Jaryd Meek
CSCI 3104-100
Ethics Essay

The Self-Driving Car Trolley Problem

It's a sunny spring afternoon, and you decide that you want to go for a ride in your brand-new Tesla Model 3 with the "Full Self Driving" upgrade. Many other people are taking advantage of the nice weather and are walking along the sides of the road. You're driving through town at 45 mph, and you decide that you want to change your music, so you enable "Autopilot" while you change songs on the touchscreen in the car. You happen to be following a flatbed truck carrying k-rails, and right at that moment, that truck hits a pothole, and a k-rail falls off the back of the truck. There's a large median in the middle of the road, leading to only two options: drive straight and hit the k-rail, killing you in your car, or swerve off the road to the right and hit a pedestrian walking, killing them. You don't have time to react, so the car must make the decision for itself. This adaptation of the popular philosophical thought exercise, the Trolley Problem, is used to evaluate different ethical ideologies but is a very real problem that surfaces as self-driving vehicles become more ubiquitous.

The price and reliability of self-driving technologies for vehicles ranging from lane assistance technology to Tesla's "Full Self Driving" (Level 2 Autonomy). Almost any new vehicle sold today will have assistance algorithms of some sort that aim to increase safety for both the driver as well as other drivers on the road. For many people, these are algorithms that will run any time they get in their cars and run the entire length of their drive. Many of these technologies are still in their infancy and act as sources of additional information that the driver may choose to act upon. Still, it is becoming more and more common for the software to act for the driver.

These algorithms take in information from many sensors, including many cameras, LIDAR, RADAR, etc., and can output through displays, speakers, or by performing an action with motors. Generally, these algorithms take in the information from these sensors and process that data to either help a driver avoid a dangerous situation or pilot the vehicle in the case of autonomous vehicles. For the purposes of this paper, we will focus on algorithms that control the car directly, bypassing the driver altogether, as these need to make decisions that have direct implications on their own, rather than passing the burden of decision off to the driver.

Today, the largest car manufacturer working on self-driving is Tesla, but even their algorithm right now wouldn't make and act on a decision in the adapted trolley problem introduced above. Today, your car would beep at you and give control back over to you, and it would be up to you to decide. But as self-driving cars become more trusted and reliable, at some point, they will no longer have steering wheels and brake pedals, meaning that the vehicle will have to make the entire decision itself.

Based on the article, *The ethics of algorithms: Mapping the debate,* This situation creates an ethical situation of inconclusive evidence. This is due to the fact that the algorithm will have to plan what to do based on what is probable rather than what is actual. What if in the situation laid out above, it happens to be an empty plastic k-rail rather than one made of concrete? The vehicle may decide to swerve and kill an innocent person rather than just hit the

plastic, that would do no more damage than to scratch the paint or maybe dent the front of the vehicle. While algorithms will eventually be able to be safer than humans even with this limitation, deliberation is necessary when considering the ethical ramifications of self-driving vehicles.

While these algorithms eventually will operate on a computer in a car, it has become common place to test them with a simulator first. Many companies, including Tesla, have simulators where they can feed their algorithms specific data for certain situations and see how the algorithm would respond. You could feed the algorithm the adapted trolley problem and see how it would react. Beyond seeing how it would respond, we can also go back and see exactly what caused it to make the decision it did. This allows us to see if there are any ethical concerns well before the software gets loaded on to millions of 2-ton death machines on the road. The most considerable ethical problem is what factors cause the car to swerve and kill an innocent person/people or kill the people in the vehicle. There are many outcomes that could be seen as ethically concerning, including seeing the possibility of it favoring one race of people over another. These are all concerns that we can test for in simulations, which can hopefully prevent us from releasing unethical algorithms to the general public on their vehicles.

There are two main ethical views concerning this situation. The Utilitarian perspective, or the Deontological perspective. Utilitarians believe that all actions should be decided based on which decision promotes the most "pleasure." This means that in the adapted trolley problem, the car would count the number of people who would die with either decision and pick the option that leads to the most "pleasure" or the least number of people killed. For example, there is a car full of 7 people, and it could kill everyone in the car or swerve to the right to kill one person; it would decide to swerve. The main opposing view to Utilitarianism regarding the trolley problem is the Deontological perspective. Deontologists believe that there is a strict set of rules about what is ethical. This means that they will pick a strict rule and stick to it. For example, regarding the trolley problem – the person/people in the car should always be the one that has to die, rather than an innocent pedestrian since the car created the situation, or perhaps assign a value to each life (doctors valued more than a retail worker, for example) and save the lives with the most value. Both lead to ethical issues and neither is a perfect decision, but these are the decisions that will need to be made before these cars become ubiquitous.

Either of these ethical theories can create situations where the algorithm has to decide the value of people. Then, either maximize "pleasure" by killing the people who generate less "pleasure" (Utilitarianism) or *always* save the lives with the most value (Deontology). This means that the algorithm can aid in the perpetuation of existing societal harm. If the car cannot precisely determine who someone is and their worth, does it start obeying stereotypes? Does the car favor a white person since historically they have been more likely to be doctors, or influential politicians? This creates many ethical issues and could lead to the exacerbation of societal harm.

I believe that the ethical concerns should be addressed with a deontological view. I believe that until the day where self-driving cars are ubiquitous and no longer have manual control, the vehicles should always choose to favor people outside the vehicle. Algorithms should never favor killing an innocent person instead of someone who is directly involved with getting into the situation. By choosing to have the vehicle drive for itself, it should be

understood that the driver is taking a risk, and as part of that the vehicle should always favor outside life. This raises other concerns, for example, which company would be willing to produce a vehicle that prefers to kill the owner rather than someone else? This leads to a major point. Whatever decision is considered the ethical decision, it needs to be enforced by an above agency, for example, the NHSTA.

Self-driving cars are the future and will become ubiquitous. There are many ethical considerations that need to be made because of this. As such, some national (or international) group needs to set the laws surrounding the ethical theory in order to manage these decisions. These decisions have enormous ramifications and leaving the ethics of these decisions to automakers could lead to the proliferation of existing societal harm, or worse.

Works Cited

Mittelstadt, Brent Daniel, et al. “The Ethics of Algorithms: Mapping the Debate.” Big Data & Society, Dec. 2016.