Ethical Understandings and Approaches to Autonomous Vehicles

Kathryn McVicker

HP 4976-11: Independent Study with Edward Sherline

May 2019

Table of Contents

Introduction

From Batman to Minority Report to Ender's Game, autonomous vehicles have been a prominent feature of science fiction, representing a futuristic society that could occur only in one's wildest imagination. However, that future is slowly becoming a reality. In 2009, the Google-run project Waymo began developing a self-driving car project. By 2013, General Motors, Ford, Mercedes Benz, BMW, and others announced development their own self-driving car technologies. Nissan committed to a launch date by announcing that it will release several driverless cars by the year 2020. By 2018, more than two million miles had been driven by Waymo's autonomous vehicle (Dormehl & Edelstein, 2019).

It is easy to understand why these companies have an interest in developing a marketable autonomous vehicle. In 2016, The National Highway Traffic Safety Administration reported that "human error is involved in 94 to 96 percent of all motor vehicle crashes" (Brown, 2017). There has also been a 5.6% increase in traffic fatalities between 2015 and 2016, increasing from 35,485 deaths in 2015 to 37,461 deaths in 2016 (National Highway Traffic Safety Administration, 2017). That means that a conservative estimate of the total traffic deaths caused by human error in 2016 is 35,213.34. Autonomous vehicles would virtually eliminate human error in traffic situations; autonomous vehicles have been predicted to eliminate 90% of traffic accidents due to their elimination of human error (Bonnefon, Shariff & Rahwan, 2016). If this prediction is accurate, then autonomous vehicles could save thousands of lives every year and could change the entire transportation system in America.

However, the development of autonomous vehicles has had its setbacks. On March 18, 2018, the world experienced its first pedestrian death by an autonomous vehicle when an Uber prototype struck a woman in Tempe, Arizona (Wakabayashi, 2018). This, combined with a

general distrust of the new technology, has led to vandalism and destruction of autonomous vehicles in Chandler, Arizona (Romero, 2018). The event has also brought the public's attention to questions about the ethics and legal responsibility of those who create and operate autonomous vehicles. What should an autonomous vehicle be programmed to do in the case of an unavoidable accident? Should autonomous vehicles have universal ethics programming?

This paper addresses the first question by comparing autonomous vehicles to the classic ethical dilemma The Trolley Problem. The author considers several traditional ethical theories, such as utilitarianism, deontology, social contract theory, and egoism. After the comparison to the Trolley Problem, the paper considers an ethically pluralist standpoint represented through the ethical knob.

An Applied Trolley Problem

The Trolley Problem is a classic ethical dilemma with countless variations and considerations. The Trolley Problem has changed over time, with Judith Thompson's modification becoming the "standard" for the modern Trolley Problem. Thompson's version involves a trolley barreling down a hill, unable to stop. If it continues on its track, it will hit and kill five workers working on that section of track. A bystander has control of a lever that could change the trolley's track from the track with five workers to a track with only one worker (Rehman & Dzionek-Kozlowska, 2018). This standard Problem maintains the central problem of knowingly sacrificing one person to save five.

Before even considering what actual ethical programming an autonomous vehicle might have, one has to consider the two approaches to ethical pre-programming that programmers might consider. Those approaches are universal ethics programming and user-selected

programming. The first approach is a universal ethics programming; with universal ethical programming, all autonomous vehicles would be programmed to respond in the same way in cases of unavoidable harm. This is the more traditional approach to pre-programming, and typically what philosophers have in mind when comparing a situation of unavoidable harm by autonomous vehicle and the Trolley Problem. Universal ethical programming does have some serious benefits. One of these benefits is a perceived fairness; if everyone's autonomous vehicle is programmed in the same way and everyone drives an autonomous vehicle, then there can be no claim of the programming being biased towards a specific group of people. Another benefit is prediction of maneuvers; if programmers know that all autonomous vehicles will act in the same way in a case of unavoidable harm, then they can theoretically program the autonomous vehicles more accurately and safely than if they had to guess what other types of programming other autonomous vehicles might have. Because of this, the first section of this paper will concern itself with what ethical theories to consider if universal ethics programming is a given.

When considering autonomous vehicles, by far the most philosophical literature has been written comparing them to the Trolley Problem. This is understandable. As one of the most famous ethical dilemmas, the source material is easy to access and can provide a framework from which to understand an aspect of the autonomous vehicle dilemma. Consider this scenario: an autonomous vehicle containing one passenger is driving on a two-lane bridge. A bus containing three passengers is coming in the other direction. Five pedestrians are walking along a viewing path to the right of the autonomous vehicle. Suddenly and without warning, the bus swerves into the lane of the autonomous vehicle. With only seconds to consider the scenario, what should the autonomous vehicle do? Or rather, what should the programmer design the autonomous vehicle to do in this scenario?

There are essentially three actions for the autonomous vehicle to take in the "AV Bridge Dilemma." First, it could hit the bus head-on, thereby sacrificing its one passenger and possibly the three bus passengers to save the five pedestrian lives. Second, it could swerve into the five pedestrians, saving its passenger's life and the lives of the three bus passengers by sacrificing five innocent lives. Third, it could swerve off the bridge, saving the most lives by surely sacrificing its passenger's life.

