Cameron Feigenbaum

The Utilitarian approach to Self-Driving Cars

Introduction

As society continues to make technological advances, new philosophical and ethical questions begin to arise as to how we should properly use these technologies. One of the emerging technologies that has been debated has been the advent of self driving cars being controlled by artificial intelligence. The premise behind self driving cars is that they would be a lot less susceptible to accidents compared to that of a vehicle driven by a human. However, even artificial intelligence is not infallible, and inevitably accidents that are unavoidable will occur. The question then becomes, how should engineers program these self driving cars to behave in the event of an unavoidable accident? Some philosophers have attempted to use variations of trolley problems to gather data as to how people naturally feel these self driving cars should prioritize innocent bystanders lives, or the passengers. However, there are clear limitations to using the trolley problems as a truly comparable means of simulating a scenario that would arise, and how a human would react if placed in the same situation. Another question that has arisen in regards to the programming of self driving cars, is who should be held responsible in the event of an accident, because typically the blame would be placed on the driver, but in this case the driver is not a person, but an artificial intelligence programmed by engineers. In order to create unanimity between these self driving cars, it is necessary for governments to establish an agency within their transportation services to standardize rules that engineers must uphold in the programming of self driving cars in order to create consistency, as well as clarify who is to be held accountable in the event of accidents. Self driving car's artificial intelligence should follow

a utilitarian model of always maximizing the statistical lives saved without any discirimination to age, gender or social status, but would require government regulation in order to hold companies and engineers to a standard, allowing for their to be culpability on their part if their technology malfunctioned.

Benefits of Self Driving Cars

There are clear indisputable advantages to introducing self-driving cars to mainstream society. It is already clear that a utilitarian would believe that the artificial intelligence of these cars should be programmed to maximized statistical lives saved, because maximizing statistical lives saved would lead to the greatest outcome in terms of utility to society. However, self-driving car technology does not only benefit society in terms of safety utility. There are other avenues for self-driving cars to enhance the utility within society. For example in Andy Lau's article "The Ethics of Self-Driving Cars," the National Science foundation discovered that self-driving cars would severely reduce the traffic and congestion we have on our streets and highways. By preventing traffic jams, people are allowed to more efficiently allocate their time to productive activities, which can only be seen as a positive for society. Not only will people have more time due to the reduction in traffic, but an individual can even be productive while in their vehicle. Rather than having to drive, the would-be driver can devote themselves to tasks that they would otherwise be unable to do because they were driving. The National Science foundation also discovered that self-driving cars would significantly reduce fuel consumption. If traffic jams are reduced, so too is fuel consumption. It is estimated that self driving cars would reduce fuel consumption of vehicles by up to 44%. This is not only saving people time,

resources, and money, but is also good for the environment. The less fuel and oil are consumed, the less greenhouse gas emissions we have to contend with, the better off our society is.

People tend to be apprehensive about the integration of self-driving car technology into society. Much of the fear stems from a lack of understanding as to how the technology works and how accurate it would be. However, we already have largely automated technologies that have proved to be more efficient than any human piloted technologies. Even within the transportation industry we see the successful application of a similar technology with airplanes. The vast majority of a plane's trajectory is automated. Pilots rarely have to take control of the plane themselves and instead focus on managing the system to make sure they are functioning properly and communicating with flight control. Yet, despite the high level of automation in flight systems, planes have been the safest way to travel for years, with the lowest number of deaths per billion passengers out of all the primary modes of transportation. This is largely due to the automated nature of flight, because it removes most of the potential for human error. Human error is overwhelmingly responsible for the majority of car accidents. Self driving cars would operate similarly to how our planes operate. The cars use a combination of cameras, sensors, and antennas to chart their course or course correct. With the introduction of 5G technology, cars will be able to instantly access databases, allowing them to make split second calculations based on statistical probabilities. Not only will these vehicles be able to instantly pull information from these databases, but they will also be able to communicate with our infrastructure, other vehicles and even pedestrians. The vast majority of people have smartphones, which the vehicles can communicate with to provide more information to allow for self driving cars to more safely navigate. The same will happen with vehicle to vehicle communications, allowing for a level of safety that would be impossible to accomplish with a human driver.

The Problem With Trolley Problems

Inconsistencies across trolley problems make it difficult to apply them in order to try to answer the question as to how we should ethically program the artificial intelligence of self driving cars in the event of an accident. For example, imagine a “pull the switch” variation of a trolley problem, where the “person” pulling the switch is actually an artificial intelligence. A self-driving car driving down the road has a vehicle swerve in front of. The artificial intelligence is now left with two options. The artificial intelligence can decide to continue on its path, which would result in the death of its five passengers, or it could swerve onto the sidewalk and hit two innocent bystanders instead. This is essentially a pull the switch trolley problem in which there are five people tied to track A, and two people tied to track B. The artificial intelligence then has to decide whether to “pull the switch” and kill the two people on track B, or allow the trolley to continue onto track A and kill five people. It is my belief that in this scenario, the maximum number of lives saved should always be the primary goal. It is true that everyone has the right to life, and that right should be violated as minimally as possible, which both utilitarians and natural rights advocates could agree on.Utilitarians will always prioritize saving the maximum number of lives given this scenario, but a natural rights theorist should also be able to recognize that as we all have the right to life, it would be worse to end that right for the lives of five people then it would be to end the lives of two. So using the pull the switch variation of the trolley problem it would appear to be clear that people prefer a utilitarian approach to the programming of self driving cars.

