EMBEDDING ETHICAL THEORY INTO AUTONOMOUS VEHICLES

ANALYZING A CONVERGENCE OF ETHICS AND ACTION

Austin Atchley

TC 660H
Plan II Honors Program
The University of Texas at Austin

May 2020

Sarah Abraham
Digitally signed by Sarah Abraham
Date: 2020.05.05 14:57:40 -05'00'

Sarah Abraham, Ph.D.
Department of Computer Science
Supervising Professor

Scott Stroud
Digitally signed by Scott Stroud
Date: 2020.05.05 13:21:32 -05'00'

Scott Stroud, Ph.D.
Department of Communication Studies
Second Reader

Abstract

Author: Austin Atchley

Title: Embedding Ethical Theory Into Autonomous Vehicles: Analyzing a Convergence of Ethics and Action

Supervising Professor: Sarah Abraham, Ph.D.

The ability of autonomous vehicle systems to collect data and, independently of passion, make split-second decisions creates a newly emerging phenomenon: developers encode their ethical views into something that will (almost deterministically) enact them in reality. This is a departure from the traditional relationship between belief and action that is present in each decision made by a human, and, as artificial intelligence becomes more widespread, perhaps a majority of day-to-day decisions will be made without human unpredictability. Instead, decision-making will boil down to the application of ethical principles (either explicitly or implicitly) to a dataset. In creating ethical autonomous vehicles, we should address the problem using both top-down and bottom-up approaches to encoding ethical beliefs. The top-down approach uses pre-defined ethical values to drive decision-making processes, and the bottom-up approach attempts to mimic human ethical conduct by using a pre-collected dataset of decisions made by real human drivers. Currently, many developers are not aware of the ethical principles they embed into their code, and by explicitly taking this approach to ethical decision-making, we can encourage morally preferable decisions to be taken by autonomous vehicles. These issues are inherently interdisciplinary, and this thesis treats them as such, borrowing from both engineering and philosophical discourse. I argue that a similarly holistic perspective should be adopted by scholars working on the many-faceted topic of autonomous driving.

Acknowledgments

To my supervisor, Dr. Sarah Abraham, and my second reader, Dr. Scott Stroud—thank you for your guidance over the course of this project. Your assistance helped me shape my topic, through both your classes and our meetings. Thank you to Prof. Joydeep Biswas and Prof. Peter Stone, as well as the various other Computer Science professors who gave me the technical background needed to write this thesis.

To my friends and family—thank you for your unwavering support, and thank you for listening to me talk about this project incessantly over the last nine months. In particular, I would like to express my gratitude to Kierra Nguyen, who served as a constant editor and source of inspiration, and Logan Zartman, who spent many long nights developing autonomous driving software with me, allowing me to gain a deeper understanding of this topic.

To those listed here and many others—thank you again for your support. I could not have completed this project without you, especially in the unusual circumstances forced on us by the global pandemic.

Table of Contents

Chapter One: Introduction

In 2020, self-driving cars are already on the street. These cars are made up of complex systems, representing the forefront of technological advancements in artificial intelligence and robotics. The fact that these cars are functioning is a testament to how far we have come, but we know that autonomous vehicles are not flawless. Even though autonomous driving technology is, on average, already safer than human drivers, there is always a chance for error in complex real-world scenarios. One of the most important goals in designing and building autonomous vehicles is to fail as rarely as possible.

About 93% of the 5.5 million crashes in the U.S. have been attributed to human error (Gogoll and Müller, 2017). Autonomous driving technology would already greatly reduce this figure if deployed in scenarios it can safely navigate today, and the technology will only continue to advance. Nevertheless, when autonomous vehicles do fail, we must also ensure that they do so as gracefully as possible. In this thesis, I will explore the moral considerations we must make while introducing autonomous vehicles into our society from philosophical and technical standpoints, synthesizing an interdisciplinary investigation into the machines that will soon have a large impact on our world. Here, I argue that this interdisciplinary approach is not only the most appropriate approach to analyzing topics in autonomous driving, but it is essential to consider both the philosophical and technical implications of these problems.

The development of new technology has historically demanded new ethical considerations. As a new piece of technology enters society, the onus is placed on thinkers

across disciplines to form a strategy that prevents harm. Yet no matter how often this pattern has presented itself throughout history, the gap between technological advancement and the ethical principles governing these advancements has grown wider. Recently, the pace at which technology is developed has turned exponential, while the pace at which we form new ethical strategy remains constant. Concerns for safety and societal benefits have always been at the forefront of engineering, but today's systems are approaching a level of complexity that requires these systems to take on a new set of responsibilities. The amount of software used in cars is growing by a factor of 10 every 5 to 7 years, and some car manufacturers are, in a sense, becoming software companies (Holstein and Dodig-Crnkovic, 2018).

Specifically, the recent development of autonomous systems (the most apparent example being self-driving cars) has created an entirely new kind of ethical dilemma: how should these systems deal with ethical decisions traditionally faced by individuals? It seems to be the case that, if we want to make any sort of intelligent decision when faced with an ethical decision, ethical principles need to be embedded into these autonomous systems. These ethical decisions will be deterministically replicated across multitudes of agents running the same software. As the human race offloads more decisions to autonomous systems, the ethical beliefs embedded in systems' code will be enacted absent of human psychological factors. Ethical belief has always guided how people act, but it has never completely dictated a person's action.

Because autonomous driving software is already being used on streets across the world, this subject is no longer theoretical. The topics discussed in this thesis will become increasingly more important as more cars become autonomous and semi-autonomous cars become more autonomous, but lives are already in danger if something goes wrong. For this reason, car manufacturers, autonomous vehicle software creators, and government agencies must ensure the best possible safety guarantees on

these machines. Because the trickiest dilemmas involving autonomous vehicles take place at the software level, I will primarily take the perspective of software creators. Software implementation issues are inseparable from ethical issues in the broad 'social dilemma' of autonomous vehicles. This is an intrinsically interdisciplinary topic, and I will treat it as such.

MIT's Moral Machine experiment demonstrates one of the critical decisions with which we hope to soon trust autonomous vehicles — an applied version of the Trolley Problem (Awad et al., 2018). Imagine a scenario in which a car is moving toward a tunnel at 60 miles per hour. A child trips and falls onto the road. There is not enough room to brake, but the car can swerve into the tunnel wall — thereby killing the driver. Should the car save the passenger or the pedestrian? This is a classic thought experiment in ethics, and its application to self-driving cars is logical, but perhaps not the most fitting. Such a scenario is a rare occurrence, and nobody can be sure of the results of either decision, but self-driving cars must have an answer to even the most impossible of questions. Nevertheless, the decisions made by autonomous vehicles go outside the context in which the technology was developed and tested. Programmers implicitly embed ethical programming into the systems they create, sometimes employing blanket strategies in situations where humans would apply specific reasoning.

Moreover, ensuring anything is difficult when one has to account for so many moving parts, uncontrollable external factors, and inconsistencies in both hardware and software systems. The noisiness of sensors is worrying considering the precision at which autonomous driving systems need to operate. I will analyze how relevant technical autonomous driving systems are implemented in the latter part of this thesis, and specifically, I will focus on the technological issues involved in making self-driving cars themselves into moral agents. We have come to trust digital technologies with

our lives in many aspects of every-day life, but this trust should not be blind. Software breaks. We should not expect anything different when applying advanced technology to cars, even if these systems are, on average, more reliable than humans.

1.1 Background

This thesis will use the terms 'autonomous vehicles,' AVs, and 'self-driving cars' interchangeably, but many of the same concepts could also be applied to autonomous motorcycles or semi-trucks. The systems I will examine here lie beneath the red line in Figure 1.1, meaning that an autonomous system both controls the car and monitors its environment. When I use the term "autonomous systems," I am referring to intelligent software systems that form and enact decision-making processes, as are used on autonomous vehicles.

We can use Figure 1.1 as a schema for reasoning about the levels of autonomy in a system. For example, systems at SAE level 1 (e.g., traffic cruise control, lane assist) allow autonomous systems the lowest level of control, and thus, do not cause many of the ethical dilemmas I will be analyzing. Of course, lane assist could malfunction and cause someone to get injured, but in such scenarios, a human driver is expected to pay attention, taking control if necessary. If something goes wrong, the human driver will be held responsible. In situations where control of the car is out of human hands, things get more complicated. If we hope to have fully autonomous vehicles, we must be able to fully trust them with critical decisions, like safely avoiding hazards at highway speeds or choosing to hit obstacles when avoidance might mean passenger death. If things go wrong without a driver, it becomes difficult to determine who should be held responsible.

The Defense Advanced Research Projects Agency (commonly known as DARPA) held three autonomous driving challenges between 2003 and 2007 (J. M. Anderson

SAE level	Name	Narrative Definition	Execution of Steering and Acceleration/ Deceleration	*Monitoring* of Driving Environment	Fallback Performance of *Dynamic Driving Task*	System Capability *(Driving Modes)*
***Human driver* monitors the driving environment**						
0	**No Automation**	the full-time performance by the *human driver* of all aspects of the *dynamic driving task*, even when enhanced by warning or intervention systems	Human driver	Human driver	Human driver	n/a
1	**Driver Assistance**	the *driving mode*-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 *human driver* perform all remaining aspects of the *dynamic driving task*	Human driver and system	Human driver	Human driver	Some driving modes
2	**Partial Automation**	the *driving mode*-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 *human driver* perform all remaining aspects of the *dynamic driving task*	**System**	Human driver	Human driver	Some driving modes
***Automated driving system* ("system") monitors the driving environment**						
3	**Conditional Automation**	the *driving mode*-specific performance by an *automated driving system* of all aspects of the dynamic driving task with the expectation that the *human driver* will respond appropriately to a *request to intervene*	System	**System**	Human driver	Some driving modes
4	**High Automation**	the *driving mode*-specific performance by an automated driving system of all aspects of the *dynamic driving task*, even if a *human driver* does not respond appropriately to a *request to intervene*	System	System	**System**	Some driving modes
5	**Full Automation**	the full-time performance by an *automated driving system* of all aspects of the *dynamic driving task* under all roadway and environmental conditions that can be managed by a *human driver*	System	System	System	**All driving modes**

Figure 1.1 SAE levels of driving automation (SAE International, 2018)

et al., 2016). These challenges worked to rapidly accelerate the pace at which autonomous driving technology was developed. The first challenge had no winners, but the second and third challenges were major milestones in the development of autonomous driving technology. The DARPA Urban Challenge, which took place in 2007, produced many of the foundational research papers in autonomous driving research. As a result of these advancements, companies developing self-driving car technology are becoming increasingly prevalent. While this is extremely exciting from a technical standpoint, we must remember that with more self-driving cars comes increased centralized risk. Driving is an intrinsically dangerous task. We can generally trust human drivers, but imagine if every Toyota driver who downloaded the latest daily car update became an overly aggressive road-rager, or if they became your 100-year-old grandmother who can barely see over the steering wheel. This is

foreseeable if self-driving car creators do not use caution with the code they send to their products.

1.2 Related Work

Recently, there has been a surge of academic work dealing with the ethical aspect of autonomous vehicles. Since the DARPA challenges, there has been a good deal of speculation on how autonomous vehicles will change our world, but most of this has taken place outside the realm of academia. In 2014, Goodall published "Machine Ethics and Automated Vehicles" (Goodall, 2014), cementing the worries with autonomous vehicle accident-avoidance software presented in non-academic settings. In 2015, Lin published the highly influential "Why Ethics Matters for Autonomous Cars" (Lin, 2015). These two works presented many of the issues that scholars have been trying to solve for the following 6 years, such as considering the Trolley Problem's similarity to autonomous vehicle accident-avoidance (Holstein and Dodig-Crnkovic, 2018), creating practical algorithms based on theoretical ethical frameworks (Leben, 2017), (Gerdes and Thornton, 2015), and allowing passengers to customize the ethical behavior of their vehicle (Contissa, Lagioia, and Sartor, 2017). These papers also hint at issues that other scholars have expanded upon, such as collective decision-making (Nyholm and Smids, 2016), collective responsibility (Liu, 2017), and differences between cultures in ethical decision-making (Gold, Colman, and Pulford, 2014).

In 2016, Bonnefon, Shariff, and Rahwan published "The social dilemma of autonomous vehicles" (Bonnefon, Shariff, and Rahwan, 2016), gathering even more attention. Bonnefon, Shariff, and Rahwan focus on the psychological dilemma in autonomous vehicles, showing that this issue is inherently interdisciplinary, and inviting analysis from alternative perspectives. Thus far, scholars have looked at this issue from the perspective of traditional ethics — drawing from Utilitarianism, Kant, and

Rawls — (Mill 1861/2012), (Kant, 1785/2002), (Rawls, 1971), policy-making (J. M. Anderson et al., 2016), (Maurer et al., 2016), (Matthias, 2004), and human-machine interaction (Floridi, 2017), (Taddeo and Floridi, 2018), (Taddeo, 2017).

However, none of this work would exist if it were not for the technical advancements that brought us to this point. Many algorithmic techniques that are employed on autonomous vehicles are published in technical papers, but the full software products that run autonomous vehicles are siloed on computers in private research labs. The technical papers published from the DARPA competition and the papers detailing autonomous vehicles' accidents are the exceptions to this rule. Autonomous car creators wish to prevent mistakes from being made again, so papers about the DARPA low-speed crash (Fletcher et al., 2009) and the Uber Arizona crash (National Transportation Safety Board, 2018) were made public. Architecture and technical details from teams Junior (Montemerlo et al., 2009), Boss (Urmson et al., 2008), Talos, and Skynet (Fletcher et al., 2009) from the DARPA Urban Challenge were also made available.

Much of the related work takes the perspective of a single discipline. Here, I argue that technical researchers should further consider the ethical ramifications of their creations, and philosophical researchers should be better acquainted with the systems they are writing about. While this might be uncomfortable, analyses of autonomous driving should take a holistic approach to the topic if they wish to circumvent the catastrophes that can result from autonomous vehicles' failures.

1.3 Outline

In Chapter 2, I begin by examining programming as an ethically-charged — and even ethically prescriptive — action. With this moral perspective on software in mind, Section 2.3 analyzes which decisions should be allowed to be made by machines and

why. Next, Section 2.2 gives an analysis of determinism in artificial intelligence and how it should change how we should think about decisions made by AI.

Next, in Chapter 3 I will examine a broad set of ethical dilemmas induced by deploying autonomous vehicles. I will begin by analyzing the Trolley Problem as it relates to autonomous vehicles. Section 3.2 deals with the broader set of ethical problems brought about by autonomous vehicles. Lastly, many of these dilemmas deal with the question "How do we allow for multiple ethical views?" In Section 3.3, I present a potential solution to this question — allowing autonomous vehicle users to customize the behavior of their vehicle according to their own beliefs.

Additionally, I hedge my ethical arguments with technical considerations, including the limitations and implications of how autonomous vehicles are built, in Chapter 4. I begin by introducing the functional limits to the abilities of the software and hardware that drive autonomous vehicles. Section 4.2 goes further into depth, explaining how the hardware and software on autonomous vehicles work. Section 4.3 introduces and analyzes two methods of embedding ethical principles into autonomous vehicles. Following this, Section 4.4 explains *how* autonomous systems fail. Lastly, I give actionable recommendations for developers seeking to improve the safety and reliability of their autonomous vehicles.

Chapter Two: Trusting Moral Machines' Decisions

Whether a human is behind the wheel or merely sitting in what would usually be called the driver's seat, driving is an ethically-charged action. We must treat delegating the act of driving to autonomous vehicles as a separate ethical decision, but the main concern ethically is with how autonomous systems will fare in replacing human drivers. If we wish to answer this question, we must look at how autonomous driving systems are built. Correctly trusting autonomous vehicles necessitates guarantees about how they work.

2.1 Programming as an Ethically Prescriptive Action

Programming should be treated as an ethically-prescriptive action. Programming is not generally seen as an ethical task in itself. Programmers discuss ethics in the abstract, and most people know that programmers are supposed to think about ethics when they write code, but it is not entirely clear what this means or at what stage one should begin to think about the ramifications of the code they write. Any code that affects the world carries with it an ethical dimension. Whether code is poorly optimized, thus wasting electricity and contributing to climate change, or it causes an autonomous vehicle to collide with an innocent pedestrian, we must admit that there are effects to the execution of code that carry ethical weight.