While there are only three actions, there are a myriad of ethical theories to justify any of those actions. The four most appealing frameworks would be utilitarianism, deontology, social contract tradition, and the egoism. I will briefly consider each of these ethical frameworks in an effort to provide background information and an understanding of how their practitioners would react to the "AV Bridge Dilemma." After those overviews, I will consider some theoretical and practical problems with each theory. It is important to consider the practical problems with each theory; due to the massive number of human lives that would be saved through the mainstream adoption of autonomous vehicles, it is clear that autonomous vehicles should be implemented as quickly and popularly as possible. Therefore, it is important to consider how each ethical framework will affect the ease of mainstream adoption of autonomous vehicles. Finally, I will conclude that the frameworks are ethically and practically wanting, leading to the next section, which will discuss a universal ethics setting versus a pluralist approach.

In terms of programming ease, two ethical theories stand out: utilitarianism and deontology. Both of these theories would be the easiest to program into autonomous vehicles given their simple principles and applicability over a variety of situations. Because it would seem that the programmer might reach for these theories for the sake of consistency and ease, let's first consider utilitarianism.

Classic utilitarianism, originally conceptualized by Jeremy Bentham and John Stuart Mill, considers the consequences of an action, and how those consequences either promote universal happiness or universal pain (Stanford Encyclopedia of Philosophy, "The History of Utilitarianism"). In utilitarianism, the philosopher can use a sort of "moral calculus" in order to determine if the action will cause more net happiness or more net unhappiness universally. In terms of the trolley problem, the utilitarian would sacrifice the one for the five as that would maximize the number of lives saved, and therefore maximize net happiness. Another point to consider is this theory is strictly objective, since it essentially tasks philosophers with forgetting any personal connection to the situation, including who is involved. For example, a utilitarian should not care if the one worker on the alternate trolley track is her sister any more than if the worker were a stranger.

When applying utilitarianism to the "AV Bridge Dilemma", the utilitarian programmer would be concerned first and foremost with minimizing death. This seems simple at first; the autonomous vehicle should sacrifice its one passenger by swerving off the bridge to save the three bus passengers and five pedestrians, since it does not owe any more to its owner than it does to the other strangers. However, there are several considerations that the autonomous vehicle must make in its moral calculus. Consider the possibility of safety measures. Perhaps the bus has better safety measures than the autonomous vehicle, so the autonomous vehicle's sensors could reasonably assume that crashing into the bus would not kill the bus passengers. Then, autonomous vehicle should crash into the bus as safely for its passenger as possible in the hope that doing so would prevent any deaths.

However, Goodall brings up a problem with the consideration of safety measures:

> If the collision was severe and injury likely, the automated vehicle would choose to collide with the vehicle with the higher safety rating, or choose to collide with a helmeted motorcyclist instead of a helmetless rider. Many would consider this decision unfair, not only because of discrimination but also because those who paid for safety were targeted while those who did not were spared. The fact that such decisions were made by a computer algorithm would be no consolation.
>
> (Goodall, 2014)

If the autonomous vehicle collides with the option it deems as "safer," then those who are trying to be safe would experience more crashes, and potentially harm, than those who are not as concerned with safety. This definitely is not fair. In the "AV Bridge Dilemma", what person in her right mind would choose to ride in the bus, knowing that because it is a safer option in terms of preventing fatalities, it has a higher chance of being plowed into by an autonomous vehicle trying to save another person's life? What incentive would there be to choose the safer option, and why should people seeking safety be unfairly targeted by autonomous vehicles?

Another problem with the utilitarian autonomous vehicle is a demographics concern. Let's presume that an autonomous vehicle has a sophisticated facial recognition system that can detect age, gender, and weight. In the autonomous vehicle's moral calculus, it would have to consider that women are 28% more likely to die in the same crash as a man (Evans, 2008). Should an autonomous vehicle do its best to avoid collisions involving women by targeting men? What about vehicles with child passengers? Heavier cars, such as those with more passengers, are less likely to produce fatalities than lighter cars with a single passenger. Does this mean that the autonomous vehicle should actually target a car with more passengers? There are simply too many demographics factors for the autonomous vehicle to consider, and none of those factors guarantee a minimization of lives lost. This resembles a common complaint of utilitarianism; how can one mortal person consider every consequence of an action?

The utilitarian might respond with yes, it is almost impossible for one mortal person to consider every consequence of an action while a situation is occurring. However, an autonomous vehicle is not a mortal person, nor is its programmer attempting to consider every consequence in the moment. The programmer could consider many more consequences over the months, or even years, of programming than one person could consider in a split second. Then, the computer in the autonomous vehicle could be programmed to make the split-second decision with the most information and considerations as possible, since computers can process complex decisions much faster than humans.

Even considering the longer period of time for consideration and the speed at which the autonomous vehicle could process the complex decision, it would still be impossible for the programmer and the autonomous car to consider every possible scenario; the possibilities are infinite. Because of the infinite possibilities, the programmer must idealize an eliminate unlikely scenarios from her consideration. Thus, the programming would encompass all the most likely or moderately likely scenarios while eliminating unlikely scenarios for the sake of simplicity and usability. In the future, perhaps the programmer would consider artificial intelligence learning programs which would allow the autonomous vehicle to learn what to do for unprogrammed scenarios based on what it does for its programmed scenarios. However, Goodall brings up a problem with learning programs:

> If not carefully designed, they risk emulation of how humans behave rather than what they believe. For example, a human may choose to push a nearby vehicle into oncoming traffic to avoid his own collision. Self-preservation instincts that do not maximize overall safety may be realistic but not ethical. Ethics addresses how humans ought or want to behave, rather than how they actually behave, and artificial intelligence techniques should capture ideal behavior.
>
> (Goodall, 2014)

If the autonomous vehicle is not carefully designed and begins driving how humans drive instead of how they wish they drive, the programmer risks making the entire point of the autonomous vehicle, to eliminate human error and create a safer traffic environment, obsolete. However, this concern might does not outweigh the utilitarian benefits outright; instead, the demographics concern just raises questions about what we should value and whether the benefits outweigh the consequences from a personal standpoint.

