black box systems.

In chapter three I asserted that laws regarding liability must be altered in order to keep pace with the changes in driver responsibilities brought about by autonomous vehicles. In this chapter I demonstrated that drivers of autonomous vehicles exhibited only a negligible contribution to their vehicle's behavior outside of select situations. In demonstrating this, I also provided arguments as to why illegal or unsafe algorithmic autonomous vehicle behavior should be considered a manufacturer defect. Using Michael Bratman's framework I also determined that the operator of an autonomous vehicle is not partaking in shared agency when the vehicle behaves in a manner which is contrary to the operator's expectations. I concluded this chapter by arguing that autonomous vehicle manufacturers should included in distribution of liability for the actions of autonomous vehicles.

4.2 LIMITATIONS

Many could argue, like Noah Goodall that there are other potential sources of regulatory inspiration for autonomous vehicles, beyond moral theory, such as legal theory, engineering standards, and legislative action by governments. However, my position is not that moral philosophy should be the sole source for the generation of new regulation, but merely that it should be consulted given its demonstrated relevance. My point is that there no reason why engineers and politicians should be trying to re-invent the wheel and develop a new isolated ethics by ignoring pre-existing work on moral philosophy. Given that moral philosophy has been

easily integrated into multiple practical fields such as medicine, science, and law with great success there is little reason as to why engineering cannot stand to benefit from it as well.

There is also the possibility that one could dismiss my determinations regarding black box algorithms on the basis that there are many types of such algorithms and my thesis does not address specific technical matters pertaining to each. However, I should point out that any system which fits the description of black box algorithms in terms of its opaque methodology suffers from the same ethical issues I outlined when operating in an operational design domain containing human actors. As well, my claims regarding explainability are based on the limitations of current technology meaning that if such algorithms develop into or are reformulated into systems which can be explained sufficiently to be regulated fairly my arguments would not apply to such systems. This means that if new technology emerges which enables different kinds of explainability and interoperability, then my view on the most prudent course of action regarding autonomous vehicle regulation could definitely change.

Finally, my claims regarding liability are claims of moral responsibility rather than precise legal claims. I acknowledge that there are other factors which may contribute to legal liability being attributed differently from attribution of moral responsibility. However, my aim here was to provide the means to attempt to consolidate the two in order to rectify what was shown to be unjust distributions of liability. In this case my response to such an objection would be to question why an algorithmically determined behavior is not the same a design defect which is already recognized in multiple jurisdictions as a factor in determinations of legal liability.

4.3 CONCLUDING REMARKS

My hope is that the ideas discussed in this thesis will be explored further as autonomous vehicle technology becomes more widespread. Given the vast potential for this technology to greatly improve lives and transportation safety it is undoubtedly only a matter a time before the majority of the vehicles on the road have significant autonomous capabilities. If the ethical concerns outlined in this thesis can be addressed as the technology develops, modern society may one day reach the ideal of the Phaeacian ship, a completely safe, predictable, and fully autonomous vehicle, one that not only serves the interests of the driver well, but also the interests of society more generally.