This runs parallel to most decisions that we make in our lives. If you decide to eat

steak instead of tofu, you are making an ethical decision, but we do not go through this process (considering the moral weight of something) for each decision we make in a given day. It is too much mental load to fully evaluate the ethics of each action we make because there are just too many. This is also true in computing. Some things are not worth our worrying, while others are well worth it. The difficultly in ethical programming is in determining which decisions might carry moral weight and predicting how much they might carry.

Programming is ethically-charged in the sense that it deals with how people should act, but it is ethically-prescriptive in the sense that it prescribes some behavior given some input. Writing code is like writing a contract. If I write a program that spins up an Amazon Web Services EC2 instance, the program that I have written assumes my ethical responsibility. If you have good intentions in writing this code, but this program is misused in a way that you could not have foreseen, then most people would say you would have no moral culpability in its misuse. However, if I was lazy and did not see how my code might be used to achieve negative results, I share the blame for the effects. This is more similar to how we treat ethical or unethical actions, as opposed to ethical or unethical beliefs.

A large goal in computer science currently is to make programming ethically as easy and achievable as possible. We must strike a balance between the respect of human rights and the developments and application of technology like data science and artificial intelligence (Floridi and Taddeo, 2016). Overlooking ethical issues may cause negative social impact, but overemphasizing these ethical issues in the wrong contexts may lead to too much rigidity, causing us to lose our chance to harness the powerful technologies we are developing today. Furthermore, we cannot effectively plan for malicious programmers, and there will always be ways for malicious actors to exploit programs to manipulate their purpose. Nonetheless, we should ensure

that those who have good intentions can translate those intentions into good effects on the world around them, preserving the rights of each individual and harnessing technological power to its fullest extent.

You can almost guarantee that a program will be run deterministically, meaning that hardware running a program will carry out the commands it has been given with much more precision than its human decision-making counterparts. People are unpredictable, and this is often what leads to things like car accidents. However, programs are not guaranteed to run correctly. Sometimes hardware fails, but also, being deterministic does not imply that a system will carry out the intentions of its creator. In other words, programs might not accomplish their goals because they incorrectly describe programmers' intentions. Programs are often executed in environments entirely different than the one in which they were developed and tested. This introduces room for things to go wrong, even when hardware does not fail.

Software's power is a double-edged sword, ethically speaking. It carries the ability to cause great social good, but if it is not evaluated carefully for its potential effects, software has just as much potential for social harm. This being said, not all programs carry the same ethical weight. All programs take on some of their creator's moral beliefs, but some programs (e.g., the software systems running on autonomous vehicles) are embedded with more explicit ethical statements, as well. While perhaps not this explicit, there is ostensibly code running on an autonomous vehicle that decides whether or not to run over a pedestrian detected in the road. Whatever the software architecture that this decision is made through, this is an intrinsically ethical decision.

In traditional cars (as in any other human-controlled decision-making process), ethical belief has always guided how one acts in the driver's seat, but it has never completely dictated one's actions. Consider a scenario: Tara is driving her car on the highway. Suppose Tara's most sincere belief is that ethical egoism is the only

correct ethical theory, meaning that she believes an action to be right if and only if performing that action maximizes her self-interest. If a pedestrian accidentally wanders onto the highway and in front of Tara's car, there is a chance that she will swerve out of the way. According to Tara's ethical beliefs, she should have allowed her car to run into this person, but she just could not bring herself to do it. There is a chance that every human will make an action that does not reflect their moral convictions. Machines do not share this quality. The ethical belief that is encoded into an autonomous vehicle will directly inform the action it will take (except in the case of a system failure or malfunction).

Actions that were considered 'good' by the creators of the system will be enforced more often if we allow autonomous vehicles to make decisions that humans would normally make. If these creators do an adequate job of choosing which values to encode into their system, the world might be better off. However, defining an adequate job of choosing values is not as trivial as it sounds.

If autonomous car creators found the perfect ethical framework and applied it to every car across the world, we would be able to create machines that enact how people think they should act instead of how they act in reality. This concept — how one thinks they ought to act — has a name: normative ethics. How one acts in actuality is the domain of psychology. The creation of autonomous vehicles necessitates considerations that reach across disciplines, converting decisions affected by psychological phenomena to decisions informed by ethical principles. People do not often think "Does this action conform to my ethical beliefs?" in complex or high-stress situations. When given enough time, our decisions often reflect our convictions, but there is simply not enough time for one to contemplate the moral weight of one's actions for split-second decisions. This is not the case for a machine, which can perform complex calculations at breakneck speeds.

However, we must concede that we do not have a perfect ethical framework. Instead of comparing themselves to a gold standard ethical framework, the questions drivers ask themselves might take the form of "What would Kant do in this scenario?" or "What is the Utilitarian choice here?" Which of these questions should autonomous vehicle developers ask of themselves? Here is where one of the biggest dilemmas with autonomous vehicles arises — how do autonomous vehicles do the 'right thing' when the 'correct' ethical framework is ambiguous? If we step into an autonomous vehicle, we trust that the creators of that vehicle have thought about these problems and to have arrived at an acceptable solution. We trust the programmers who worked on the system to have created principled and robust code because if they have not, their code will have real and perhaps unintended implications on the world around us.

2.2 Predictability in Autonomous Systems

If we run the same code twice, we might observe entirely different results in each trial. Code does not tell the whole story. The text a programmer types into their machine makes up only one part of a tripartite relationship between hardware, software, and data. Even if a piece of hardware is perfectly built and there are no bugs in a program running on it, this program is only as good as its input. The output of a function on an input it was not designed to handle is called "undefined behavior." There are no promises on what a vehicle will do when it experiences undefined behavior. Errors can filter from system to system, causing additional problems down the line. If a system is not properly designed to deal with falling into this state, we cannot expect it to, for example, safely obey traffic laws. Figure 2.1 shows, at a high level of abstraction, how an autonomous vehicle's model of computation might be defined. Each of the layers (hardware, software, and data) depends on the other two, and the correctness of the end result (e.g., an autonomous car safely moving along a road) depends on correctly

dealing with errors in each of the three layers. In the end, errors are inevitable, but a robust autonomous system can deal with unexpected input safely.

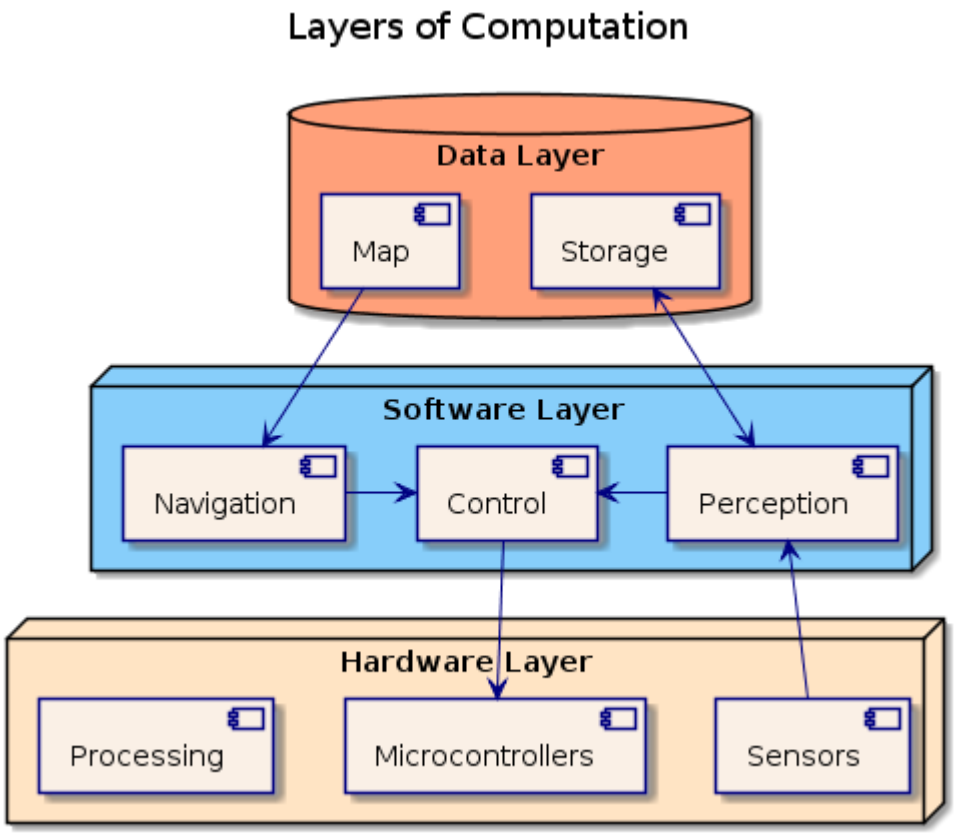

Figure 2.1 High-level computation hierarchy

Moreover, the presence of uncertainty in an autonomous system affects the level of determinism in that system. In the computer science sense, determinism implies that, given a particular input, a function will always produce the same output. We might be able to guarantee this on the software level, but as Figure 2.1 shows, we can never separate software from the hardware it runs on and the data it interprets. In real-world scenarios, we cannot control the consistency of hardware, nor that the data we will receive as input will make sense. In fact, it is impossible to perfectly replicate the state of the hardware and the sensor data we receive in real-world scenarios.

Deep learning is a subset of artificial intelligence through which neural network

'models' learn how to perform functions on data by being trained on a provided set of data. Computer vision systems, like those used on autonomous vehicles, use deep learning techniques to perform such tasks as determining whether a stop sign lies ahead. Even though there are several extremely important advantages to using deep learning, systems that use this technique sacrifice a degree of predictability in exchange. Deep learning algorithms can treat raw data as input instead of performing pre-processing. This is essential when interpreting real-time data because it would be very computationally infeasible to pre-process a large amount of data in $\frac{1}{20}$ of a second, the time that many autonomous driving systems provide between each update.

Furthermore, deep learning models cannot perform any better than the data they were trained on. If there is a situation that was not accounted for in the training data, the model will produce undefined behavior. It would be virtually impossible to perform the type of calculations necessary to safely operate a car without using this technology (and thereby sacrificing predictability). If there exists an input that was not accounted for in a model's training data, the model will enter undefined behavior. Undefined behavior is not favorable when lives are at stake, but it is inevitable if we wish to reap the benefits that deep learning brings. Not all autonomous vehicle sensors use deep learning, so when some systems produce undefined behavior, autonomous vehicles give more weight to the systems it trusts.

Undefined behavior in the presence of unfamiliar input is not the only vulnerability that deep learning brings. Machine learning, which includes deep learning, is a tool for building models that accurately represent input training data. The blind application of machine learning runs the risk of amplifying biases present in the provided data (Bolukbasi et al., 2016). When undesired biases concerning demographic groups are in the training data, well-trained models will reflect those biases (Zhang, Lemoine, and Mitchell, 2018). Mitigating bias in machine learning is currently undergoing heavy

research, and some potential solutions have been suggested, but there is not yet a clear way forward. This amplification of bias is intrinsic to how machine learning works, and if we wish to receive its benefits, we must prepare to receive the drawbacks, as well.

Fault tolerance is the key to handling these drawbacks and thus preserving the lives of self-driving car passengers and pedestrians. We must accept that some systems will not act predictably in real-world scenarios because real-world data is not predictable. Furthermore, as the neural networks running on cars continue to improve, their behavior will change. In a sense, predictability is one of the most important factors of graceful failure. If pedestrians can predict *how* a vehicle will fail, they might not have a deer-in-the-headlights reaction to being faced by a rogue vehicle. Human drivers are very unpredictable, and removing that from the equation of road safety has the potential to give each person the intuition for how to react in the unfortunate event of a car accident.

In a broader sense, the interactions between software, hardware, and data are complex. As a result, some scholars of data ethics have proposed forming a macroethical framework, regulating how we ought to interact with data (Floridi and Taddeo, 2016). Within this framework, we can begin to solve individual ethical problems. In other words, we are not seeing the big picture in data ethics yet, and we need to do this before attempting to fix specific problems. By creating a consistent macroethical framework, we will find solutions to problems across boundaries, treating the root problem instead of its symptoms. In doing this, we would shift the level of abstraction of ethical inquiries from being information-centric to being data-centric. Some data will never translate into information, but it will support actions or generate behaviors, and changing the paradigm of data ethics will help us take a holistic approach to the problems presented in this thesis. In the context of autonomous driving, this

means that we should begin by correctly interacting with the data layer in Figure 2.1. This approach can be applied to future problems that arise, which is preferable to solving similar problems individually.

Just as the proliferation of AI has stimulated new interest in the philosophy of the mind, it has the potential to stimulate new ways of thinking about ethics. AI laboratories could become experimental centers for testing theories of moral decision-making (Wallach and Allen, 2009). Moral agents, such as autonomous vehicles, are not simply agents who obey the rules of morality. Instead, they are in a bi-directional relationship with the moral decision-making process. Autonomous vehicles can surpass the ethical principles with which they have been created when they are presented with new scenarios. They even affect our own concepts of morality with their actions, and their effect will continue to strengthen as increasingly more decisions are made by autonomous agents.

2.3 Offloading Decisions to Machines

Some decisions are deemed too critical to allow artificial intelligence to make them. We might be hesitant to offload decisions in this category to machines for various reasons, but underlying each of these is that we are incapable of deeming the AI we are asked to trust as a trustworthy agent (Taddeo, 2017).

The relationship between a person and an instance of artificial intelligence can be modeled as a first-order relation. Goods or actions are exchanged between the two parties. Trust between the two (or rather, trust by the human involved with artificial intelligence) is a second-order property that affects the first-order relation (Taddeo, 2010). That is to say, if the human involved does not possess this second-order property of trust, then they will cease to continue in the first-order relation.

Luciano Floridi defines 'mature information societies' as having members who

possess unreflective and implicit expectations to be able to rely on digital technologies (Floridi, 2016b). However, trusting implicitly can lead to forgetting how exactly to trust a piece of technology or in which scenarios it should be trusted. In today's age, we must trust digital technologies if we wish to do anything, but we should ensure that we trust these technologies correctly. If we are to offload decisions to machines effectively, we must identify the correct way to trust the technologies involved in creating autonomous vehicles so that we can harness their value while protecting fundamental rights and fostering the development of our society (Floridi and Taddeo, 2016).

Identifying the correct way to trust digital technologies is no small undertaking. We need to have the right people thinking about these issues. If the people leading our society do not foresee problems that autonomous vehicles might cause down the road, it could mean catastrophe. If a small bug on a web server can take down internet service availability, the same small bug on a self-driving car can mean the death of dozens. The number of devices we have that are connected to the internet is only growing, and software developers are rolling out code faster and across more platforms than ever.

Mariarosaria Taddeo has argued that simply resorting to better technical design for autonomous vehicles would be similar to a Band-Aid solution for the greater problem (Taddeo, 2017). It will not solve the medium and long term problems that our society will soon begin to face on many fronts. There are critical issues within autonomous driving, but these issues arise out of only one application of the type of technology that will soon begin to overhaul how our society functions. Taddeo argues that our society requires a "normative infrastructure" to truly trust digital technologies correctly (Taddeo, 2017).

This infrastructure would define an overarching strategy, and its specifics could be

fine-tuned to fit the needs of its society. For example, we could enforce transparency into the ethical decision-making process of autonomous vehicles or require human oversight into the way that some digital technologies are deployed (e.g., those that make decisions concerning human beings). It would also be fruitful to define policies to ascribe liability to designers, providers, and users of digital technologies (Floridi, 2016a).

This might seem to be too abstract to implement successfully, but it is necessary to propose an overarching shift in how we interact with technology if we wish to appropriately solve data ethical problems that arise. We cannot tweak certain actions within the context of the same relationship. We need to rethink everything, starting from the broad strategy, and ending with the finer details. These details will fall into place once we have established expectations on how digital technologies should be trusted for the societal scale.

The act of delegating a decision to artificial intelligence is complex and multi-faceted. It involves creating a decision-making model, an algorithm that translates that model into code, datasets for the algorithm to operate on, the 'learning' process that model goes through if using machine learning at any stage, and the shift itself from human to artificial decision-making. There are also human factors involved: the social, political, and economic environment in which the technology is developed and deployed, the act of deciding to delegate the decision (i.e. a boardroom of directors found it to be more cost-efficient), and the techniques used in development and deployment. If anything goes wrong along the way, it can mean disaster for machine-made critical decisions.