Finally, there is a practical problem with the utilitarian autonomous vehicle: how will the car company get anyone to buy it? First, forget all that was just discussed and assume that a utilitarian car will perform the simplest calculation to minimize deaths. In the "AV Bridge Dilemma", this would mean that the utilitarian car would sacrifice its passenger in order to save the three bus passengers and five pedestrians. From a public opinion perspective, people support this idea when others are the ones purchasing the autonomous vehicle: "76% of participants thought that it would be more moral for AVs to sacrifice one passenger rather than kill 10 pedestrians" (Bonnefon, Shariff & Rahwan, 2016). However, when asked what type of autonomous vehicle they would want to drive, people overwhelmingly stated they would want an autonomous vehicle that prioritizes the safety of its passenger. As Bonnefon et al. write, "In other words, even though participants still agreed that utilitarian AVs were the most moral, they preferred the self-protective model for themselves" (Bonnefon et al., 2016). Given the public's contradictory view of utilitarian autonomous vehicles, autonomous vehicle manufacturers have a problem; the public at large will be angry if the autonomous vehicle is programmed with anything less than full utilitarianism, but no one will purchase a fully utilitarian autonomous vehicle. If no one will buy the utilitarian autonomous vehicle, there is no reason for the car manufacturer to continue to produce it in terms of a cost-benefit analysis. This paradox with

utilitarian autonomous vehicles could prevent autonomous vehicles from ever becoming mainstream.

The other ethical theory that would be as easy to program as utilitarianism would be Kantian deontology. Made famous by Immanuel Kant, deontology essentially posits that moral people must operate based on a set of pre-determined rules, meaning that no matter the consequences, some actions are always morally impermissible. For example, the classic deontological norm that is discussed in relation to the Trolley Problem states that one should never knowingly do harm to others. In the terms of the Trolley Problem, the Kantian deontologist following this norm would not change the track, as doing so would be to knowingly do harm to the one person. For the deontologist, it is more morally acceptable to maintain one's norms in any situation than to minimize harm. This idea rests on the concept of doing vs. allowing harm. This means that for the Kantian deontologist, there is a moral difference between actually doing harm rather than just allowing harm to happen. The Stanford Encyclopedia of Philosophy offers this reasoning to justify that doing harm is worse than allowing harm to happen:

> It seems true by definition (almost) that you can be causally responsible only for upshots that you cause. And it is arguably true that you can be morally responsible only for what you are causally responsible for. So, if you cause a bad state of affairs, you've probably done wrong; whereas if you don't cause a bad state of affairs, you haven't. In choosing between killing and letting die, you are choosing between doing wrong and not doing wrong.
>
> (Stanford Encyclopedia of Philosophy, "Doing vs. Allowing Harm")

Thus, the Kantian deontologist would expand on the doing vs. allowing distinction by saying that doing harm is morally worse than allowing harm, so allowing harm to come to the five workers is morally permissible while doing harm to the one worker is morally impermissible.

When applying deontology to the "AV Bridge Dilemma," it is easy to see what the deontologist would program the autonomous vehicle to do. The autonomous vehicle would be programmed to do nothing, causing it to crash into the bus, killing its passenger and potentially the three bus passengers. This is because allowing harm to occur is more morally permissible than doing any action that would kill any amount of people, including sacrificing the one passenger to save the other eight strangers.

There are, of course, criticisms of deontology and the norm of never knowingly doing harm to others. The first criticism would be that the doing vs. allowing principle is inherently flawed. This is because "allowing" harm to occur could be understood as an inaction instead of no action at all. Inaction can be defined as "having a reasonable choice, and choosing to do nothing." This is slightly different from doing no action at all, which applied to the Trolley Problem would mean that the bystander had no ability to change the track of the trolley. Simply, the difference between inaction and no action at all is the element of choice. One could argue that an inaction from which harm occurs is actually a form of action, since the decision-maker must still decide to do nothing, allowing the action to occur. By making the conscious decision to not act, the decision-maker is still making a moral choice that could be considered an action. Therefore, the conclusion would be that allowing harm to come to someone still violates the norm of never knowingly doing harm to others.

Another problem with doing vs. allowing in the "AV Bridge Dilemma" is that the understanding of doing vs. allowing does not really work when a programmer is pre-deciding the autonomous vehicle's actions. A human driver might be justified in claiming that she was "allowing" harm to come to the pedestrians because everything was happening in that moment, and because in the American legal system it is far worse to knowingly harm someone (murder)

than to accidentally or unavoidably harm someone (manslaughter). In the case of a human driver, it would appear that allowing harm is less offensive than doing harm. However, the programmer does not have the option of allowing harm; all of her actions would be doing harm. Since the programmer is not acting at all within the moment that the Dilemma is occurring and is always choosing what the autonomous vehicle should do before it enters a situation like the "AV Bridge Dilemma", the programmer is never really "allowing" harm to come to the pedestrians; her predetermination means she is always doing or making harm come to the pedestrians. Then, the distinction between doing and allowing does not apply to the programmer, and the programmer is far less morally justified in programming the autonomous vehicle to stay on course and kill the pedestrians.