BIBLIOGRAPHY

1) Baker, Steven, et al. “Connected and Autonomous Vehicles – a Cross-Jurisdictional Comparison of Regulatory Developments.” *White & Case*, 2020.
2) Bickenbach, Jerome, "Disability and Health Care Rationing", The Stanford Encyclopedia of Philosophy (Spring 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2016/entries/disability-care-rationing/>.
3) Bonnefon, J.-F., et al. “The Social Dilemma of Autonomous Vehicles.” Science,
4) vol. 352, no.6293, 2016, pp. 1573–1576., doi:10.1126/science.aaf2654.
5) Bojarski, Mariusz, et al. “Explaining How End-to-End Deep Learning Steers a Self-Driving Car.” *NVIDIA Developer Blog*, 4 Sept. 2018, devblogs.nvidia.com/explaining-deep-learning-self-driving-car/.
6) Bratman, Michael E. “Shared Intention.” *Ethics*, vol. 104, no. 1, 1993, pp. 97–113., doi:10.1086/293577.
7) Bushwick, Sophie. “How NIST Tested Facial Recognition Algorithms for Racial Bias.” *Scientific American*, Scientific American, 27 Dec. 2019, www.scientificamerican.com/article/how-nist-tested-facial-recognition-algorithms-for-racial-bias/.
8) “Defense Advanced Research Projects Agency.” *Defense Advanced Research Projects Agency*, 2020, www.darpa.mil/program/explainable-artificial-intelligence.
9) Ehmer, Mohd, and Farmeena Khan. “A Comparative Study of White Box, Black Box and Grey Box Testing Techniques.” *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 6, 2012, doi:10.14569/ijacsa.2012.030603.
10) Fleetwood, Janet. “Public Health, Ethics, and Autonomous Vehicles.” American Journal of Public Health, vol. 107, no. 4, 2017, pp. 532–537., doi:10.2105/ajph.2016.303628.
11) Foot, Philippa (1967). “The Problem of Abortion and the Doctrine of Double Effect”. Oxford Review 5: 5-15. Reprinted VV 19–32.
12) Goodall, Noah J. “From Trolleys to Risk: Models for Ethical Autonomous Driving.” *American Journal of Public Health*, vol. 107, no. 4, 2017, pp. 496–496., doi:10.2105/ajph.2017.303672.
13) Hotten, Russell. “Volkswagen: The Scandal Explained.” *BBC News*, BBC, 10 Dec. 2015, www.bbc.com/news/business-34324772.
14) “Introducing AI Explainability 360.” *IBM Research Blog*, 19 Sept. 2019, www.ibm.com/blogs/research/2019/08/ai-explainability-360/.
15) Leggett, Christopher. *THE FORD PINTO CASE: THE VALUATION OF LIFE AS IT APPLIES TO THE NEGLIGENCE-EFFICIENCY ARGUMENT*. 1999, users.wfu.edu/palmitar/Law&Valuation/Papers/1999/Leggett-pinto.html.
16) Madiega, Tambiama. “EU Guidelines on Ethics in Artificial Intelligence: Context and Implementation.” *EU Guidelines on Ethics in Artificial Intelligence: Context and Implementation*, 2019.
17) Marshall, Aarian. “What Can the Trolley Problem Teach Self-Driving Car Engineers?” *Wired*, Conde Nast, 25 Oct. 2018, www.wired.com/story/trolley-problem-teach-self-driving-car-engineers/.
18) National Assembly of Québec. “Bill 165, An Act to Amend the Highway Safety Code and Other Provisions - National Assembly of Québec.” *Bill 165, An Act to Amend the Highway Safety Code and Other Provisions* , 18 Apr. 2018,

www.assnat.qc.ca/en/travaux-parlementaires/projets-loi/projet-loi-165-41-1.html?appelant=MC.
19) Ontario Ministry of Transportation. “Automated Vehicles – Driving Innovation in Ontario.” Automated Vehicles – Driving Innovation in Ontario, 25 Oct. 2013, www.mto.gov.on.ca/english/vehicles/automated-vehicles.shtml.
20) Ross, Alex. “Autonomous Vehicles In Canada: Are Liability Rules Being Affected By Horses, Elevators And Autopilots? - Transport - Canada.” *Articles on All Regions Including Law, Accountancy, Management Consultancy Issues*, Gowling WLG, 25 July 2019, www.mondaq.com/canada/rail-road-cycling/828982/autonomous-vehicles-in-canada-are-liability-rules-being-affected-by-horses-elevators-and-autopilots.
21) SAE International. “Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles.” *SURFACE VEHICLE RECOMMENDED PRACTICE J3016*, June 2018, pp. 1–35., doi:10.1149/2.0031815jes.
22) Schlosser, Markus, "Agency", The Stanford Encyclopedia of Philosophy (Winter 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2019/entries/agency/>.
23) Silver, David, et al. “A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play.” *Science*, vol. 362, no. 6419, 2018, pp. 1140–1144., doi:10.1126/science.aar6404.
24) Slotnick, David. “Nearly a Year after It Began, the Boeing 737 Max Crisis Still Drags on. Here's the Complete History of the Plane That's Been Grounded since 2 Crashes Killed 346 People 5 Months Apart.” *Business Insider*, Business Insider, 5 Mar. 2020, www.businessinsider.com/boeing-737-max-timeline-history-full-details-2019-9.
25) “Toyota Sudden Acceleration Timeline.” *Safety Research & Strategies*, 2015, www.safetyresearch.net/toyota-sudden-acceleration-timeline.
26) Vivek, Skanda, et al. “Cyberphysical Risks of Hacked Internet-Connected Vehicles.” *Physical Review E*, vol. 100, no. 1, 2019, doi:10.1103/physreve.100.012316.
27) Vlachojannis, J., et al. “Willow Species and Aspirin: Different Mechanism of Actions.” *Phytotherapy Research*, vol. 25, no. 7, 2011, pp. 1102–1104., doi:10.1002/ptr.3386.