Furthermore, the autonomous systems that have already seen roads and those that are still under development are almost entirely opaque. We have access to research papers that describe techniques for implementing only pieces of overarching

autonomous systems. We do not know if systems do use these techniques, how they are implemented, or how those implementations fit into the overarching systems. It is hard to trust something when you do not know how it works. Many people are under the illusion that computers will be better at driving than humans, but in reality, they are better at a specialized set of operations. Autonomous vehicles are proficient at collecting sensor data and performing complex calculations on this data, but they do not have human intuition. It takes precision to create a safe autonomous system, and there are no guarantees that a system has been implemented with the necessary precision without transparency. Even if a system has been on the road for three years with no crashes, how are we to determine if it is safe to get into our cars on a leap day? Software is not inherently robust. It takes eyes on code, testing, and iterative improvement if it hopes to perform to standards. We should not blindly trust digital technologies that do not carry a guarantee for each of these factors. This must be considered in the macroethical shift we hope to take as a society.

Beyond safety, we should also consider each user's comfort with using autonomous vehicles. We will never reap the benefits of self-driving cars if nobody uses them. Mimicking the user's driving style is one way to make self-driving car users more comfortable with the technology (Kuderer, Gulati, and Burgard, 2015). Perceptions of safety vary between users. Some users might prefer a large safety margin on each side of the car, but others might prioritize slower acceleration. By allowing self-driving car owners to train a model with their personal driving style, we do not have to find a solution that matches every user's perception of comfort.

This is also a different type of offloading decisions to autonomous vehicles. The vehicle will make decisions for the user after it has been engaged, but these decisions will ostensibly result in the same behavior as if a human was driving. Thus, the results are the same, and the difference lies in the agent who makes the decision. This may

seem to be an inconsequential distinction, but digital technologies' behaviors can change quickly. If a developer accidentally creates a bug in a new software update, each car that receives this update has the potential for catastrophic failure. This is not the case for human drivers, who, despite their varying driving styles, for the most part, drive safely. Imitating human driving styles might be an effective way of circumventing bugs in this functionality of autonomous vehicles, but it does not change the potential for other systems to break. As a result, we must learn to trust digital technologies correctly, even if we put safeguards in place to reduce the amount of trust necessary.

Chapter Three: Ethical Dilemmas Faced by Autonomous Vehicles

If we are to trust autonomous vehicles with our lives, we ought to hope that they work very well. These vehicles hopefully will work well once they reach the consumer market, but there is always the chance that something will go wrong. With autonomous vehicles on the road, many people could potentially get hurt because of the scale on which they will be deployed. On the slim chance that something does go wrong, we should seek to minimize the impact of the failure. Sometimes harm is unavoidable, and in this chapter, I will discuss the dilemmas faced by autonomous vehicles and how to minimize harm in several specific scenarios.

3.1 Dealing With Trolley Problems

Situations faced by autonomous vehicles' accident-avoidance algorithms are often compared to the Trolley Problem, one of the most commonly presented ethical dilemmas. There are two widely discussed variants of the Trolley Problem. In the "switch" version of the problem, a driverless trolley is heading towards five people who are stuck on the tracks. These people will be killed unless the trolley is redirected to a side track, on which another person is stuck. You are standing next to a switch, and if you pull the switch, the trolley is redirected from the main track to the side track (Nyholm and Smids, 2016). A common response to this situation is to pull the switch, minimizing the number of people who are killed (Greene, 2013).

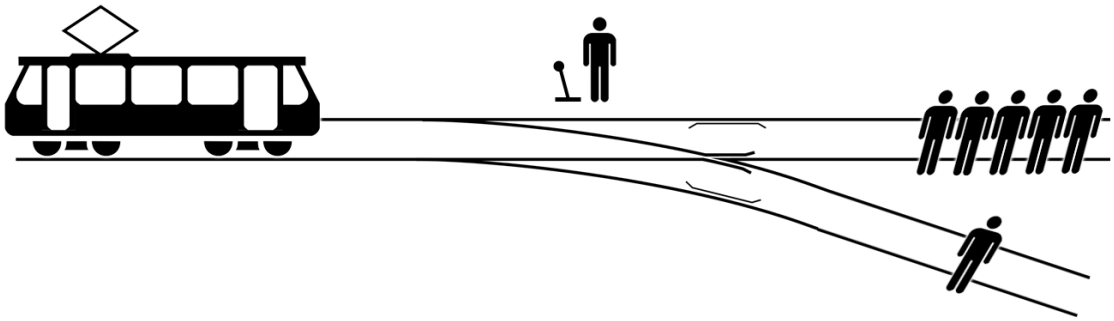

Figure 3.1 The "switch" version of the Trolley Problem

In the "fat man" variant, (Thomson, 1985) saving the five people on the tracks requires a different action. In this case, you are standing on a footbridge, and in front of you, there is a very large, heavy man. If pushed in front of the trolley, his mass would be sufficient to prevent the five people on the tracks from getting killed, but he would be killed in the process. In contrast to the first scenario, the most common response to the "fat man" Trolley Problem is to refrain from pushing the man (Greene, 2013). Most people explain that they do this to avoid actively killing anyone in the scenario, and pulling a switch is less direct.

Not every situation that is faced by autonomous vehicles can be reduced to a variant of the Trolley Problem, but the problems that can are often the most ethically compelling. For example, imagine an autonomous car driving toward a tunnel when suddenly a child runs into the road Goodall, 2016a. The car begins to brake, but it realizes that it will not be able to stop before striking the child. It has two options: hit the child (likely killing her), or swerve and hit the tunnel wall (likely killing the passenger). It is difficult to find an acceptable solution when faced with this decision. Each of the outcomes in this dreadful scenario is unfavorable, but harm is unavoidable in such a scenario, and creators of autonomous vehicles must decide how vehicles should act when faced with a comparable decision.

The Trolley Problem does not have an easy solution, as is the case with any ethical dilemma. In fact, it might not have a solution at all. This can be disconcerting to artificial intelligence programmers, who have spent their careers solving problems with right and wrong answers. Instead, ethical dilemmas are often used as a method of delineating between different ideological frameworks (Nyholm and Smids, 2016). Specifically, the Trolley Problem explores the differences between "positive" and "negative" duties, killing and letting die, and consequentialism and non-consequentialism. The value of the problem is not in finding an objective solution but in bringing one's ethical beliefs to light. It helps identify intuitions about the correct course of action and areas of strong agreement or disagreement. And by altering the scenario's rules, philosophers can begin to explore the reasoning behind responses, even if they are unable to articulate them explicitly (Goodall, 2016a).

However, the Trolley Problem might not be a good model for scenarios faced by autonomous vehicles. Holstein and Dodig-Crnkovic (2018) argue that the Trolley Problem is built on assumptions that are neither technically nor ethically justifiable, and thus, it is "intrinsically unfair." Real-world engineering problems are substantially different from hypothetical ethical dilemmas, but these dilemmas can help us isolate certain problems, even if there is often no good solution. Solutions to engineering problems must be (by construction) differentiable between better and worse solutions. While Holstein and Dodig-Crnkovic are correct in their argument that research around autonomous vehicles should focus on accident-avoidance instead of solving the Trolley Problem, the authors must admit that attempting to deal with the Trolley Problem is a part of creating a robust accident-avoidance strategy.

Additionally, the authors argue that, instead of treating the dilemmas faced by autonomous vehicles as the Trolley Problem, we should treat all human lives equally and focus more on scenarios where we must decide between hitting a correctly identified

obstacle or an unknown or incorrectly identified object. This argument is premised on the practical difficulty inherent to this problem. We might never have enough data to make an informed decision, so how should we proceed? This is another problem that cannot be solved per se. There is not much we can do in dealing with uncertainty besides practicing robust default actions and working to gain as much additional information as possible. However, this point is insightful insofar as we will never have a complete set of data for the environment in which autonomous vehicles operate. The models these cars use to make decisions must be calibrated to work when supplied with a highly variable stream of data.

The Trolley Problem must not obfuscate other ethical challenges faced by these systems. If we are successful in preventing our focus from becoming too narrow, the techniques presented in response to the Trolley Problem might be applicable further down the line. This falls in line with the type of macroethical framework recommended by Taddeo. In attempting to engineer an acceptable solution to the Trolley Problem, we are making progress on the question "How can we perform ethical analysis on a computer?" Solving additional problems presented by researchers (and those presented in Section 3.2) contribute to the same goal.

If our end goal in programming autonomous vehicles is to create the morally best solution for this real-world problem, we must nevertheless push on. The engineers who work on these vehicles must find a solution to the problems they face, and their solutions will be deployed at scale. Additionally, when autonomous vehicles reach consumer markets, one accident-avoidance algorithm could gain more market share than others, causing a sort of 'ethical monopoly.' If, for example, Waymo's software became ubiquitous, Waymo's executives might find themselves attempting to find a unilateral solution to this applied Trolley Problem.

From a philosophical standpoint, this is worrying. Within certain constraints, it

seems that multiple ethical viewpoints are permissible to all of us. This is to say that it is quite obvious that you would be committing a morally reprehensible act by intentionally running a pedestrian over in your car, but there is a multitude of acceptable responses to the scenario presented earlier involving the child stepping out in front of your car. No matter if you decided to swerve or continue straight, there is very little chance that you would be prosecuted for making either choice.

If we take this to be true, nobody can reasonably determine the 'right' answer to the Trolley Problem. The same can be said for the related autonomous vehicle accident avoidance problem. We could look to the law for guidance, but at the same time, this is an emerging issue, and the law might not yet be properly calibrated to deal with self-driving cars. Thus, we will explore how we should treat the issue morally instead of applying the letter of the law.

Those who identify with subjectivism or intersubjectivism believe that the ethicality of a decision is inherently subjective. In some extreme views, morality is entirely subjective to cultures or even individuals. On the other hand, less radical views hold that there are some things that will always be morally reprehensible, such as the example given above — intentionally running a pedestrian over with your car. Some things might always be considered morally commendable, like saving a puppy from drowning. Between these two extremes, we are left with a gray area. We have learned to accept this concept of moral ambiguity in our everyday lives, and we will come to accept it in the context of autonomous vehicles. The challenge we are faced with in autonomous driving is that the decision made by programmers is replicated across thousands of computers, enforcing programming decisions on a massive scale.

Cultural relativism is a subsection of subjectivism which argues that ethics are relative to the society in which one lives. Most ethicists reject subjectivism (and cultural relativism, by extension) because it implies that one cannot criticize the actions

of societies approved by a majority (or individuals acting per their own beliefs) (J. M. Anderson et al., 2016). For example, we do not generally consider human sacrifice or the Holocaust to be morally acceptable, but they were at one time approved by a majority in their respective societies. However, relativism does merit consideration to an extent in the context of autonomous vehicles. This is especially true when we consider autonomous vehicles in a situation relatable to the Trolley Problem. Different cultures (in particular, Chinese culture) make vastly different decisions when faced with the Trolley Problem. These differences affect cognitive processes, which then lead to differences in decision-making, judgment, and philosophical intuition (Gold, Colman, and Pulford, 2014). For example, only 52% of Chinese participants in a study agreed that it is "morally permissible" to flip the switch in the classic formulation of the Trolley Problem, whereas 81% of Americans and 63% of Russians held the same position (Ahlenius and Tännsjö, 2012). Since the debate is not settled on whether moral relativism or objectivism is the correct metaethical framework, we must compromise. Autonomous vehicles should probably reflect the values of the cultures in which they are operating, but only to an extent. They should operate within the bounds of some pre-determined level of moral ambiguity.

All this being said, we need an answer to the problem at hand, and we need a way of dealing with unforeseen problems as they come up. Several answers have been presented by others, and I will give a brief outline of these here. I will begin with the consequentialist approach to accident avoidance. The most common consequentialist ethical framework is Utilitarianism, and this is the one I will explore here. Utilitarianism holds that the morally right action is the action that produces the most good. According to Utilitarianism, we ought to bring about the greatest amount of good for the greatest number of people. A Utilitarian approach to accident-avoidance might apply this principle to the survival probability of each person an autonomous vehicle

detects. In other words, the vehicle will monitor its surroundings for people, keep a list of possible decisions to make, and estimate how much suffering each decision would cause. The algorithm will choose the decision that entails the least suffering. Such a system is outlined in M. Anderson, S. L. Anderson, and Armen (2005).

Suffering is difficult to quantify, and it is dangerous to base an approach off something imprecise. It might be more feasible to predict bodily harm, perhaps assigning a certain value to certain types of injury. Calculating the probability of survival would be even easier. If we achieve reasonable accuracy in using survival probability as a proxy for social utility, computational Utilitarianism is promising because machines can abide by the theory at least as well as human beings and, perhaps, even better given that humans are not able to gather the data necessary to perform moral calculus (M. Anderson and S. L. Anderson, 2007).

Most people prefer autonomous vehicles to implement the Utilitarian approach, but they would not want to buy one themselves (Bonnefon, Shariff, and Rahwan, 2016). Instead, when told that they would play the role of passenger, participants preferred an avoidance-algorithm that protected the passenger at all costs. This study dubbed this asymmetry in opinion the "social dilemma of autonomous vehicles," arguing that everyone has a temptation to 'free-ride' instead of adopting the behavior that would lead to the best global outcome. However, characterizing it as such presumes that Utilitarianism is the only optimal approach.

Criticisms of Utilitarianism are well-known, even if it is the most popular choice for an autonomous driving ethical framework (Bonnefon, Shariff, and Rahwan, 2016). Utilitarianism does not always provide the answer most people would consider 'right.' In most recreations, the ethical framework considers the sum of pains and pleasures, not their distribution. Imagine the following scenario: society at large functions normally, but unhappiest 1% of the population is instantaneously executed at the

beginning of every year. For those who belong to the 99%, life goes on, and for those unlucky few, they feel no fear before death. On average, happiness increases dramatically in such a society, but most people would not consider this to be morally 'right.' Utilitarianism is but a single potential answer to a complex set of questions in the ethical dilemmas that face autonomous vehicles. Many researchers who work with autonomous vehicles (and even in this sub-field of ethics in autonomous driving) take Utilitarianism's superiority for fact. In reality, many people have contrasting beliefs, and these beliefs can be modeled by autonomous vehicles, as well.

The avoidance-algorithm that prefers protecting the passenger could be likened to the moral theory of egoism, representing an alternative to Utilitarianism. Because participants in Bonnefon, Shariff, and Rahwan, 2016 would prefer to ride in a vehicle that implements a similar accident-avoidance algorithm, market forces might make this option might be the most appealing to car manufacturers. However, it is not clear whether egoism would be a sufficient candidate to solve the problem at hand. Consider the scenario in which the child appears in front of the tunnel. If the autonomous vehicle determines that it will not be able to safely brake in time, the egoist accident-avoidance algorithm will choose to protect the passenger by staying on course. Choosing one's life over another's might seem like a tough choice, but it is not clearly problematic in itself. Let's modify the scenario: suppose five children appeared in front of the tunnel — now 100, or 1,000. The egoist accident-avoidance algorithm will make the same decision in each of these scenarios. Most people would consider killing 1,000 others to save yourself to be morally reprehensible, so we must either practice a reduced version of egoism, or we must look elsewhere for an acceptable accident-avoidance solution.

Perhaps taking a deontological approach will help circumvent some of the issues outlined above. Deontology is a normative theory that argues that some actions ought

or ought not to be performed, regardless of how they might affect others. Immanuel Kant's is one of the most respected deontological ethical theories, and rule-based ethical theories like his appear to be promising because they offer a computational structure for judgment (Powers, 2006). Kant's Formula of Universal Law says that we should "act only in accordance with that maxim through which you can at the same time will that it become a universal law" (Kant et al., 2002).

I will not give my interpretation of Kant's ethics here, but I will present a machine-computable version, as demonstrated in Powers (2006). Kant defines a maxim as a subjective principle of the volition, but our interpretation of this term in an algorithmic setting will more closely resemble a plan for how to proceed from an initial state. According to Kant's theory, maxims should be universalized to evaluate their ethical permissibility, and we can do the same with the concept of plans. We might universalize these plans according to the technique presented in Powers (2006), and we could determine their permissibility as either forbidden, permissible, or obligatory. In some scenarios, it might make sense to keep track of a set of forbidden and obligatory actions. If this was the case, we could directly check to see if our universalized maxim belongs to either set instead of performing this universalization process every time, thereby saving computational resources.