Finally, there is the utilitarian's criticism with deontology and the norm of never knowingly doing harm to others. The utilitarian would say that the deontologist is acting contrary to rationality and even selfishly when protecting her personal moral integrity over saving the lives of others. This comes from the distinction that deontologists draw between what is "right" and what is "good." Deontologists argue that the conduct of people, i.e. following their moral rules is "right," while the "good" would be a utilitarian understanding of minimizing harm and maximizing happiness. Utilitarians draw no such lines between the "right" and the "good", simply arguing that whatever is for the greater "good" must be the morally "right" decision. "While consequentialism judges acts according to whether they bring about the best state of affairs, deontology judges acts according to whether the actors conduct themselves in ways that maintain their moral integrity" (Zamir & Medina, 2010). However, by focusing on the "right" instead of the common "good", deontologists risk appearing as if they do not care for other people but only wish to protect their own moral integrity. This is easy to see in the Trolley

Problem; the deontologist lets five people die to protect some abstract idea about their own morality. This also appears to be contrary to rationality, since most people when presented with the simplest form of the Trolley Problem decide that they must save the five people.

In terms of the "AV Bridge Dilemma", deontology also has its problems. The deontologist would choose to not pull the lever to protect her moral integrity in the classic Trolley Problem, but whose moral integrity should be considered in the "AV Bridge Dilemma"? The autonomous vehicle itself has no semblance of moral integrity, since it is not capable of independent thought. It must be programmed in advance by its programmer, who chooses an ethical framework for the autonomous vehicle. Is it the programmer's moral integrity that is considered then? It would seem so, since there is no driver or bystander in control of the autonomous vehicle in the "AV Bridge Dilemma." But the programmer is so far removed from the situation, likely being physically and temporally distant, since it is likely that she programmed the autonomous vehicle's computer far before it was even sent to manufacturing, let alone sold to the consumer. If the programmer is still following the norm of "never knowingly do harm to others," can it even be considered knowingly doing harm when the programmer is that far removed from the actual situation? I assume that the deontologist's answer would be that the programmer knows that it is impossible for an autonomous vehicle to never experience a crash; autonomous vehicles are predicted to reduce traffic accidents by 90%, but it could never reduce accidents by 100% (Nyholm, 2018). The programmer can also assume that some of those traffic accidents will involve unavoidable death. Therefore, the programmer recognizes that some harm must occur, bringing the norm of never knowingly doing harm into the fold. Then the programmer can program the autonomous vehicle to do nothing in the "AV Bridge Dilemma,"

protecting her moral integrity by knowingly allowing, not causing, some harm to inevitably happen, even if she does not know when, where, or how.

Another ethical theory to consider in the "AV Bridge Dilemma" is social contract tradition. This theory can be traced back to Plato, and includes philosophers such as Hobbes, Rousseau, and Scanlon. Social contract tradition involves a group of people mutually agreeing upon certain moral rules for the group as a whole. What might be considered a legitimate agreement varies within the tradition. For example, one might believe that a binding, legitimate agreement must involve unanimous consent, informed consent, and non-forced consent. However, another less strict practitioner might find a simple majority vote to be enough justification. Under social contract tradition, a decision is ethically justified if the group has mutually decided to see the action as morally correct. Thus, right vs. wrong is not an inherent moral concept but is instead constructed by the group that will follow its moral rules. In this way, practitioners of social contract tradition avoid any strict notions of the "right" or the "good" by allowing the group to construct its own ethical framework. This also justifies any punishment of those who do not abide by the rules, since the rule-breaker must have agreed with the rule during its establishment for it to have passed unanimously. In contractualism, one can see a kind of pluralist understanding of ethics; social contract tradition does not dictate any one right way to act, but instead dictates one correct, moral way to decide what is the right way to act. In this way, social contract theory is less about the action itself and more about how the rule for the action is mutually decided upon.

For the Trolley Problem, this means that purely social contract tradition does not offer the right way to act. In fact, the practitioner of social contract tradition probably would not care if the person pulls the lever or not; she would only care about whether the action the decision-

maker takes aligns with the social contract to which the decision-maker agreed. This means that social contract tradition is the most flexible with which action to take. The same would go for the "AV Bridge Dilemma."

However, the problem with the social contract in general is that it works best in smaller and more homogenous groups. How could rules be effectively agreed upon over a larger area, such as the United States as a whole? Bonnefon, Shariff, & Rahwan have already proven this to be difficult. Even when considering the most ethical position and following reasonable rejectability, here was never a 100% consensus on which ethical theory should be programmed into autonomous vehicles (Bonnefon, Shariff, & Rahwan, 2016). If the group cannot unanimously agree, or at least not reject, a set of rules, then there would be more moral justification for any programming and autonomous vehicles would likely be outlawed.