John Rawls presented another ethical theory called Contractarianism (Rawls, 1971). The term 'Contractarianism' can refer to either a meta-ethical or normative theory, but the latter is more relevant to autonomous vehicles and will be the focus of my analysis here. In its most basic form, normative Contractarianism says that the best solution to a problem will be based on the hypothetical agreement of the participants involved. This theory has also been applied to self-driving cars, and the approach is different from either Kantian or Utilitarian techniques in accident-avoidance (Leben, 2017). The Rawlsian approach gathers the vehicle's estimation

for the probability of survival of each involved person, then it calculates which action most people would agree to if each person did not know who they were in the scenario. In Rawls' words, we consider the perspective of each person if they were placed in a hypothetical bargaining position and under an imaginary "veil of ignorance" as to their identity in the situation.

This seems feasible in theory, but how should we determine what participants will agree to without directly asking them? The decisions made by self-driving cars must be made in split-second intervals, so each person's expected decision must be computed efficiently. Rawls suggests we use the "Maximin" heuristic, a strategy for maximizing the minimum payoffs. Essentially, this heuristic attempts to make the worst-off person as well-off as possible. According to Rawls, every self-interested player will follow this criterion. By using each involved person's probability for survival as the measure to which we apply Maximin, this algorithm is readily quantifiable and applicable to autonomous vehicles.

However, by reducing our criterion in each of these decision-making algorithms to survival probability, we lose nuance. If a car had an option between leaving 1,000 people paralyzed and killing one person, it would avoid killing the person. This seems counter-intuitive from several perspectives. For example, Utilitarians might argue that paralyzing 1,000 people produces more social harm. However, the design of each decision-making algorithms is not dependent on the criteria it takes into account. Survival probability is an incomplete criterion, but it gives us a rough recreation of an ethical theory. If there are other criteria available to a system, it could take these into account as well. For some ethical frameworks, survival probability might suffice, and for others yet (like egoism), the survival of others in the scenario might not even have a place in the moral calculus.

Additionally, some might object to this approach's 'targeting' of safer people in

collisions because they probably have a higher probability of survival. For example, self-driving cars might choose to hit motorcyclists who wear safer helmets (Santoni de Sio, 2017). However, the semantics of this criticism itself is questionable. It is misleading to say that any of the accident-avoidance algorithms we are considering here are programmed to hit anyone. Using the terms 'hit' or 'target' suggest that this is the intention, but we are instead trying to minimize negative outcomes.

Thus, the question remains: which of these ethical frameworks (or perhaps one unlisted) should autonomous vehicles follow? In my opinion, this question is unanswerable. The premises on which the question is based are shaky. First, there is nothing anyone could say or do that could convince everyone affected by autonomous vehicles that one solution is the best. Ethics is an unsolved disciplined, and it will remain that way. It does not have the type of clear answers to which engineers are accustomed. Second, it would be unethical to subject autonomous vehicle passengers to decisions made based on ethical reasoning with which they do not agree. We should, within certain limits, allow autonomous vehicle owners to choose the ethical framework their car will follow.

Autonomous vehicle manufacturers might install ethical knobs in their cars to allow users to customize the cars' behavior (Contissa, Lagioia, and Sartor, 2017). If ethical knobs were installed, passengers of autonomous vehicles could pick between the aforementioned ethical frameworks. Within certain guidelines, passengers could even create their own ethical framework. Car owners might take a quiz assessing their ethical beliefs, and autonomous vehicle manufacturers could tailor the behavior of the car to these beliefs. We must remember that a machine will not be able to replicate the thought process that led to these ethical decisions, but it will make decisions based on similarity in the result. In other words, autonomous vehicles will learn to mimic the causal effects of their human driver counterparts, but they will not come

to the same conclusions every time because they are not undergoing the same ethical deliberation.

However, embedding ethics in self-driving cars necessitates the consideration of topics outside the Trolley Problem. Thus, we should broaden our scope before moving forward.

3.2 The Broader Set of Ethical Dilemmas

Not every ethical decision can be reduced to the Trolley Problem. Once realizing how many decisions an autonomous vehicle makes and that each of these decisions carries moral weight, we realize just how rarely the Trolley Problem should cause us to worry. Risk calculations are based on a combination of severity and likelihood. Being placed in the Trolley Problem carries the potential for high risk, but it will be unlikely as long as autonomous cars function appropriately. It would be negligent to focus so much on this one dilemma when the large majority of decisions carry just as much significance yet are understudied.

Before worrying about whether a specific ethical framework is carried out in the rare situation that a child jumps into the road and a self-driving car is going too fast to stop in time, we should ensure the protection of common goals that exist between all ethical frameworks. In analyzing this broader set of ethical problems for autonomous vehicles, I will focus on dealing with ambiguity and selfishness. Other topics such as environmental, infrastructural, and social effects are less relevant to the central topic of this thesis, but they are still pressing concerns to our society and merit a brief discussion each.

Ambiguity in a system like an autonomous vehicle is unavoidable. In Chapter 4, I will expand on the technical grounds for this claim, but for the sake of argument, let us assume this to be true. When trying to ensure safety in the presence of ambiguity, it is

essential to make the best decisions with whatever resources are available. Sometimes, there is little to no data available, and we must rely on a default action to ensure safety.

Default options are ubiquitous in computer science, and, moral rectitude aside, a default option should be something that makes the most sense most of the time. Regardless of the ethical framework involved in an autonomous vehicles' decision, there are common things we seek to maximize (e.g., passenger and pedestrian safety) and minimize (e.g. performing risky actions in the presence of uncertainty). Taking these common values, we can provide better default actions that do not require an ethical framework to be programmed into the car's code or an ethical knob to be turned. The 2018 Uber self-driving car crash in Tempe, Arizona was caused by a problem in its default actions (National Transportation Safety Board, 2018).

An Uber self-driving car was driving in Tempe when it detected a pedestrian walking alongside a bicycle. The car classified this pedestrian correctly initially, but on subsequent calculations, it classified her as a cyclist. The Uber software created a timer after each new classification, and after the timer expired, the car would take the appropriate action to deal with the classified obstacle. However, this autonomous vehicle kept re-classifying the pedestrian, and thus, kept resetting the timer. The default action was to keep driving at full speed, and unfortunately, the pedestrian was hit by the vehicle as a result of this software bug.

If an autonomous vehicle detects an obstacle in the road, it should always try to avoid hitting the said obstacle. This is a common goal, regardless of the ethical framework used in reaching the ultimate decision. However, there is not always a clear cut action to avoid hitting an obstacle. In some scenarios, such as driving on the highway, braking is one of the worst actions to take. No matter if the car stops safely, or if the passenger dies from being rear-ended by a vehicle traveling at 60

miles per hour on the highway, the decision to stop carries ethical weight because of its potential implications on those surrounding an autonomous vehicle.

Luckily, this sort of decision is often not as difficult to make as those which are similar to the Trolley Problem. If we have a common goal to protect passengers and pedestrians, the solution to seeing an obstacle in the road with no other cars around is straightforward — stop the car if you can. Otherwise, take evasive action, but make it as safe as possible. Kantians, Utilitarians, and Rawlsians will all agree with this solution. Admittedly, most situations are not so contrived, and their solutions require more reasoning than "brake if you can." Nevertheless, this illustration demonstrates that there are many decisions that autonomous vehicles make that do not involve debate on the correct approach. Decisions that are common between ethical frameworks, therefore, merit different treatment than Trolley Problem decisions.

Selfishness is also a necessary consideration in our discussion on autonomous vehicles. Previously, I presented an egoistic approach to creating accident-avoidance algorithms. If a human was driving a car, and they consciously made the decision to not avoid hitting a person in the road to save their own life, they would not be held liable on account of the state-of-necessity defense. If a programmer made this decision beforehand, the state-of-necessity defense might not hold up, and this degree of selfishness might not be permissible (Contissa, Lagioia, and Sartor, 2017). Since different standards apply to pre-programmed autonomous driving software, we need to regulate the degree of selfishness that we allow on the part of whoever determines a car's ethical framework.

However, selfishness is a hard concept to define. With the type of explicit ethical encoding that I propose in Section 4.3, the difficulty in measuring selfishness might be lessened, but there is no indication that the problem would be fixed by an architectural overhaul. Perhaps a first attempt at defining over-selfishness might be "the favoring

of one person's life significantly over that of others," but the term "significantly" is vague, and this is an outcome-oriented approach. In other words, something bad would have to happen before the creators of an autonomous vehicle's software would know that their system was too selfish. To avoid this problem, a testable, code-oriented approach would be ideal. Crafting such an approach is a significant ethical goal that can also be applied to other dilemmas faced by autonomous systems.

In a different vein, there are some ethical problems that self-driving cars will face that do not involve decisions made while driving. For example, how will autonomous vehicles affect the environment? Once cars can drive themselves, searching for parking might cease to exist. Instead, car owners might instruct their vehicles to drive in circles until they are ready to return to their vehicles. Drivers would not have to pay for parking, but they would expend more fossil fuels. Selfishness manifests this time in the choice not to pay for parking. This is a broader problem facing society rather than those near a self-driving car. If selfishness is allowed to run rampant, air-quality would greatly suffer, causing harm to us all (Fox, 2016).

However, it is possible that city structures would shift to accommodate self-driving cars, thereby removing the need to instruct one's car to circle the city. High-efficiency parking structures could communicate to cars, directing them to get off the road as soon as possible to cut down on traffic. Additionally, cars that do not need the room for a driver's side door to swing open can park much closer to each other. Autonomous vehicle parking lots or structures can decrease parking space by an average of 62% (Nourinejad, Bahrami, and Roorda, 2018).

If all cars could communicate with each other, traffic lights might not be needed anymore. Cars can avoid each other — even at high speeds — much more efficiently than our current human-oriented traffic control systems. This would greatly cut down on traffic congestion, perhaps giving each commuter more time in the day. In 2016,

the average American drove a reported 50.6 minutes per day (AAA Foundation for Traffic Safety, 2016). If commuters spent less time in traffic, and they were free from driving when they were in traffic, it would cause a societal shift simply by giving each person an extra 50.6 minutes per day. As the cost of commuting drops to zero, pressure for workers to live near the city center would reduce, perhaps also resulting in increased suburban sprawl (Fox, 2016). In effect, emissions will both increase and decrease from different effects of autonomous vehicles reaching mass usage, so it is difficult to predict the results.

Furthermore, self-driving cars do not require passengers to be sober if they wish to transport themselves. In 2011, alcohol was involved in more than 39 percent of motorist fatalities (J. M. Anderson et al., 2016). Autonomous vehicles could eliminate more than a third of traffic deaths just by taking control from alcohol-impaired drivers. However, removing this barrier to consuming alcohol might cause alcoholism to become more prevalent, or it might cause people to become more reckless when they are out (Lin, 2015). Many people rely on the prospect of negative consequences to hold them back from poor behavior, and autonomous vehicles will remove some of these consequences, like DUIs. While this is probably a net positive, there are going to be secondary effects which we cannot reliably predict. We must be prepared for these secondary effects to have an impact at least as large as the primary effects we anticipate now.

We must think about the institutional effects this shift might have on our society. If there are many fewer car accidents, then car insurance companies might cease to be profitable. Police departments might not be able to support themselves with traffic ticket revenue. Additionally, 16% of organ donations come from car accidents (U.S. Department of Health & Human Services, 2020). If there are radically fewer car accidents because vehicles are so safe, there will be fewer organs available for those

who are in dire need of them. Despite their exciting potential, autonomous vehicles have the potential to cause a lot of ethical problems if we are not careful about how we transition.

Carjacking is another of these secondary ethical dilemmas. In some countries (such as South Africa), carjacking is a regular occurrence (Davis, 2003). Suppose autonomous cars are programmed to always stop (if they can safely) when they detect a pedestrian in the road. Carjackers might take advantage of this by jumping into the road in front of an autonomous vehicle, making it stop, and forcibly removing the passenger. This might be something that developers from South Africa, for example, might think about in the process of programming accident-avoidance software, but it is not apparent to those of us who are not from a culture where carjackers are prevalent. Is this something that software developers should be expected to think about if it is not a problem in their society? Or perhaps there should be an "ethical localization" package, similar to those for translating video games. This problem further reinforces the point that there is no blanket solution to ethical dilemmas. Cultural values vary, but the needs of each society vary, as well.

Moreover, autonomous cars will not be cheap in the foreseeable future. Those who can afford autonomous vehicles might be safer than those who cannot. Inequity could be caused by the capability for autonomous vehicles to avoid crashes, or it could be caused by something less obvious. Self-driving cars equipped with the technology necessary to communicate with other vehicles might be programmed to avoid those with which they can communicate more often than vehicles with which they cannot. If inequity in safety between autonomous and human-driven vehicles was enough, it might cause an entirely separate social dilemma. In this scenario, should consumers even buy self-driving cars knowing that they are safer, but that they also unfairly target those who cannot afford the same luxuries?

Lastly, many job sectors will be affected by automation in the coming decades, and driving is no different. There are approximately 2,000,000 truck drivers and roughly 400,000 other professional drivers in the United States (Bureau of Labor Statistics, 2019a), (Bureau of Labor Statistics, 2019b). If autonomous vehicles were hypothetically found to save 10% of lives from traffic-related incidents, but they were also found to cause unemployment to rise by 1%, it is unclear whether they have created a net positive effect on society. Traffic-related deaths make up a relatively small number of deaths per year, but every American citizen relies on the economy to survive. These drivers might move on to take up higher-skill jobs, creating a positive effect, but they might not. There is a great deal of speculation, but we cannot yet say.

3.3 The Ethical Knob: A Potential Solution

As previously mentioned, allowing users to customize the ethical framework of their autonomous vehicle, as demonstrated by Contissa, Lagioia, and Sartor (2017), represents a potential solution to our inability to choose an outright "best" ethical framework. With this feature installed in self-driving cars, we move from a single mandatory ethical setting to providing for personal ethics settings.

The knob presented in the paper by Contissa, Lagioia, and Sartor provides a spectrum of ethical settings between the two extremes of altruism and egoism, as shown in Figure 3.2. After a user selects their desired setting, the autonomous vehicle makes decisions based on a combination of the data available to it (e.g., survival probabilities of those nearby) and the passenger's altruism setting. If the knob was turned toward altruism, the car would favor pedestrians more often, and if the knob was turned toward egoism, the car would favor the passenger more often. However, if such a system was implemented, who would decide the permissible level of egoism

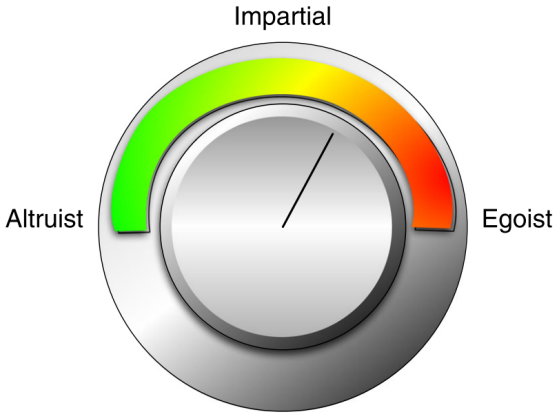

Figure 3.2 An ethical knob operating on a single criterion: altruism (Contissa, Lagioia, and Sartor, 2017)

and altruism? And what exactly would the "Impartial" setting represent?

If too much egoism was permitted, the car would act as a pedestrian killing machine. If too much altruism was permitted, the car would act as a suicide device. Neither of these two options seems desirable, so even after adding an ethical knob, we are still faced with ethical dilemmas. If the ethical knob ever saw use, we would need to set reasonable limits for the acceptable level of altruism and egoism. Regulatory agencies, either government or professional, might set these limits. However, the technology will probably be developed before regulations are created, so developers will have to be careful about setting safe limits until standardization occurs. Millar compares the decision autonomous vehicle users must make by turning the ethical knob to the kind of decision a patient might make when dealing with today's healthcare system. Healthcare professionals are seen as having an ethical responsibility to "provide an appropriate set of healthcare options for the patient to choose from, and to reasonably counsel patients on the benefits and risks of each option" (Millar, 2015). Perhaps car manufacturers carry the same responsibility in the future.