One solution might be loosening the requirements from 100% consent to each rule to a democratic majority vote. Since unanimous consent is basically impossible over such a large and diverse group of people, and the United States has a political and cultural appreciation of the idea of the fair and democratic vote, this seems to make the most sense. The vote would be considered an entry into the social contract with the knowledge and consent to abiding by whatever ethical framework for autonomous vehicles gains the majority of votes. Even in the case of a 51/49 split in the votes nationwide, there would still be a consensus and a justification to prosecute rule-breakers based on the fact that they consented to the decision by consenting to the democratic nature by which the nation came to that decision.

However, a typical complaint towards the social contract theory and a democratic vote is that it is inherently selfish and does not really address the common good. In a typical social contract, a person usually only considers what conditions would be optimal for her own

situation; if the person is more privileged than the others who would be involved in the social contract, then the person might not consider what would be best for the entire community since it might negatively affect them personally. For example, if a person knows that she will eventually be the passenger in the autonomous vehicle in the case of the "AV Bridge Dilemma," then she will most likely argue for a protectionist autonomous vehicle. This is demonstrated in the public opinion survey by Bonnefon et al (2016). However, this is where the Veil of Ignorance would come into play. The Veil of Ignorance is essentially a state in which "no one knows his or her place in society, class position or social status, nor does any one know his or her race or gender, fortune in the distribution of natural assets and abilities, level of intelligence, strength, education, and the like." (Freeman, 2019). Operating under the Veil of Ignorance means that people are unable to act in their own self-interest, because they simply have no idea what their personal characteristics are and therefore cannot identify what would be in their self-interest.

Using the Veil of Ignorance in relation to a social contract on autonomous vehicles would help the deliberation, especially because of the public opinion paradox outlined earlier in the paper. If people do not know if they will be the passenger or pedestrian in the "AV Bridge Dilemma," then they would be more likely to reach a fair settlement that benefits everyone. This condition would also be relatively simple to implement during this time before autonomous vehicles are a mainstream driving option. While most people consider themselves the pedestrians as of now, they recognize that autonomous vehicles will most likely be widely driven over the course of the next ten years. This creates a sort of Veil of Ignorance; no one really knows what position they will be in the "AV Bridge Dilemma" in ten years, since there is only a theoretical understanding of how widespread this technology will be in ten years. Because of this, the most

fair social contract about autonomous vehicles would be decided upon in the next five years or so, so that people can make a decision without acting in their own self-interest.

Another solution would be to allow states to decide their regulations for the ethical frameworks of autonomous vehicles in that state. That would allow for a smaller group to decide upon its rules and would probably lead to faster unanimous consent. But there is a problem with this solution as well. Autonomous vehicles, and vehicles in general, are made to go places. If different states have different rules for autonomous vehicles, would this prevent autonomous vehicles from crossing state lines? If the autonomous vehicle has a "moral dial" that allows the passenger to change the moral framework based on the state that it is in, is the passenger really a part of the community that decided upon different rules? Perhaps we can conceptualize a hierarchy of rules, resembling something like this:

- National rule: states can decide upon the ethical framework of their autonomous vehicles, and all visitors agree to change their autonomous vehicles to obey the rules of the state in which they currently are. This should be unanimously decided.
    - State rule: democratically decided upon by the residents of that state through a majority vote

If the national rule can be unanimously agreed upon, visitors can be considered a part of the larger decision-making process and would be obligated to follow the rules of the state in which they are in the same way that residents are obligated to follow the rules to which they consented.

One practical benefit of this theory as related to the "AV Bridge Dilemma" is that it would be easy to program and justify once the rules are mutually decided upon. For example, if the group decides that a basic rule of their society is to minimize harm, the autonomous vehicle

could be programmed in a utilitarian way. Or, if the group decides that a basic rule should be to protect its children at all costs, the autonomous vehicle could be programmed to use facial recognition and do everything in its power to not harm a child, even if that is at the expense of several adult lives. This would make responsibility quite simple too; the autonomous vehicle must be programmed to follow the group's rules, so the group at large is responsible for whatever happens in the "AV Bridge Dilemma." In that way, no one is at fault because the rules were unanimously agreed upon. When the group becomes responsible instead of any one individual, it is unnecessary and in fact irrational to assign blame to any individual following the rules. A social contract would also make legal responsibility, specifically lawsuits, simple. The only justifiable reason to bring a lawsuit would be if the programmer, for some reason, purposefully programmed the autonomous vehicle to defy the rules; she would of course be found guilty and punished, since she consented to the rules or the rules-making process during their formation.

The final ethical approach to consider is egoism. Moral egoism essentially states that one morally ought to perform an action that supports one's self-interest (Shaver, 2019). In short, the egoist prioritizes her good over the good of all others and is not required to even consider other people when making a decision unless considering them would be better for her in the long run. In terms of the "AV Bridge Dilemma", the egoist would clearly argue for the autonomous vehicle being programmed to protect its passenger(s). This would mean sacrificing the pedestrians in order to save the passenger's single life. It is clear that the egoist would argue for a protectionist autonomous vehicle because of the view of other parties in egoism; in this line of reasoning, the programmer owes nothing to the five pedestrians because she must prioritize the lives of the car's passengers above all else.