However, setting limits on both altruism and egoism would not determine what the middle of the dial should represent. The manufacturer would have to provide

some default value for the "Impartial" setting. This is the very problem that the ethical knob seeks to avoid, as this would create a unilateral ethical setting for passengers who choose not to move the knob. Even though this solution has created new problems, these problems are not inherently unsolvable. Perhaps there is no perfect placement for the middle of the dial, but car manufacturers could be confident that their default setting lies within the realm of moral permissibility if it is between the chosen "Altruist" and "Egoist" limits. To avoid choosing a unilateral setting for users, car manufacturers could require autonomous vehicle owners to turn the knob before the car starts. In this case, responsibility for the placement of the knob is shared between the manufacturer and passenger. Each can negotiate their own views within certain limits.

However, there is good reason to believe that many people would not give thought to turning the ethical knob in their car (e.g., turning to it the minimum amount necessary to start the car). These users might trust the manufacturer's default setting, thinking that they have more expertise on the matter. Alternatively, they might simply not want to exert enough effort to put thought into the matter. Forcing passengers to turn the knob is not a surefire solution.

Additionally, requiring users to turn the knob would imply some liability on the part of the passenger for the actions the car takes. We 'allow' human drivers to make terrible decisions (as long as they are not legally negligent or malicious) while operating traditional vehicles, although perhaps this is because of a lack of effective preventative methods. How would we extend the same ethical leeway to autonomous vehicle manufacturers (who build the ethical knobs) and the users (who turn them)? The share of liability in this scenario is a departure from traditional models, and it will need to be examined closer if customizable ethics become popular.

Furthermore, we must ask ourselves if passengers would refuse to drive such cars

if the liability model changes. If consumers had to choose between cars whose manufacturers took sole responsibility for crashes and cars that included settings that, if chosen, might come back to have legal repercussions, they would probably choose the former. We know that human-driven cars fall into the former category, and some autonomous vehicles might fall into the latter (Contissa, Lagioia, and Sartor, 2017). If consumers were faced with this decision, they might choose against autonomous vehicles, thereby cutting off funding and slowing technological advancement. Bonnefon, Shariff, and Rahwan argue that manufacturers and regulators face a major design challenge in balancing competing public preferences between a moral preference for "utilitarian" algorithms, a consumer preference for vehicles that prioritize passenger safety, and a policy preference for minimum government regulation of vehicle algorithm design. The ethical knob will create a similar dilemma.

Autonomous vehicle manufacturers should prevent such a choice between liability and immunity from being available to consumers, or they will be at the mercy of market forces. This being said, the question of liability for self-driving car companies, passengers, and part manufacturers is still open. Courts have not begun to form an opinion on this issue, and we can almost be sure that the technology will be deployed before any legal adjustments are made.

Moreover, if users of the ethical knob do not gain full visibility into the decision-making process of their car, it would be easy for users to become manipulated or confused over a matter of life and death. For example, suppose we represented the spectrum covered by an ethical knob as an interval and altruism values as numbers, where 0 is limit for egoism as determined by some regulatory agency, and 1 is full altruism. Manufacturer A's ethical knob might operate on the full interval (0, 1), but Manufacturer B's ethical knob might only allow users to choose on the interval (0.1, 0.3). The "Impartial" setting for each knob would be set to 0.5 and 0.2, respectively.

These are considerable differences in ethical beliefs, but even more drastic is the difference in their behaviors. Even though Manufacturer A and B's knobs could look entirely identical, Manufacturer B's car could kill 30% more pedestrians. For society to be able to trust customizable ethics settings, there needs to be more transparency in the decision-making process than that provided by the labels 'less altruistic' or 'more altruistic.' As suggested by Millar, car manufacturers should take the same approach as healthcare professionals when it comes to informing their customers on the ramifications of their decision.

In its current formulation, the ethical knob is limited to a single criterion. In effect, its only input is a setting between egoism and altruism. However, this is only a limitation of the implementation presented by Contissa, Lagioia, and Sartor. One could apply the ethical knob to other criteria in the ethical frameworks discussed in this chapter. For example, imagine a knob with "Consequentialism" and "Deontology" on opposite sides, or option to choose between favoring "Pedestrians" or "Cyclists." The possibilities are endless because the actions of autonomous vehicles make endless implicit ethical decisions. We usually make these decisions without a second thought, but the process of creating autonomous software makes our biologically automatic processes more explicit. Realistically, these knobs would probably be tied to a digital interface, and new car owners could take a survey with many questions about their ethical beliefs to configure their vehicle before it starts.

However, the more ethical features we make customizable, the more moral limits we have to determine. We have already struggled enough with the Trolley Problem — can we handle considering countless other dilemmas brought up by increasingly ethically-complex designs? While it is exciting to think about cars that drive for us and make the same decisions that we would make, we must proceed carefully down the road of introducing additional ethical criteria in autonomous vehicles. If additional

ethical configuration features are added to AVs, they should be added slowly to gauge their effects — both direct and indirect.

Lastly, if vehicles with different ethics settings could communicate with each other, how should multi-agent systems make decisions? In its current iteration, self-driving car technology runs on individual cars, but it is feasible that systems make decisions collectively (for example, intersections communicate with individual cars to safely interleave the crossing vehicles). If different cars have different ethical configurations, is it enough to act based on some average of the individual values? Perhaps this is fairer than our current human-driven system, wherein many people can be left at the mercy of a reckless driver. In a sense, taking the average of ethical knob settings is the democratic approach to accident-avoidance — everyone's beliefs are taken into account.

However, democracy is not always an appropriate system. Imagine 9 cars driving down a highway are set to full altruism, and one car's ethical knob has been hacked, and its egoism setting is past the morally permissible limit. This car might have learned that the fastest way to traverse its route is by driving on the wrong side of the road. The 9 altruist cars might swerve uncontrollably off the highway in an attempt to save the recklessly driven care, leaving it to profit off their altruism. In this scenario, a multi-agent system might override this one car's decision, preventing a major risk to the lives of many others.

Chapter Four: Practical Considerations in Autonomous Vehicle Decision-Making

Many misconceptions around autonomous vehicles and their potential effects on society can be dispelled with knowledge of their technical limitations. Anyone can tell you that artificial intelligence is not the same as human intelligence, but few can tell you the difference between the two. The differences between autonomous vehicles and their human-driven counterparts are essential to understanding how our society will change. In this chapter, I seek to contextualize ethical analyses from Chapters 2 and 3 by explaining how autonomous vehicles function and where ethics bleeds into the technical process of decision-making.

4.1 Limits to Artificial Intelligence

At its heart, the software that runs autonomous vehicles runs by sensing the vehicle's surroundings, analyzing this input data, and finally classifying potential decisions. At a high level of abstraction, these systems work by assessing their current state and making the best decision according to some heuristic. The heuristics used by autonomous driving systems are specialized according to the system. For example, a navigation system might prioritize maintaining space around the vehicle it is running on. If it has no choice, a vehicle might decide to take a path with low clearance (for example, between two semi-trucks), but it is safer to avoid this, and the heuristic represents this.

As shown in Section 4.2, these systems can be very complex, and they might execute their programmers' intentions flawlessly, but they will never be able to overcome the limits inherent to the way they are built — and the way they must be built. Engineers cannot consider algorithms if they are not computationally feasible. We might have the perfect algorithm for a given problem, but if it is incapable of running at the required speed, we cannot use it. The goal is to find techniques that are both robust and feasible.

As close as its end behavior might resemble ours, artificial intelligence is not human. Artificial intelligence cannot even simulate true humanity. Computers and brains are both complicated and powerful, but humans have evolved to be good at tasks essential to our survival. Our society has developed in response to humans strengths that came from evolution. For example, humans make extensive use of eye contact while driving in parking lots (Fletcher et al., 2009). We could design parking structures for computer-driven cars without much trouble, but until it becomes profitable to stop providing parking for human drivers, autonomous vehicles will operate in mixed AI/human scenarios. Perhaps in the future, the government will provide incentives for autonomous parking structures like today's parking spots designated for electric vehicles, but this is will remain far off until self-driving cars are much more common. If we expect autonomous systems to pick up on nonverbal gestures like eye contact, we project human capabilities onto machines that do not possess the same qualities. Even though we should try to minimize the time autonomous vehicles spend in hybrid human-autonomous driving environments, we must design for the scenarios autonomous vehicles drive in now before we design large-scale systems optimized for AI.

This being said, autonomous systems are better equipped than humans for most tasks involved in driving. According to a study by the ENO Center of Transportation,

about 93% of the 5.5 million crashes in the U.S. have been attributed to human error (Gogoll and Müller, 2017). Autonomous vehicles can prevent many of these accidents, but only once they have been thoroughly tested and are deployed on a large enough scale to make a difference. Until that day, systems like Tesla's Autopilot mode (as opposed to the Full-Self Driving mode) will be more common.

Tesla's Autopilot lies at SAE Automation Level 3, meaning that it controls steering/acceleration and monitors the environment around the car, but it requires active driver supervision. If a Tesla running this system detects something it does not know how to handle, it will hand control back to the human driver. However, this fail-safe maneuver is not perfect. Some experiments suggest that human drivers need up to 40 seconds to regain situational awareness (Lin, 2015). We cannot expect vehicles to predict a minimum of 40 seconds into the future, and we cannot expect humans to acclimate to their environment any faster than that. Thus, there is a limitation in the application of the current state of this technology, given the way our society currently functions.

Ethical theory comes from a human-centered perspective. While some think that ethics are inherent to existence itself, we are unable to interpret ethics from any other point of view. Machines do not share the qualities that make us human, and thus, they have no concept of this perspective. While it is theoretically possible to encode an artificial human-centered perspective into autonomous systems, the values and concerns expressed in the world's religious and philosophical traditions are not easily applied to machines (Wallach and Allen, 2009). Nevertheless, we have no better option than to give AI our best estimation of our values.

Until this point, I have not asked an important question: can machines even be moral agents? Here, the term "moral agent" refers to something that can discern right from wrong and be held accountable for its actions. While full-blown moral agency

may be beyond the current (or even future) state of technology, there is a spectrum between operational morality (which we are after in programming autonomous vehicles) and "genuine" moral agency.

Some agents might have explicit moral actions programmed into them (e.g., "If a pedestrian walking a bike is detected, treat them as a pedestrian"), and others might merely take on the beliefs of their creators (e.g., pedestrians with bikes are identical to pedestrians according to a machine learning-based classifier). In the former case, decisions are made according to explicit rules. In the latter case, decisions are made according to the interaction of decisions made in the design process and the implementation process. These agents might produce the same end behavior, but their decision-making processes differ. In reality, autonomous systems are programmed with both explicit and implicit moral decision-making, and regardless of the exact level of moral agency in an artificially intelligent system, we must treat it as lying somewhere between two extremes: full moral agency and carrying the residual beliefs of their creators.

Furthermore, predictions of the future will almost always prove to be incorrect, but this thesis would become outdated quickly if it did not at least attempt to predict how cutting-edge and future technological advancements will affect the problems at hand. Some of the technical limitations that exist now will disappear as computing power becomes even cheaper. The most noticeable limitation put on the algorithms discussed in Chapters 2 and 3 is the degree to which autonomous cars can simulate and predict what will happen in the environment they find themselves in at the current point in time. It might be possible to compute the risk of death or injury for every person in a car's view. Having the ability to perform better predictions improves the safety of autonomous vehicles, but it does not necessarily make autonomous vehicles "more moral." As demonstrated previously, moral agency comes from the ability to

discern between wrong and right. Autonomous vehicles will continue to face most of the ethical dilemmas they face today, even if they are more technically advanced. Thus, we must focus on finding robust solutions to ethical dilemmas regardless of the pace of technical innovation.

That being said, technical development will not be the only change in the next ten years. It is impossible to predict how the public will respond to the further development of autonomous vehicle technology, and the amount of public attention given to the technology will affect how much attention policymakers give it, too. Some countries will adopt this technology (and regulation targeted at it) before others, and these early adopters will serve as examples for the rest of the world. For example, AI treating people differently based on their age, social status, or other data was already made illegal in Germany in 2017 (Bundesministerium für Verkehr und digitale Infrastruktur, 2017). However, increased regulation might prevent autonomous vehicles from coming to market. Bonnefon, Shariff, and Rahwan argue that autonomous vehicle manufacturers should configure their products appropriately to avoid running into regulation based on the behavior of their vehicles. Otherwise, we, as a society, will lose time in reaping the benefits of removing human drivers. No matter when autonomous vehicles become popular, we will not have comprehensive laws targeted at autonomous vehicles before these cars make their way into wide usage. As a result, we must be wary of the problems that arise in this transition period, attempting to catch problems in the implementation of the technology even before standardization occurs.

4.2 Exploring the Salient Implementation Details of Autonomous Vehicles

Autonomous vehicle software can be broadly decomposed into the following components: perception, planning, and control (Siciliano and Khatib, 2016). After reading this chapter, you should not necessarily know how to build an autonomous vehicle, but you should have a better idea of how they work. Understanding key technical concepts is necessary for understanding the social dilemma of autonomous driving.

I will approach this explanation in the same order that data flows through an autonomous driving system. First, autonomous vehicles enter the sensing stage, making use of the many pieces of hardware available to them. These sensors almost always include RADAR, LIDAR (which uses laser light pulses, whereas RADAR uses radar waves), and cameras, which sometimes have depth-sensing capabilities (J. M. Anderson et al., 2016). Furthermore, there are sensors connected to the motors that propel autonomous vehicles. These sensors monitor and publish odometry data (e.g., estimated position, velocity, acceleration, and orientation). These sensors represent the link between virtual autonomous systems and the physical world around them. Lastly, autonomous vehicles make extensive use of GPS technology. GPS tracking provides valuable data for use in navigation, but it is not reliable enough to base the entire perception system on. If a car goes into a tunnel where it cannot contact GPS satellites, the car needs to be able to navigate its new environment safely. It will use the sensors it can get reliable data from to navigate through the tunnel. There is a failure case for each of the listed sensors, and we need to make use of all these tools if we wish to ensure the highest level of safety. Generally, this technique is given the name "sensor fusion."

Next, an autonomous system takes the data collected by its sensors and interprets

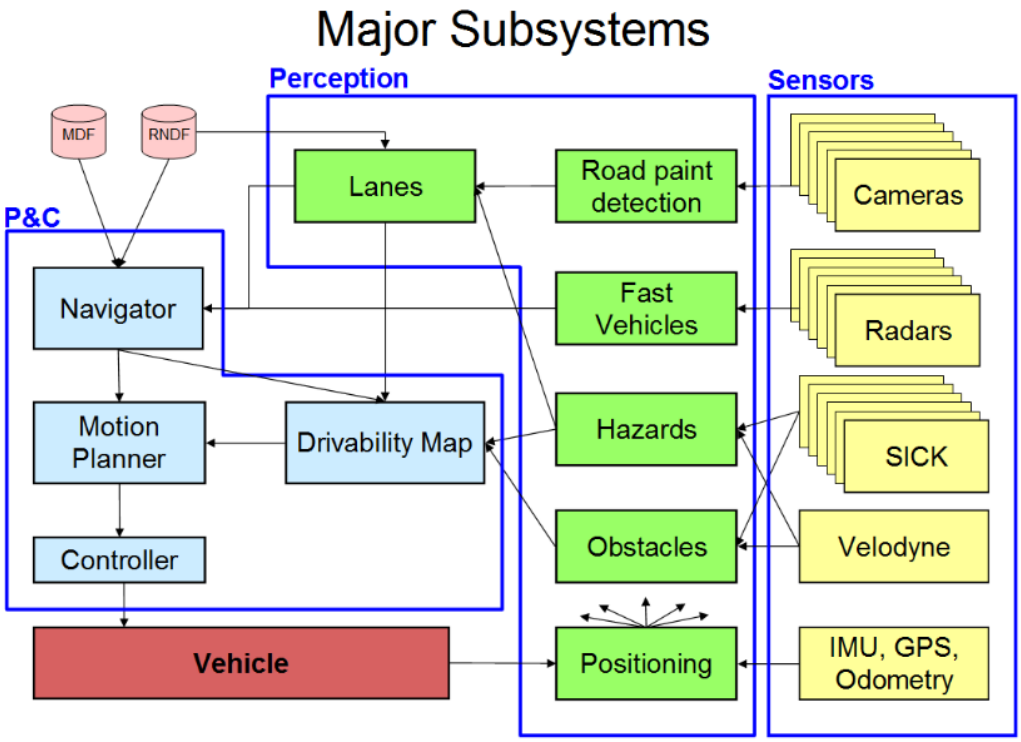

Figure 4.1 An example of an autonomous vehicle's architecture: MIT's *Talos* system architecture (Fletcher et al., 2009)

it. This is the perception stage. In this stage, one of the most important tasks is localization. Localization is the process of situating oneself within an environment. If all sensors were perfect, this would be a relatively straightforward task. However, they are not, and depending on the technology used and the number of sensors, the type and quality of information gathered differs (Holstein, Dodig-Crnkovic, and Pelliccione, 2018). Luckily, there has been a great amount of recent technological advancement in this area, including the highly influential SLAM algorithm, which performs simultaneous localization and mapping of a robot's environment (Grisetti et al., 2010).