This egoist standpoint would also be a good decision practically. As discussed before in the section on utilitarianism, people generally want other people to drive utilitarian autonomous vehicles but would greatly prefer to ride in protectionist vehicles themselves (Bonnefon et al., 2016). If programmers programmed their autonomous vehicles following the egoism, where the autonomous vehicles were programmed to protect their passengers first, it is likely more consumers would purchase the product. This is obviously a good decision for the manufacturer; more customers mean more money for the company. It is also a good decision from the practical standpoint of trying to reduce traffic deaths. When everyone uses an autonomous vehicle, then an estimated 90% of traffic deaths a year will be prevented by this use (Bonnefon et al. 2016). If more people make the switch over to autonomous vehicles quicker given a protectionist vehicle over a utilitarian vehicle, than more lives will be saved in the long run. One could consider this a "sneaky utilitarian" line of thinking.

One potential benefit of egoist autonomous vehicles would be that they could possibly be safer. Consider a situation in which two autonomous vehicles are headed towards collision and possible harm. Both of the autonomous vehicles in this case are programmed to protect the lives of their passengers, following ethical egoism. Theoretically, if both autonomous vehicles are attempting to save the lives of their passengers above all else, then they could potentially avoid a collision and save the lives in both of the vehicles. This might differ from two autonomous vehicles attempting to protect the lives of the other passengers, as the situation might become more complicated and lead to a collision anyway. However, this benefit is highly theoretical, and has been doubted by some game theorists. For example, consider Richard Tay's conclusions in his study on whether it is safer for people to operate larger perceivably "safer" cars such as SUVs or smaller "dangerous" cars such as Smart Cars. In the world of regular vehicles, people

tend to believe that driving a larger car will be safer for them, thereby prioritizing the protection of the vehicle's passengers. However, Tay showed that if both drivers prioritized the lives of the other vehicle and chose to drive a small car, the chances of a fatality in both of the small cars is much lower than the chance of fatality between a large and a small car, or even two large cars (Tay, 2000). The lowering of fatalities in this case requires all parties to prioritize the lives of the others more than their own lives. While this obviously does not deal with autonomous vehicles directly, one could speculate that the same principle might apply to ethical programming in these types of driverless vehicles. Once again, both of these considerations are highly theoretical, and so the benefits or consequences as related to egoism should be considered only possible.

One flaw with egoism is obvious; it seems incredibly selfish. To elaborate, it would seem that ethical egoists have no duty to other people in the same ways that utilitarians, deontologists, and social contract theorists have towards the others involved in the "AV Bridge Dilemma" and other ethical concerns. For example, egoism does not require its practitioners to perform any action in which there is no benefit for the self, no matter how small the sacrifice or how large the benefit for others; if a person were drowning, the egoist would not be required to save the drowning person even if the consequence was as small as getting her clothes wet (Shaver, 2019). Of course, this seems almost unethical. However, the egoist would have a response to the claim that egoism is selfish. The egoist would reply by saying there is a real self-interest involved in considering the feelings and humanity of others. This argument can be explained as follows: "Each person needs the cooperation of others to obtain goods such as defense or friendship. If I act as if I give no weight to others, others will not cooperate with me. If, say, I break my promises whenever it is in my direct self-interest to do so, others will not accept my promises, and may even attack me. I do best, then, by acting as if others have weight (provided they act as

if I have weight in return)" (Shaver, 2019, pp. 20). In short, the egoist has an interest in thinking about the good of others because if she fails to do this, she will fail to receive certain benefits that are vital to her own self-interest. Although the ethical egoist might have a reason to consider the good of others, egoism can still be seen as selfish since there is no requirement by egoism to do so and the incentive for this is still to serve one's self-interest. It is up to each individual to consider whether the motive of the action affects the action itself.

Another problem with egoism is that it does not make a lot of sense in the case of the "AV Bridge Dilemma" where a third-party programmer is making the decision of what the autonomous vehicle will do. Ethical egoists will do whatever action is in their own self-interest, while considering the interest of others to a far lesser degree, if at all. In fact, as demonstrated in the point earlier, ethical egoists only consider the interest of others because it can further their own self-interest. In the "AV Bridge Dilemma," the programmer who is making the decision is not even a factor in the actual problem; she makes the decision beforehand knowing that she will likely never be in that autonomous vehicle in a situation of unavoidable harm. So how could the egoist programmer justify the protectionist autonomous vehicle when she herself has no direct self-interest in the case of the "AV Bridge Dilemma"? One answer might be that the programmer is acting in her own self-interest because she will probably be one of the first people to drive an autonomous vehicle. Since egoism here would be the universal ethics setting, she is acting in her own self-interest to program the autonomous vehicle in a protectionist manner, and it is a secondary result that all autonomous vehicles are programmed this way. Another answer might relate to the marketability of the vehicles; the programmer recognizes that people are more willing to buy the protectionist vehicle, so she programs the vehicles in such a way that she will maintain her job. In any case, the relationship between egoism and the "AV Bridge Dilemma"

becomes rather contrived when considering the programmer has no direct stake in a specific case of unavoidable harm.

Should Autonomous Vehicles Have Universal Ethics Programming?

After discussing the four most appealing ethical frameworks in the first section, along with their theoretical and practical problems, it is easy to see that there is no perfect ethical framework for handling autonomous cars; in fact, there is no perfect ethical framework in general. This can sour the overarching framework of universal ethics programming; without a perfect ethical theory, how can one justify one specific theory as being universal and applicable to everyone? If one rejects universal ethics programming, this forces the consideration of user-selected programming.

To justify user-selected programming, there are two moral frameworks that must be considered. The first framework is a minimal deontological theory referred to as moral libertarianism. Moral libertarians state that there is only an ethical obligation to do no harm to others, but there is no ethical obligation to help others. This means that outside of knowingly doing harm to others, moral libertarians are free to act in whatever way they choose, justified by the fact that those actions do not knowingly do harm to anyone else. Using a framework of moral libertarianism, any action the autonomous vehicle does in a situation of unavoidable harm is justified, because the autonomous vehicle has already done everything in its power to prevent harm and now must act outside of deontology.

The other moral framework to consider in this case is pluralism. The justification for user-selected programming is very similar to the justification for pluralism; the philosopher has explored the other moral frameworks and found all of them wanting, so there must be a problem

interpersonally (disagreement between different people) or intrapersonally (conflict between the philsopher's own decision-making). In the case of user-selected programming, it would appear that the pluralist intention takes form as theory indeterminism, in which all moral theories are equally true. However, to avoid the contradiction involved in affirming two distinctly opposing frameworks, it is more helpful to adopt pragmatism and claim that all moral theories are useful in different contexts instead of that they are all true for all contexts.

Another way to think about pluralism in the case of autonomous vehicle programming is to consider value pluralism. Value pluralism recognizes the essential values in all of the moral theories and allows those values to conflict by claiming they are not direct conflicts (Mason, 2018). For example, a person might place value on both utilitarian justice and deontological beneficence; those two might conflict in a situation, but through value pluralism, it is not as serious of a conflict as in theory indeterminism. After distilling values in this way, the pluralist can weigh each value against the others until one ethical decision proves itself better (Mason, 2018). In cases where there is an equal weight, the pluralist should let other contextual considerations break the tie; this could come in the form of self-interest or perhaps societal pressures through the expectation of those watching the person making the decision.

The Ethical Knob

Because of the theoretical and practical shortcomings of a single ethical programming for autonomous vehicles, autonomous vehicles should be programmed in a more ethically pluralist way; this would be physically manifested as an "ethical dial" of sorts. The dial would allow the passengers and/or owners of the autonomous vehicle to decide how the autonomous vehicle should react in cases of unavoidable death. Essentially, the passengers and/or owners could decide to program the car in the following modes:

- Altruistic mode: preference for third parties
- Impartial/utilitarian mode: equal weight given to the passengers and third parties
- Protectionist/egoist/partialist mode: preference for passengers

(Contissa, Lagioia, and Sartor, 2017).

After listing those three settings for the autonomous vehicle, it is easy to see which traditional ethical theory would line up with each setting. However, one might be wondering, why is there no deontological setting of "do no harm"? This is quite simple once it is pointed out; the autonomous vehicle is primarily operating under a deontological setting of "do no harm," up until it encounters a situation of unavoidable harm. The ethical knob only comes into play in situations of unavoidable harm because it would be horribly unethical for the autonomous vehicle to be programmed in any other way than deontological "do no harm" in cases where harm is avoidable. Consider that the ethical knob is programmed in an egoist way. This would not mean that the autonomous vehicle would mow down any pedestrian in its way just to ensure that the passengers reached their destination quickly. The egoist setting would only come into play if the autonomous vehicle was forced to make a decision in which harm would come to somebody. Another reason why there is no deontological setting on the ethical knob is because the deontological norm of "do no harm" breaks down when harm is unavoidable. The clause of doing vs. allowing also does not apply here, because the passenger is essentially always "doing" harm when she pre-selects the setting of the ethical knob. With those considerations, the paper can move on to the actual discussion of the ethical knob itself.

One strong benefit of the ethical knob approach to autonomous vehicles is that responsibility could be easily assigned, much like how responsibility is assigned in cases of human drivers. In the examples listed in the previous section, responsibility in the "AV Bridge

Dilemma" was always assigned to the programmer. This is because the programmer would ultimately decide and program the autonomous vehicle with the ethical framework. However, this might not necessarily be fair. Perhaps the programmer was pressured by the company she works for to program the autonomous vehicle in such a way that more customers would buy it. This example also illuminates the financial barriers associated with this kind of responsibility; if the company assumes responsibility for the ethical programming of the autonomous vehicle, then it must budget for lawsuits, which will in turn increase the price for autonomous vehicles (Gogoll and Müller, 2016). In the case of the ethical knob, the passengers, or setters of the knob, would clearly be responsible for the outcomes of the "AV Bridge Dilemma" because the passenger is the one who decided what the autonomous vehicle would do. As mentioned before, because the passengers set the knob in advance, the doing vs. allowing clause would not come into play. This would make litigation around car accidents such as these very simple, since the setter of the knob would always be somewhat at fault.

Moral responsibility would also be easy to assign. Give a morally libertarian framework, the person exercising their free will in setting the ethical knob can then be praised or blamed since she had a completely free choice between her duty to others and her strong desire to protect her own life at all costs. Once moral responsibility can be assigned in this case, people can generalize and establish a moral principle for cases of unavoidable harm, creating a stronger desire to act based on duty.