Also during this phase, autonomous vehicles apply computer vision techniques to camera input to identify objects in their environment. Perhaps the most important task of computer vision is determining the location of road paint. Cars need to be

predictable and stay in their lanes if they wish to keep their passengers safe. Also, self-driving cars must analyze their environment for signs. Some signs are standard, such as those that instruct drivers to keep below a certain speed limit or to stop, but others are not so simple. Once these signs are identified, other systems on the car must be notified so they can adjust their behavior accordingly.

Obstacles and hazards are another important set of properties autonomous vehicles must detect. This can be done through a combination of techniques, utilizing both computer vision and LIDAR or RADAR sensing. However, some objects are harder to detect than others, such as potholes, rocks in the road, pedestrians, or fast-moving vehicles. Autonomous vehicles need to classify obstacles appropriately (treat pedestrians as pedestrians and cars as cars) because different behavior is necessary for different obstacles. For example, we are more likely to swerve off the road when a pedestrian is detected than when a rock is detected. We value saving the lives of pedestrians more than avoiding rocks. Furthermore, some of these obstacles require different sensing techniques. Fast vehicles can be identified better by RADAR, and using cameras and computer vision is the best way to classify pedestrians as pedestrians and cyclists as cyclists. While we can tolerate faults to a point by combining several methods, each specialized sensing system should be kept up at all costs.

After the perception phase has been completed, planning begins. This is perhaps the trickiest stage because one cannot test for errors the same way one can test for sensor malfunctions or braking failures. Planning malfunctions are more subtle. These problems lie at a higher level of abstraction, but they have serious consequences if implemented incorrectly. The planning phase can be roughly split into navigation (or long-term planning) and short-term planning. Navigation takes the current state of the car (position, orientation, etc.) and determines what steps are necessary to reach some long-term goal. Long-term goals are usually locations.

Short-term planning deals with more immediate concerns. Should the car accelerate? Should it turn? These questions are answered based on the results of the perception stage, and other data is collected from GPS data and local drivability maps. Short-term goals need to avoid immediate danger, but they also need to make some progress is satisfying the long-term goals, as well. In short, balancing short-term and long-term goals is difficult but not impossible. It takes a fine-tuned model to both ensure safety and travel from point A to B.

Additionally, there is a great amount of research being done currently on V2X communication, especially with the advent of 5G cellular technology (5G Automotive Association, 2016). 'V2X' is a technology providing an interface for vehicles to communicate with their surroundings, standing for vehicle-to-everything. Here, 'everything' encompasses vehicle-to-infrastructure communication, vehicle-to-vehicle, vehicle-to-pedestrian, vehicle-to-device, and vehicle-to-grid communication. I will use the term 'V2X' to refer to general communicative technology between a vehicle and anything else. With V2X communication, vehicles could nearly instantaneously survey the safety of the road they are traveling down miles ahead of their current position, whereas human drivers are limited to the extent of their vision and any road signs placed around hazards. V2X needs a large bandwidth to work effectively, but it could have a large impact on the effectiveness of autonomous accident-avoidance and the overall safety of self-driving cars.

With the ability to communicate between actors on the road, we are faced with the opportunity for cooperative driving. This compensates for at least one of autonomous vehicles' weaknesses: the inability to predict the actions of human beings through intuition, for example, when eye contact is made to ensure a shared understanding between human drivers (Fletcher et al., 2009). With V2X communication, vehicles can communicate with each other about their plans before they even begin to execute

them. Moreover, vehicles could make collectively safer decisions in the planning phase, giving an even stronger safety guarantee to those on the road and taking each car's ethical settings into account.

Finally, the control phase begins. Autonomous vehicles translate the plans they have made into control commands. These commands differ based on the hardware available to a given autonomous vehicle, but every vehicle will have control over acceleration, braking, and steering. It is the combination of the control phase with the previous steps (sensing, perception, and planning) that makes programming autonomous vehicles difficult. Time passes between each step, and assumptions that were made in previous computations may no longer hold. To make matters even more complicated, several of these different stages could be running concurrently, depending on the architecture of the system. There is a tradeoff between simplicity and performance, and many of the operations outlined above are computationally expensive. Autonomous vehicles need to ensure safety for their passengers, but they also need to react quickly to changes in their environment. The tradeoff between performance and simplicity can be seen throughout computer science, and self-driving cars are no exception.

If all these stages function correctly, and they are fitted together correctly, self-driving cars will be safer than traditional human-driven cars. If these cars avoid danger much better than human drivers, we might find ourselves in a situation where there might not even be a steering wheel in vehicles. Should we trust our vehicles enough for this? There is a tangible benefit: the driver of a car without a steering wheel becomes just a passenger during the autonomous journey, and he or she can take their hands off the steering wheel and pedals and pursue other activities (Maurer et al., 2016). Those with a long commute would get years of free time they would have spent driving.

However, people are already wary of autonomous vehicles. Knowing that there is no human fail-safe only adds to that fear. The NHTSA's regulations make rear-view mirrors and steering wheels mandatory (Adkisson, 2018). These regulations no longer make sense when cars can drive themselves, but we must get to that point first. Either the regulations should be changed (as they are expected to be by the end of the year), or we should not allow self-driving cars to operate without the use of a human-steering fail-safe.

Complexity is another way to think about ethics in autonomous systems. That is, as the amount of input (like sensor data, data from GPS, and V2X communication) increases, the difficulty of solving a given ethical dilemma increases. Suppose a car is in a scenario where it must choose between crashing into car A or car B. Further, suppose that this car can determine the identity of the passengers of each car. Car A contains three sixteen-year-old girls, and car B contains a sixty-year-old renowned international human rights lawyer. If the autonomous vehicle's accident avoidance algorithm takes this data into account, the ethical dilemma begins to take on more features. From a moral perspective, treating people differently based on their age, race, socioeconomic status, or profession — as done in Awad et al. (2018) — seems unethical. In fact, this very practice is illegal for AI to perform in Germany as of 2017, as previously mentioned (Bundesministerium für Verkehr und digitale Infrastruktur, 2017).

We must consider the complexity of the scenarios we find ourselves in because there are practical limits to both computational and ethical complexity in the type of solutions AI can provide. Identifying the most impactful ethical and computational considerations is one of the most difficult goals in modeling solutions to ethical dilemmas. The most impactful goals will become the highest weighted values in the weight function autonomous vehicles use to make decisions. It is probably impossible to make

the morally best decision in each scenario even without a time constraint. In reality, we are constrained by time, as well, and if we try to compute too many features in the decision-making process, we may never make a decision. Most autonomous vehicles run with less than 0.05 seconds between each update. All of the systems described in this section (sensing, perception, planning, and control) need to execute in this time frame, so autonomous driving software engineers must consider efficiency as well as correctness. And while decisions can be carried over between frames, conflicting data might present itself, further complicating the process. Efficiency is a persistent factor in designing autonomous driving software.

4.3 Programming Autonomous Vehicles With Ethical Principles

There are two general types of approaches we can take to programming autonomous vehicles with ethical principles: top-down and bottom-up. The top-down approach operates using some top-level rules and applying these logically to various scenarios to make a decision (Wallach and Allen, 2009). The bottom-up approach operates by taking specific examples of decisions and connecting the principles behind these scenarios to "learn" how to act ethically. Each of these techniques has strengths and weaknesses, but their implementations are radically different. Because there has not yet been enough work in this area, we must take both ethical viability and ease of implementation into account.

When faced with scenarios like the Trolley Problem, autonomous vehicles do not decide between two choices, as experiments like the Moral Machine might make it seem (Awad et al., 2018). If allowed enough time, humans might analyze the problem like this (i.e. choosing between two difficult choices), but they would not have enough time if they were driving. Humans react primarily according to instinct in such

scenarios. In contrast, autonomous vehicles create a model of their environment in the perception stage and form a set of potential actions to take based on this model of their environment. Each of these actions might be represented as a steering angle, an acceleration value, and a braking value. For each of the potential actions available, autonomous vehicles predict the results of the action. They use these predicted results as input to some weight function. Results that are valued by a system are given positive weight value (e.g., 5 points for braking a safe distance from the car in front of you, 10 points for staying within a meter of the planned route). Unfavorable results are given a negative value (e.g., -1000 points for each person with a predicted 100% chance of death, -500 points for each person with a predicted 50% chance of death).

After simply selecting the highest weight value, an autonomous vehicle decides which action it will take. Thus, the weight function contains most of the ethically-charged information in an autonomous system. When we discuss top-down and bottom-up approaches to encoding ethical principles, we are really discussing how we construct the weight functions autonomous vehicles use. This is an important distinction because it is easy to confuse explicit ethical programming for making an explicitly ethical decision (like we do in the Trolley Problem).

Top-down ethical programming is clear, but it suffers from many of the issues discussed in Chapter 3. This class of techniques focuses on the principles a system follows, but what if we cannot agree on what the correct principles should be? After all, is this not what philosophers have been trying to do for millennia? In Section 3.1, I presented and analyzed consequentialist and deontological frameworks for autonomous decision-making. Each of these frameworks was a top-down approach. Because the ethical knob configures the settings of these frameworks, it is a top-down approach, as well. The ethical knob alleviates problems that arise from picking how

to model ethics in a top-down system.

Logic-based approaches to top-down ethical programming are particularly promising. If we can manage to describe a system's ethical principles in a rigid, logical form, we can use established theorem proving software to prove the ethicality of a system. Most software engineers would not want to write functionality for their applications with the level of verbosity this logical form requires, however, so perhaps user-friendly applications could be developed to help this process. Otherwise, another role on software teams could appear with the sole purpose of translating code functionality into logical expressions. Researchers have proposed methods through which "standard deontic logic" (one of the most commonly used systems for expressing logical statements) can be adapted into a more AI-friendly version, allowing engineers to describe their AI's actions in provable, clear terms (Bringsjord, Arkoudas, and Bello, 2006).

Correctly translating the functionality into logic is not trivial (and might not even be possible with machine learning involved), but if we assume that logical statements for an artificial agent reasonably match its functionality, we can be more sure of its reliability. If one were to implement provable Benthamite Utilitarian moral calculus in an autonomous vehicle's software system, we would be able to conclude that this car can follow the theory of act utilitarianism at least as well as human beings and, perhaps, even better, given the small amount of information that humans use to routinely make decisions (J. M. Anderson et al., 2016). This is not to say that machines might be morally better than humans if we can prove their actions adhere to some theory. In fact, the claim that artificial intelligence can even be moral is debatable. We can, however, trust provable algorithms more than their human counterparts.

Furthermore, any description of top-down ethical programming would be incomplete without mentioning Asimov's Three Robot Laws (Asimov, 1950):

First Law: A robot may not injure a human being or, through inaction,

allow a human being to come to harm.

Second Law: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws

These laws represent a high-level set of principles that robots in *I, Robot* were expected to uphold. Each of the robots' smaller decisions flows from its permissibility in the context of the Three Laws they have been given. This is an excellent example of a top-down ethical programming paradigm, but it does not tell us how to formulate these laws in more concrete, provable terms. Ostensibly, the semantics of the translation of these plain English laws into laws that robots can understand would be debatable.

Additionally, if these ethical rules are separate from the code that facilitates the decision-making process itself (i.e. the principles are a separate architectural element), then they could be treated differently than the decision-making code. In other words, we might design ethical machines modularly, so that ethical principles could be modified by those who should be making the ethical decisions for a system. Most software engineers have little knowledge of ethics, and they might not feel comfortable making decisions for others. If ethics were treated as a separate architectural element, ethicists would be able to work closely with someone familiar with the given system and create a suite of application-specific principles. These decisions are critical, and those who make them should be qualified. The people who are good at creating autonomous systems are not necessarily the people who are good at determining how to program ethics into autonomous vehicles.

Moreover, some of the principles used in one autonomous system might be similar to those in another autonomous system. These similar principles could be central-

ized, forming a set of common ethical principles. These common principles might even be offered in an "ethics-as-a-service" architecture. Latency would need to be accounted for, so the ethics service could be stored locally and downloaded on each new update. Ethics services could either be created by for-profit businesses or some open-source/non-profit organizations. Getting all of the ethicality of an application into one condensed project would help make its ethics explicit, and it would allow people who can reason about ethical decisions to take part in building autonomous systems.

On the other hand, bottom-up ethical programming is less transparent, but it is easier to implement and avoids having to take a stance on difficult ethical dilemmas. This approach to programming ethical values takes individual cases and fills in the gaps between them to form a sort of ethical mesh. There are no hard principles with bottom-up ethical programming — only examples to follow. Novel scenarios are evaluated for their similarity to previously evaluated examples. In a sense, this is the more organic approach to decision-making (Wallach and Allen, 2009). After all, most humans do not cite their preferred ethical principles when making routine decisions. They relate potential decisions to decisions they have made or seen others make.

Using this approach, we could rig human-driven cars with equipment that measures their actions and the environment around the car to which they are responding. Once this human driver makes a decision that we might want to teach autonomous vehicles to make (or not make), we could capture the environment's current state and create a data point out of the decision made. An interesting application of this technology would be to allow a car owner to train his or her car on their personal driving style (Kuderer, Gulati, and Burgard, 2015). Because humans are capable of unethical behavior, this might not be preferable when trying to create the best moral agents possible. Nevertheless, it is an interesting and immediately feasible technique

to embedding bottom-up ethics.

However, we must consider the mechanics of moving from cases to ethical principles. If this process made use of neural networks, the technology, by its nature, would obscure our view into the decision-making process. Neural networks essentially match patterns, regardless of the reasoning behind this matching. If the decision-making process does not use neural networks, we must classify the features of the data ourselves. With neural networks, we had the advantage of considering each feature in the data, albeit implicitly. When considering features explicitly, we run the risk of being overly reductive, thereby lowering the accuracy of our results. Thus, we are faced with another tradeoff — this time between transparency and accuracy. Believers of moral particularism would find this technique appealing because of its implication that principles should not be our focus, but rather actions in particular situations. However, one can simultaneously reject moral particularism (which says that no moral principles are defensible) and support bottom-up ethical programming as a practical solution to this theoretical problem.

Suppose a bottom-up ethical programming system compares a car's current state to its database of states and their respective human-made decisions, and it outputs the state in the database that is closest match to the current state. Further suppose it also outputs the degree of similarity between these two situations. If this degree of similarity is very low (maybe the most similar state is only a 10% match), then it might not make sense to take the decision that is tied to this state. There are two ways of attempting to circumvent this problem. First, we could simply add more entries to the database in an attempt to raise the degree of similarity in the worst case. However, this is not a perfect solution because we cannot have an infinite number of entries. There will always be a situation where we do not have a similar precedent.

The second way of solving this problem is to implement a similarity threshold.

With this technique, we might use another technique to make a decision if the degree of similarity is too low. Top-down ethical programming is the most apparent technique to use here. In the end, a combination of top-down and bottom-up techniques is probably the safest strategy. Each technique would be able to cover the other's weaknesses, and we would be able to simultaneously mimic the average driver while allowing explicit ethical customizability for autonomous car passengers.

4.4 What Goes Wrong

Previously, I have analyzed how autonomous vehicles make ethical decisions and how we should correctly trust autonomous vehicles to make those decisions. In this section, I will analyze scenarios where these two problems have been acceptably solved, but there is an error in their execution. Undesirable outcomes might occur as a result of what happens after these decisions are made.

Broadly, autonomous vehicles produce negative results when they are placed in situations for which they were not appropriately prepared. Sometimes the decision to deploy an unprepared autonomous vehicle is knowingly made, but most of the time, a lack of transparency into the decision-making process prevents the creators of autonomous systems from knowing how their systems will react in risky scenarios. Cars are seldom faced with decisions similar to the Trolley Problem, and their behavior in such scenarios might not even be known before an autonomous vehicle is allowed out of the testing phase. Additionally, even if each part of an autonomous software system is tested extensively, differences between test environments and real-world environments can cause errors with sensors or computational latency, leading to serious malfunctions.