Since the responsibility of a crash will still be assigned to the passengers or setters of the knob, there could be some practical incentives to attempt to balance a person's natural tendency towards self-interest and the egoist knob setting. For instance, consider insurance. Since the passengers will most likely be deemed responsible in an "AV Bridge Dilemma" in which they

pre-set the ethical knob to prioritize their lives over the pedestrians, thereby causing the deaths of innocents, their insurance premiums will most likely be much higher than passengers who use an altruistic or impartial setting. This monetary incentivization might persuade people to not select the highest protectionist mode, or even to select the impartial or altruistic mode for a reduced rate. However, there is an ethical dilemma within this scenario as well; would this allow wealthy people to always select the egoist setting while less wealthy people must select impartial or altruist due to financial necessity, and will this in turn continue to prioritize the lives of the wealthy over the lives of the poor? This incentivization through insurance might work in the very short-term in which only the wealthy own autonomous vehicles, but it would need further review once autonomous vehicles are more available to people in lower wealth brackets.

Another Option

If the ethical knob seems suspect in some way, then perhaps there needs to be another option. One option is to forget all notions of ethical programming in an autonomous vehicle in cases of unavoidable harm. This would mean removing the facial recognition or counting ability in autonomous vehicles so that they would not be able to consider how many lives would be saved or lost in its actions. Then the vehicle would be programmed with accident avoidance in those situations and would operate much like a human driver, with no consideration of numbers.

This could be considered more ethical when thinking about paternalism and technology. If technology is treated as a godlike figure, then the technology should strive to be as moral and perfect as possible. However, if the technology acts like humans act, then could it reasonably be expected to be any more moral or perfect than we expect humans to be? It is clear that people are not usually held up to this level in cases of instant decision-making; this is typically the

difference between a murder charge and a manslaughter charge. Perhaps the lack of programming for these cases would also exempt autonomous vehicles.

However, this is a disservice to the technology and the people who will benefit from it. Human error, as discussed before, is what causes the vast majority of traffic-related accidents and deaths. Programming an autonomous vehicle with only human capabilities in cases of unavoidable harm will not reduce traffic deaths to the extent that the technology could. If autonomous vehicles can be programmed with technology that might make them safer, then they should be programmed in that way regardless of more theoretical ethical concerns.

A Caveat

The considerations brought forth in this paper are only important in the short-term of autonomous vehicle existence. This is because the considerations only consider autonomous vehicles under a personal ownership model, in which people still primarily buy and own personal vehicles. While there are some transportation options that subvert this, such as public transportation like buses and rideshare options like Uber, automobiles are still primarily purchased for individuals or family units to use personally.

When autonomous vehicles become mainstream to the point where regular automobiles are no longer on the road, this model of personal ownership might change. For example, given the increasing rate of climate change and the increasing population across the world, it might no longer make sense for individuals to own vehicles when an environmental conscious is necessary and there are possible resource shortages. For those reasons, the United States might move towards a full collective rideshare model in which people request transportation from companies like Uber for all of their transportation needs. If this becomes the case, there will be different

considerations of whether all vehicles from the companies are programmed universally or if the passengers can select their preferred programming upon request.

Works Cited

Administration, U. D. (2017, October). 2016 Fatal Motor Vehicle Crashes: Overview.

Brown, B. (2017, October 6). Evidence stacks up in favor of self-driving cars in 2016 NHTSA fatality report.

Contissa, G., Lagioia, F. & Sartor, G. Artif Intell Law (2017) 25: 365. https://doi.org/10.1007/s10506-017-9211-z

Edelstein, L. D. (2019, February 3). Sit back, relax, and enjoy a ride through the history of self-driving cars.

Gogoll, J., & Müller, J.,F. (2017). Autonomous cars: In favor of a mandatory ethics setting. Science and Engineering Ethics, 23(3), 681-700. doi:http://dx.doi.org.libproxy.uwyo.edu/10.1007/s11948-016-9806-x

Goodall, N. (2014). Ethical Decision Making During Automated Vehicle Crashes. *Transportation Research Record: Journal of the Transportation Research Board, 2424*, 58-65.

Jean-François Bonnefon, A. S. (2016). The social dilemma of autonomous vehicles. *Science Magazine*, 1573-1576.

Maclagan, W. (1951). Symposium: Freedom of the Will. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *25*, 161-216.

Mason, Elinor. "Value Pluralism", The Stanford Encyclopedia of Philosophy (Spring 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2018/entries/value-pluralism/>.

Nyholm, S. (2018). Ethics of Crashes with Self-Driving Cars: A Roadmap I. *Philosophy Compass*, 13:e12507.

Rehman, S., & Dzionek-Kozłowska, J. (2018). The trolley problem revisited. an exploratory study. Annales.Ethics in Economic Life, 21(3), 23. doi:http://dx.doi.org.libproxy.uwyo.edu/10.18778/1899-2226.21.3.02

Romero, S. (2018, December 31). Wielding Rocks and Knives, Arizonans Attack Self-Driving Cars. *New York Times*.

Shaver, Robert, "Egoism", The Stanford Encyclopedia of Philosophy (Spring 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2019/entries/egoism/>.

Tay, Richard. (2019). The prisoners' dilemma: A game theoretic approach to vehicle safety.

Wakabayashi, D. (2018, March 19). Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam. *New York Times*.

Zamir, E., & Medina, B. (2010-01-21). Threshold Deontology and Its Critique. In (Ed.), Law, Economics, and Morality: Oxford University Press,. Retrieved 8 May. 2019, from https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195372168.001.0001/acprof-9780195372168-chapter-02.