The set of autonomous vehicle failures mainly comprises the following categories: errors in each of the individual systems outlined in Section 4.2, integration errors,

and malicious attacks on an otherwise functioning system. Each of these respective types of error can be attributed to an underlying issue with latency, hardware failure, a software bug, or a mismatch between human and AI expectations.

In general, sensors in robotics are configured to deal with what they encounter often. When sensors detect something unexpected, there is more potential for things to go wrong. Some sensor systems are extremely robust, while others cannot be trusted to the same extent. For example, LIDAR sensors have trouble with detecting reflective or extremely dark surfaces (e.g., black rubber or coal). Unreliable sensors should be treated as such, and weaknesses in individual sensors can be avoided with sensor fusion, weaving together a mesh of sensors with different weaknesses. Additionally, every sensor has noise, and its output needs to be tuned to the correct threshold. Autonomous driving engineers are experienced at solving these problems, yet mistakes are still made. There is a diverse, multi-dimensional set of failure vectors, and they interact differently based on the hardware, software system design, and the sensor data on a vehicle.

Errors occur at the perception stage, as well. These errors are generally visible, but their results remain in the software layer, so detecting them takes more effort. The perception stage uses computer vision techniques to make sense of camera input data. Similarly, autonomous vehicles use computer vision to make sense of street signs (e.g., speed limit, stop signs, yield, etc.). One of the worst ways an autonomous vehicle could fail is to lose track of lines on the road or miss a sign that would have alerted a human driver of danger. And because computer vision makes extensive use of deep learning, we cannot be sure that cars will not make these errors.

For an example, we need not look further than the Uber crash Tempe, Arizona. The Uber car classified a pedestrian correctly initially, but on subsequent frames, it classified her as a cyclist. However, this autonomous vehicle kept re-classifying

the pedestrian, and because of this, it kept resetting a timer. This manifested as a control phase bug because the car did not stop, but it originated from a problem with perception. The pedestrian was not classified correctly, and something terrible happened because of it.

Perception systems are often unprepared when anything unexpected happens. For example, take two cars driving down a road under construction. There is a temporary electronic street sign alerting drivers of construction a quarter mile down the road. One car is human-driven, and the other is autonomous. The human driver will be able to respond appropriately to the situation by reading the sign, but autonomous vehicles do not have the same response by default. We expect human drivers to be able to adapt to this scenario, but we cannot expect the same from autonomous vehicles. Even if the software system on a car adapts to this scenario, there will always be another to which they have not. Additionally, we cannot easily correct these errors during runtime. The only way to correct an error like not reading a sign is for a programmer to revisit the code base and diagnose the issue.

The perception stage also performs the localization process, through which an autonomous vehicle estimates its position. Errors in localization (i.e. a car believes itself to be in the incorrect location) can easily confuse autonomous vehicles. Most of the time when driving through unknown territory, autonomous vehicles will not crash into anything if their obstacle avoidance system is still running, but the chance of a crash does increase. A disoriented autonomous vehicle will struggle to reorient itself. While most autonomous vehicles make use of robust localization algorithms like Graph SLAM, it is difficult to create a localization strategy that works in all scenarios. Maps, LIDAR/RADAR sensor data, and camera data must be adequate to perform localization effectively.

Planning errors can occur with insufficient map information (e.g., new traffic pat-

terns, detours) or with an incorrect balance between long and short-term goals. Short-term goals that involve avoiding danger should always take precedence, but if we expect to take the steering wheels out of cars, we need to be able to trust that they will satisfy long-term goals, as well. If all other autonomous vehicle software systems are working appropriately, errors in planning will not usually result in immediate danger. Nevertheless, danger could occur if cars take passengers on a risky route, either from interpersonal violence or physical features of roads. While these potential risks merit consideration, they are secondary concerns to ensuring that autonomous vehicles do not harm their passengers or surrounding people in the act of driving. Additionally, if a car takes a bad path, it will drive farther. Driving more means consuming more fuel, and by extension, creating a bigger impact on the environment. Thus, errors in path planning directly correlate to self-driving car emission levels.

Control errors have more impactful, directly physical effects. Errors in control can occur if software and hardware believe themselves to be in the same conditions when they are actually in different conditions. A car's environment could have changed in the time it took for the software to run, or the software could have been misconfigured from the start. Odometry data scaling is a simple example of a configuration that affects the reliability of the larger system. If a car's control module tells it to travel at 60 miles per hour, but, in reality, the car is traveling at 70 miles per hour, the distance the car needs to safely brake will dramatically increase. When the car needs to stop, it might not have enough room because the calculations were made for a different set of conditions. Other control errors can occur when the environment changes within the latency period. Cars predict their environments during this latency period and act accordingly, but predictions will not be accurate 100% of the time. Thus, we get an element of error from this process, as well.

I have presented many error cases for autonomous vehicles in this section. Some

are more pressing than others, but the sheer number of places to go wrong in programming autonomous software should be telling of the difficulty involved. While there are theoretically immense benefits to putting self-driving cars on the road, we must ensure that they are safer than human-drivers if we wish to make good use of the technology. Releasing fully autonomous vehicles too early will create public distrust, and it would be hard to regain this trust, no matter how much data they are shown.

Lastly, autonomous cars can fail because of malicious attacks. In this thesis, I argue that more transparency in autonomous driving systems will be ultimately beneficial, but this causes a side effect: malicious actors also benefit from transparency. So far in this section, I have only considered failures caused by implementation mistakes, but each of the systems above could also be sabotaged. If we wish to reap the benefits of transparency (which probably outweigh the detriments), we must prepare for the worst to happen as well.

In computer security, it is widely accepted that a system is not truly secure if it relies on so-called "security through obscurity." Security is not impossible for digital technologies to achieve, but those that are secure are often not vulnerable to the same kinds of physical attacks one can launch on autonomous vehicles. For example, someone could fly a drone above the road and project false white lines, directing a self-driving car into oncoming traffic. Computer vision algorithms cannot tell the difference between fake and real road lines. After all, most humans might get confused in the same situation. The major difference between human and computer reactions in this scenario is reaction time. If fake lines or a speed limit sign are projected for even an eighth of a second, an autonomous vehicle might sense the projected images and change its behavior before the next second has even started (Nassi et al., 2020).

4.5 Actionable Recommendations for Developers

The insights in this thesis can be translated into various strategies for better designing autonomous vehicles and ensuring their reliability. In this section, I recommend several ways to make a positive impact, I do not pretend to know the only ways to accomplish this, and without a doubt, our techniques for designing and implementing 'moral machines' will evolve, just as our techniques for designing and implementing traditional software have evolved. Here, I recommend the addition of ethics testing to software testing pipelines, techniques for stronger certainty for reliability. I raise concerns with autonomous vehicles' use of continuous deployment, a software development process that delivers changes to customers quickly and automatically.

To navigate an ethical dilemma with AI is a complex problem in itself, but to make matters worse, negative effects are difficult to effectively test without an accident happening. Because the systems we build to solve these problems must be complex, there are multiple components, each of which is probably owned by a team made up of several people. These systems are so large that one person cannot reasonably understand everything that happens during runtime. As a result, developers' intuitions are dampened.

However, this is not a new problem. Complex software systems have existed, worked well (most of the time), and become the backbone of our society in the past few decades. One method of dealing with the problem of complexity is the creation of automated software testing, which is applied to each proposed change to a system and monitored for errors. This concept of automated testing encompasses many different tools, ranging from static code analysis to the simulation of a production environment. Running a collection of tests like this (sometimes abstracted into a single concept: the "testing pipeline") is vital to ensuring any semblance of a guarantee of reliability

in one's software.

For the most part, however, testing has only been applied to reliability in execution. Here, we are concerned with the reliability of ethical decision-making in the course of execution. I propose that software developers add ethics tests to testing pipelines, detecting inconsistencies between system-determined principles and the results of system execution. While it is true that detecting ethical inconsistencies is generally not as clear cut as detecting bugs in execution, we can focus on several aspects of ethics for which negative effects can be determined in testing. In short, it is not possible to solve ethical dilemmas through testing, but ethical testing will allow one to see errors in implicit ethical principles or bugs in the execution of these opaque ethical decision-making processes that were not previously visible.

Bias can appear in several ways when discussing autonomous vehicles. In the course of accident avoidance, the concept of ethical complexity comes back into play. The number and type of features that we consider in a scenario can allow biases to seep into the decision-making process. To illustrate this, consider a V2X interface that allows communication between vehicles to include the number of passengers inside each car, its safety rating, and the probability that its passengers will survive a 60 mile per hour crash. Another V2X interface might contain these features but add in options to specify the race, gender, and net worth of its passengers. If an accident avoidance algorithm discriminated based on this data (e.g., decided to run into the person with the lowest net worth when looking for a tie-breaker), most people would agree that it is implicitly biased.

One of the most promising ethical testing tools, called Themis, detects for discrimination in an application (Galhotra, Brun, and Meliou, 2017). It works by asking you to model your own application in a configuration file, rendering the ethical decisions your system makes explicit. It takes this configuration file and creates a specialized

Figure 4.2 Four scenes from CARLA, a simulator for urban driving (Dosovitskiy et al., 2017)

testing suite, which is tailored to your application's needs. However, notice that the configuration must be self-reported. This is a double-edged sword. On one hand, it forces system creators to be aware of the ethical decisions made by their system, but on the other, there is potential for abuse by either unintentional or willful ignorance. Even if we assume that there are no malicious actors, these systems are complex, and there likely is not a single person who can faithfully complete this analysis. If the work is spread across several teams, it is much more likely that this sort of testing will be the first thing to be thrown out when in a time crunch.

One of the most important things in ensuring ethical actions by autonomous vehicles will be to create a focus on making embedded values explicit. We cannot possibly weed out malicious actors, but by making the values in a system and their effects more explicit, we can make a societal move toward transparency and responsibility in designing and building autonomous vehicles. I will explore an architectural approach to

solving this problem in Section 4.3, but at the very least, employing an explicit, provable language to express the ethical decisions autonomous vehicles make will greatly improve transparency with these systems (Bringsjord, Arkoudas, and Bello, 2006). Even if the formalizations used to describe the decision-making process were not used as a part of the computation itself (i.e. they were included in the documentation), developers would be able to better reason about the ethics of their systems. Greater transparency will help developers make fewer mistakes and foster greater public trust in the technology.

Nevertheless, we cannot rely on self-reporting to detect all ethical breaches in autonomous-decision-making processes. Simulation is another technique that holds promise. Nvidia Drive Constellation is a virtual reality autonomous vehicle safety testing simulator (Corporation, 2019). Developers can make use of this framework or others like CARLA to test for negative effects once they modify their autonomous driving software (Dosovitskiy et al., 2017). These simulators can put vehicle software in rare or dangerous scenarios, testing for their robustness in handling unexpected situations without having to risk human lives. Developers could analyze the results of these simulations to detect behavior that does not mesh with their intended ethical rules. Simulation also has promise in the process of making real-time ethical decisions as well. An 'ethical layer' can be used to apply the simulation theory of cognition to artificial intelligence (Vanderelst and Winfield, 2018). This form of 'real-time testing' of ethical decisions lessens the gap between the simulator and the real-world.

However, if one hopes to ensure anything through simulation, it must have a good model of the environment the system will be run in. Many simulation trials with this model are necessary to achieve any semblances of assurance for ethical behavior or reliability. Running these trials becomes computationally expensive quickly, having a non-negligible temporal and financial impact.

Furthermore, software is never perfect the first time, nor will it ever truly be perfect. With this many variables, it is hard to determine what "correct" even is. Nevertheless, we must try to get there. If changes need to be made, they should be made reliably. Updating a 3,000 pound 120-mph top speed vehicle is not the time to "move fast and break things." Safety should not be compromised for development speed or cost. Simulators like Nvidia Drive Constellation or CARLA could be added to deployment pipelines and automatically spin up when a new change is deployed. They could test the system for a sufficient number of trials and pass it along the deployment pipeline. As of now, there are no guarantees for how well autonomous driving software will work, and it would help solve the problem of opacity if such testing frameworks were utilized in this manner. However, it is estimated that vehicles should cover around 11 billion miles to demonstrate with 95% confidence and 80% power that they fail less often than human drivers (Kalra and Paddock, 2016). Any time the system is modified, another 11 billion miles would be required to have the same degree of certainty. While it sounds less than ideal to trust autonomous vehicles without guarantees to their safety, we must make do with the tools presented here until researchers create tools with better accuracy.

Lastly, software developers should not use continuous deployment for autonomous driving software when testing environments are so inaccurate. In the process of continuous deployment, every change that passes the testing phase gets released to customers automatically. Often, continuously deployed software is less stable because the feedback loop between the developer, tester, and customer is rapidly accelerated. Testing for autonomous vehicle software is inconsistent, and errors cannot be detected as easily as with web applications. Autonomous driving systems have the potential to cause catastrophe once they reach widespread use. A single bug could be replicated across thousands or millions of vehicles. In my opinion, only emergency updates

should be allowed to be pushed automatically.

In all, there are many ways we can improve autonomous driving software. Software has become ubiquitous in our society and the importance of its quality has increased. Today, automation, advances in machine learning, and the availability of vast amounts of data are leading to a shift in how software is used, enabling the software to make more autonomous decisions (Goodall, 2016b). We must focus on making these autonomous decisions as safe as possible through the techniques I have described and more.

Chapter Five: Conclusion

In all, introducing autonomous vehicles into our society will have wide-ranging and deep impacts on many different aspects of our lives. In this thesis, I have pulled in topics from philosophy, computer science, robotics, and psychology. This topic demands interdisciplinary treatment because it will affect each of these disciplines, as well as each of us personally. Because there is so much trans-disciplinary flow on topics like creating accident-avoidance models, we need to understand the technical complexities of such a system, to understand the ethical complexity of a car's actions in various scenarios, and, ideally, to see the topic from both sides simultaneously.

Complex problems do not always necessitate complex solutions, but they do need solutions. With the analysis given in this thesis, hopefully, one can see that, even if it is difficult, we have ways of codifying solutions to ethical dilemmas that are better than others. While it is difficult to say whether to kill fewer people or obey Kant's Categorical Imperative when faced with the Trolley Problem, we can determine that either of these solutions is better than turning autonomous vehicles into killing machines — whether intentionally or not. With autonomous vehicles, we are generally worried about the unintentional ethical mistakes in decision-making processes that might have substantial ramifications. Autonomous driving software will be replicated across multitudes of vehicles globally.

As much as we wish to debate whether the Utilitarian or Kantian view is correct, we must find a way to proceed if we wish to avoid making the world's collection of vehicles into an oversized arsenal. The development of new technology has histori-

cally demanded new ethical considerations, and because programming is an ethically prescriptive action, we must soon decide how we wish to encode the decisions we make into the technology that will soon be so impactful. Currently, we are functioning with implicit ethical systems integrated into our society, but as decision-making is increasingly taken over by artificial intelligence, we must ensure that intelligent systems do not become misguided.

There are promising strategies for implementing ethical principles in both the top-down and bottom-up approaches, but the most realistic option is a hybrid system. Basic ethical principles like "try not to kill" and "never steal" are agreed upon by many ethical viewpoints, and these can be combined with the ethical knob. This approach avoids the problem of prescribing a specific belief on all users of a product, and it is clear about the things important enough to make explicit. We can combine this approach with bottom-up ethical programming, simulating the human decision-making process using pre-approved cases.

Allowing users to configure the ethical settings on their vehicles (within a pre-defined range of acceptable behavior) also seems like a promising way forward. The success of autonomous vehicles depends on reaching popular appeal, but we should not let vehicles into consumer markets until they employ the ethical techniques described in this thesis, and we are sure they will make a positive impact on the world. We cannot allow a car with "full egoism" to reach markets.

Furthermore, we must perform extensive testing on autonomous systems to find the limits of what we are willing to allow, and we must attempt to ensure reliably that vehicles operate within these limits. The continued development of autonomous driving simulators and ethical testing frameworks will allow us to come closer to achieving these goals, even if we may never find perfect answers.

Lastly, the Trolley Problem is not the only ethical dilemma faced by autonomous

vehicles. We must stay vigilant in the course of researching, implementing, and living with autonomous vehicles, staying alert for any ripple effects. Self-driving cars will bring benefits to the world, but we do not know the full extent of their effects until we reach the point of wide-scale adoption.

Lastly, if for-profit companies should be required to implement any of the recommendations I have given in this thesis (or any like them), government agencies should enforce them to do so. Additionally, some of the tools I have discussed might improve the reliability of autonomous vehicle software, but they are not appropriate for government agencies to mandate. Professional societies like IEEE or ACM or open-source initiatives could offer these supplementary tools to developers.

The scholars across fields who have already begun the monumental task of determining how to correctly trust and implement autonomous vehicles have done our future society a great service. However, there is still a great deal of work to be done. I would like to call for even more attention to be brought to several key issues.

5.1 Future Work

Moving forward, we should focus on making ethical programming models easier to understand and more accurately represent our human values. There are several ways to go about this: we could implement more user-friendly top-down ethical programming systems (i.e. improving the input method for abstract principles) or hybrid ethical programming systems (combining top-down and bottom-up). Perhaps events led by non-profits or the government (like the DARPA challenge) would inspire developers to focus on the most pressing problems.

It would also be useful to see more algorithms based on more experimental ethical frameworks. Currently, many projects employ Utilitarian or Kantian ethics, and only a few implement ethical frameworks outside these. We also currently have a

large focus on individual-based ethical frameworks. We need to find better ways to collectively resolve dilemmas — democracy might not always be the appropriate solution.

Lastly, there will be a variety of secondary effects and dilemmas that will arise as autonomous vehicles further integrate into our society. It is difficult to predict these issues now, but we are safe in assuming that there is much work to be done on issues we have not yet conceived.

References

5G Automotive Association (2016). "The Case for Cellular V2X for Safety and Cooperative Driving". In: *5GAA Whitepaper*, pp. 1–8. URL: http://5gaa.org/..

AAA Foundation for Traffic Safety (2016). *RESULTS General Driving*. Tech. rep.

Adkisson, Samuel (2018). "System-Level Standards: Driverless Cars and the Future of Regulatory Design". In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.3122393.

Ahlenius, Henrik and Torbjörn Tännsjö (2012). "Chinese and westerners respond differently to the trolley dilemmas". In: *Journal of Cognition and Culture* 12.3-4, pp. 195–201. ISSN: 15677095. DOI: 10.1163/15685373-12342073.

Anderson, James M et al. (2016). *Autonomous Vehicle Technology: A Guide for Policymakers*. ISBN: 9780833083982. URL: www.rand.org/giving/contribute.

Anderson, Michael and Susan Leigh Anderson (Dec. 2007). "Machine Ethics: Creating an Ethical Intelligent Agent". In: *AI Magazine* 28.4, pp. 15–15. ISSN: 2371-9621. DOI: 10.1609/AIMAG.V28I4.2065.

Anderson, Michael, Susan Leigh Anderson, and Chris Armen (2005). *Towards machine ethics: Implementing two action-based ethical theories*. Tech. rep., pp. 1–7. URL: http://www.boeing.com/phantom/ucav.html.

Asimov, Isaac (1950). *I, Robot (The Isaac Asimov Collection)*. New York City: Doubleday, p. 40. ISBN: 0-385-42304-7. URL: https://isbnsearch.org/isbn/0385423047.

Awad, Edmond et al. (Nov. 2018). "The Moral Machine experiment". In: *Nature* 563.7729, pp. 59–64. ISSN: 14764687. DOI: 10.1038/s41586-018-0637-6.

Bolukbasi, Tolga et al. (2016). *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*. Tech. rep., pp. 4349–4357. URL: https://code.google.com/archive/p/word2vec/.

Bonnefon, Jean François, Azim Shariff, and Iyad Rahwan (June 2016). "The social dilemma of autonomous vehicles". In: *Science* 352.6293, pp. 1573–1576. ISSN: 10959203. DOI: 10.1126/science.aaf2654. arXiv: 1510.03346.

Bringsjord, Selmer, Konstantine Arkoudas, and Paul Bello (July 2006). "Toward a general logicist methodology for engineering ethically correct robots". In: *IEEE Intelligent Systems* 21.4, pp. 38–44. ISSN: 15411672. DOI: 10.1109/MIS.2006.82.

Bundesministerium für Verkehr und digitale Infrastruktur (2017). *Ethics Commission Automated and Connected Driving*. Tech. rep. 4158, pp. 38–38. DOI: 10.1126/science.186.4158.38.

Bureau of Labor Statistics (2019a). *Heavy and Tractor-trailer Truck Drivers : Occupational Outlook Handbook: : U.S. Bureau of Labor Statistics*. URL: https://www.bls.gov/ooh/transportation-and-material-moving/heavy-and-tractor-trailer-truck-drivers.htm%7B%5C#%7Dtab-1 (visited on 03/02/2020).

— (2019b). *Taxi Drivers, Ride-Hailing Drivers, and Chauffeurs : Occupational Outlook Handbook: : U.S. Bureau of Labor Statistics*. URL: https://www.bls.gov/ooh/transportation-and-material-moving/taxi-drivers-and-chauffeurs.htm (visited on 03/02/2020).

Contissa, Giuseppe, Francesca Lagioia, and Giovanni Sartor (Sept. 2017). "The Ethical Knob: ethically-customisable automated vehicles and the law". In: *Artificial Intelligence and Law* 25.3, pp. 365–378. ISSN: 15728382. DOI: 10.1007/s10506-017-9211-z.

Corporation, Nvidia (2019). *NVIDIA DRIVE CONSTELLATION VIRTUAL REALITY AUTONOMOUS VEHICLE VALIDATION PLATFORM DRIVE CONSTELLATION IS THE ONLY HARDWARE-IN-THE-LOOP PLATFORM THAT COMBINES THE FOLLOWING FEATURES*. Tech. rep. URL: www.nvidia.com/constellation.

Davis, Linda (Aug. 2003). "Carjacking — Insights from South Africa to a New Crime Problem". In: *Australian & New Zealand Journal of Criminology* 36.2, pp. 173–191. ISSN: 0004-8658. DOI: 10.1375/acri.36.2.173.

Dosovitskiy, Alexey et al. (2017). *CARLA: An Open Urban Driving Simulator*. Tech. rep., p. 16.

Fletcher, Luke et al. (2009). "The MIT - Cornell collision and why it happened". In: *Springer Tracts in Advanced Robotics* 56, pp. 509–548. ISSN: 16107438. DOI: 10.1007/978-3-642-03991-1_12.

Floridi, Luciano (Dec. 2016a). "Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2083. ISSN: 1364503X. DOI: 10.1098/rsta.2016.0112.

— (Mar. 2016b). *Mature Information Societies—a Matter of Expectations*. DOI: 10.1007/s13347-016-0214-6.

Floridi, Luciano (June 2017). "Digital's Cleaving Power and Its Consequences". In: *Philosophy & Technology* 30.2, pp. 123–129. ISSN: 22105441. DOI: 10.1007/s13347-017-0259-1.

Floridi, Luciano and Mariarosaria Taddeo (Dec. 2016). "What is data ethics?" In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2083. ISSN: 1364503X. DOI: 10.1098/rsta.2016.0360.

Fox, Sarah (2016). "Planning for Density in a Driverless World". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.2735148. URL: http://scholarship.law.georgetown.edu/ipr%7B%5C_%7Dpapers/1http://scholarship.law.georgetown.edu/ipr%7B%5C_%7Dpapers.

Galhotra, Sainyam, Yuriy Brun, and Alexandra Meliou (2017). "Fairness testing: Testing software for discrimination". In: *Proceedings of the ACM SIGSOFT Symposium on the Foundations of Software Engineering*. Vol. Part F1301, pp. 498–510. ISBN: 9781450351058. DOI: 10.1145/3106237.3106277. arXiv: 1709.03221.

Gerdes, J. Christian and Sarah M. Thornton (2015). "Implementable Ethics for Autonomous Vehicles". In: *Autonomes Fahren*. Berlin, Heidelberg: Springer Vieweg, Berlin, Heidelberg, pp. 87–102. ISBN: 978-3-662-45854-9. DOI: 10.1007/978-3-662-45854-9_5. URL: http://link.springer.com/10.1007/978-3-662-45854-9%7B%5C_%7D5.

Gogoll, Jan and Julian F. Müller (June 2017). "Autonomous Cars: In Favor of a Mandatory Ethics Setting". In: *Science and Engineering Ethics* 23.3, pp. 681–700. ISSN: 14715546. DOI: 10.1007/s11948-016-9806-x.

Gold, Natalie, Andrew M. Colman, and Briony D. Pulford (2014). "Cultural differences in responses to real-life and hypothetical trolley problems". In: *Judgment and Decision Making* 9.1, pp. 65–76.

Goodall, Noah J. (2014). "Machine Ethics and Automated Vehicles". In: *Road Vehicle Automation*, pp. 93–102. DOI: 10.1007/978-3-319-05990-7_9.

— (2016a). "Away from Trolley Problems and Toward Risk Management". In: *Applied Artificial Intelligence* 30.8, pp. 810–821. ISSN: 10876545. DOI: 10.1080/08839514.2016.1229922.

— (June 2016b). "Can you program ethics into a self-driving car?" In: *IEEE Spectrum* 53.6. ISSN: 00189235. DOI: 10.1109/MSPEC.2016.7473149.

Greene, Joshua (2013). *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*.

Grisetti, Giorgio et al. (2010). *A Tutorial on Graph-Based SLAM*. Tech. rep.

Holstein, Tobias and Gordana Dodig-Crnkovic (2018). "Avoiding the intrinsic unfairness of the trolley problem". In: *Proceedings - International Conference on Software Engineering*, pp. 32–37. ISSN: 02705257. DOI: 10.1145/3194770.3194772.

Holstein, Tobias, Gordana Dodig-Crnkovic, and Patrizio Pelliccione (Feb. 2018). "Ethical and Social Aspects of Self-Driving Cars". In: arXiv: 1802.04103. URL: http://arxiv.org/abs/1802.04103.

Kalra, Nidhi and Susan M Paddock (2016). "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" In: *Transportation Research Part A: Policy and Practice* 94, pp. 182–193. ISSN: 09658564. DOI: 10.1016/j.tra.2016.09.010.

Kant, Immanuel et al. (2002). *Groundwork for the metaphysics of morals*. Yale University Press, p. 194. ISBN: 9780300094879.

Kuderer, Markus, Shilpa Gulati, and Wolfram Burgard (2015). "Learning Driving Styles for Autonomous Vehicles from Demonstration". In:

Leben, Derek (June 2017). "A Rawlsian algorithm for autonomous vehicles". In: *Ethics and Information Technology* 19.2, pp. 107–115. ISSN: 15728439. DOI: 10.1007/s10676-017-9419-3.

Lin, Patrick (2015). "Why Ethics Matters for Autonomous Cars". In: *Autonomes Fahren*. Springer Berlin Heidelberg, pp. 69–85. DOI: 10.1007/978-3-662-45854-9_4.

Liu, Hin Yan (Sept. 2017). "Irresponsibilities, inequalities and injustice for autonomous vehicles". In: *Ethics and Information Technology* 19.3, pp. 193–207. ISSN: 15728439. DOI: 10.1007/s10676-017-9436-2.

Matthias, Andreas (2004). "The responsibility gap: Ascribing responsibility for the actions of learning automata". In: *Ethics and Information Technology* 6.3, pp. 175–183. ISSN: 15728439. DOI: 10.1007/s10676-004-3422-1.

Maurer, Markus et al. (2016). *Autonomous Driving Technical, Legal and Social Aspects*. Ed. by Markus Maurer et al. 1st ed. Springer-Verlag Berlin Heidelberg, p. 706. ISBN: 978-3-662-48847-8. DOI: 10.1007/978-3-662-48847-8.

Millar, Jason (2015). "Technology as Moral Proxy: Autonomy and Paternalism by Design". In: *IEEE Technology and Society Magazine* 34.2, pp. 47–55. ISSN: 02780097. DOI: 10.1109/MTS.2015.2425612.

Montemerlo, Michael et al. (2009). *Junior: The Stanford Entry in the Urban Challenge*. Tech. rep.

Nassi, Ben et al. (2020). *Phantom of the ADAS: Phantom Attacks on Driver-Assistance Systems*. Tech. rep. Cryptology ePrint Archive. URL: https://eprint.iacr.org/2020/085.

National Transportation Safety Board (2018). *PRELIMINARY REPORT HIGHWAY HWY18MH010*. Tech. rep.

Nourinejad, Mehdi, Sina Bahrami, and Matthew J. Roorda (Mar. 2018). "Designing parking facilities for autonomous vehicles". In: *Transportation Research Part B: Methodological* 109, pp. 110–127. ISSN: 01912615. DOI: 10.1016/j.trb.2017.12.017.

Nyholm, Sven and Jilles Smids (Nov. 2016). "The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem?" In: *Ethical Theory and Moral Practice* 19.5, pp. 1275–1289. ISSN: 15728447. DOI: 10.1007/s10677-016-9745-2.

Powers, Thomas M. (July 2006). "Prospects for a kantian machine". In: *IEEE Intelligent Systems* 21.4, pp. 46–51. ISSN: 15411672. DOI: 10.1109/MIS.2006.77.

Rawls, John (1971). *A Theory of Justice*. Harvard University Press, p. 538. ISBN: 9780674000780. URL: http://www.hup.harvard.edu/catalog.php?isbn=9780674000780.

SAE International (2018). *J3016B: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles - SAE International*. Tech. rep. SAE International. URL: https://www.sae.org/standards/content/j3016%7B%5C_%7D201806/.

Santoni de Sio, Filippo (Apr. 2017). "Killing by Autonomous Vehicles and the Legal Doctrine of Necessity". In: *Ethical Theory and Moral Practice* 20.2, pp. 411–429. ISSN: 15728447. DOI: 10.1007/s10677-017-9780-7.

Siciliano, Bruno and Oussama Khatib (Jan. 2016). *Springer handbook of robotics*. Springer International Publishing, pp. 1–2227. ISBN: 9783319325521. DOI: 10.1007/978-3-319-32552-1.

Taddeo, Mariarosaria (July 2010). "Modelling trust in artificial agents, a first step toward the analysis of e-trust". In: *Minds and Machines* 20.2, pp. 243–257. ISSN: 09246495. DOI: 10.1007/s11023-010-9201-3.

— (Dec. 2017). "Trusting Digital Technologies Correctly". In: *Minds and Machines* 27.4, pp. 565–568. ISSN: 15728641. DOI: 10.1007/s11023-017-9450-5.

Taddeo, Mariarosaria and Luciano Floridi (Aug. 2018). "How AI can be a force for good". In: *Science* 361.6404, pp. 751–752. ISSN: 10959203. DOI: 10.1126/science.aat5991.

Thomson, Judith Jarvis (1985). "The Trolley Problem". In: *Yale Law Journal* 94.6, pp. 1395–1415. ISSN: 20416962. DOI: 10.1111/j.2041-6962.1986.tb01581.x.

U.S. Department of Health & Human Services (2020). *Organ Procurement and Transplantation Network National Data*. URL: https://optn.transplant.hrsa.gov/data/view-data-reports/national-data/ (visited on 03/02/2020).

Urmson, Chris et al. (2008). "Autonomous Driving in Urban Environments: Boss and the Urban Challenge". In: *www.interscience.wiley.com). • Journal of Field Robotics* 25.8, pp. 425–466. DOI: 10.1002/rob.20255. URL: www.interscience.wiley.com.

Vanderelst, Dieter and Alan Winfield (May 2018). "An architecture for ethical robots inspired by the simulation theory of cognition". In: *Cognitive Systems Research* 48, pp. 56–66. ISSN: 13890417. DOI: 10.1016/j.cogsys.2017.04.002.

Wallach, Wendell and Colin Allen (Jan. 2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, pp. 1–288. ISBN: 9780199871889. DOI: 10.1093/acprof:oso/9780195374049.001.0001.

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell (Jan. 2018). "Mitigating Unwanted Biases with Adversarial Learning". In: *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340. arXiv: 1801.07593. URL: http://arxiv.org/abs/1801.07593.

Austin Atchley was born in Dallas, Texas on March 17, 1998 and moved to Austin while in elementary school. From the fall of 2016 to the spring of 2020, he studied Plan II Honors and Computer Science while enrolled at the University of Texas at Austin. He completed a study abroad program at SciencesPo Paris, Campus de Reims in the spring of 2019. After graduation, he will begin a position as a Software Development Engineer at Amazon Web Services in Seattle, Washington.