When using the “fatman” variation of the trolley problem however, humans tend to believe that we should not use a utilitarian model, and instead, spare one persons life at the cost of others. Natural rights theorists in particular would believe it to be highly unethical to murder the one person to save others, and in turn, might believe that autonomous cars should not strike innocent bystanders in order to save passengers, as this would be highly unfair to those innocents that become collateral damage and have their right to life violated. However, one of the primary ethical issues with the fat man trolley problem is that people see pushing the fat man onto the train tracks as murder, whereas flipping the switch was not seen as murder. Yet, with a self driving car, it cannot truly be seen as murder for the car to elect to hit one person to spare the lives of four, because there is not a human being who is making a conscious decision to kill another individual. Instead it is an impartial artificial intelligence that is making the decision. Murder requires intent to end someone’s life, but the artificial intelligence has no intent to kill people, and the death should simply be seen as the result of an artificial intelligence making the best out of a bad situation. In this scenario, it cannot be looked at as murdering one person to save another, rather an artificial intelligence is simply saving the maximum number of lives possible, which was seen as typically acceptable in the “pull the switch” variations of trolley problems.

Another issue with using trolley problems to attempt to infer what kind of ethics should be programmed into self driving cars in the event of an unavoidable accident is that in trolley car problems, the outcomes are typically guaranteed, whereas in reality, this would not be the case. In a typical trolley problem, it is guaranteed that if you choose option A, all these people will die, and if you choose option B, all those people will die. However, that would almost never be the case when applied to reality. These artificial intelligences, in deciding what course of action

would be best to take, would be working with various levels of probabilities of lives lost, not guarantees. Therefore it would be the fairest and most consistent approach to program these artificial intelligence to maximize statistical lives saved. People place too much of a value on certainty where certainty rarely truly exists. We find this to be the case when looking at the phenomena of what people would be willing to pay in order to reduce their chance of death. People are willing to pay more for certainty but this is not rational. A person might be willing to pay 4000 dollars to reduce their chance of death from one percent to zero percent, making it certain that they would not die. However, that same person would only be willing to pay 2000 dollars to reduce their chance of death from 4 percent to 2 percent. There is no rational reason to pay more to reduce your risk of death by one percent then you would pay to reduce their risk of two percent, but people put a disproportionate value on certainty. What should matter most is the proportion by which one increases their odds of living, and self driving car technology clearly does that the most. You can claim something to be nearly certain, but anything other than facts can truly be certain or guaranteed. So statistical probability is the only relevant standard.

Government Oversight

In order to properly use this technology without creating social disruption, the ethics programmed into self driving cars artificial intelligence needs to be standardized by governments, and consented to by people in the form of a social contract. Even now, when walking along the sidewalk with human drivers, we are already essentially taking part in a social contract, where bystanders understand that drivers will attempt to do their best to not hit people, and the drivers understand that in the event they do, they will be held morally as well as legally

responsible, except under specific circumstances. Yet humans are far more fallible then a properly programmed artificial intelligence would be. We have a duty to the human race to preserve the maximum number of human lives when possible. The use of self driving cars would eliminate all deaths that have occured due to the result of a human driver being under the influence of alcohol or drugs, or who had a medical episode such as a seizure while they were operating their vehicle. So although innocent bystanders may feel afraid of being "targeted" by a self-driving car inorder to save the life of its passengers, overall, these innocent bystanders as well as the passengers, would be safer. Human beings already tend to prioritize self preservation, so there is no reason for bystanders to think that a human driver would hesitate to hit them if it saved their own life. However, with a artificial intelligence, not only would the number of opportunities when a driver would have to choose between the lives of themself and their passengers or innocent bystanders would be reduced, there may even be times when the car will choose to sacrifice its passengers in order to save a greater number of people. In this way, it is completely impartial as to who is saved, and only the greatest number of lives saved matters. This is also why it would require a sort of social contract among people, as both parties benefit from the self driving car technology, because both parties would see that there were fewer deaths, however, these people must be willing to accept that on the rare occasion of an accident, that they might find themselves to be collateral damage, but that is no different then walking around with human drivers right now. According to the National Highway Traffic Safety Administration, there are more than 10,000 deaths a year caused by humans driving under the influence of alcohol or drugs, with 30 people dying per day in the United States alone. The proper application of self driving car technology would completely abolish this type of death. Not only does it save lives, but it also reduces the cost on society that result from these types of

accidents. According to the National Highway Transportation Safety Administration, in 2010 it cost society 44 billion dollars worth of damages. So not only would more lives be saved, but the cost on society would also be reduced substantially. All parties stand to benefit.

The Ethical Issues With Self Driving Cars

The distinction between killing someone versus letting someone die comes into play in regards as to how we should ethically program our self driving cars. Utilitarians believe that there is typically no moral difference between killing versus letting die. However, natural rights theorists and non consequentialists believe there is a distinction and provides a reason as to why a self driving car should not "target" an innocent bystander, even if it meant saving those lives of several passengers. These same people would elect not to "flip the switch" in a trolley problem, as they would look at it as murder. However this would be a mischaracterization of the situation. The only factor that should matter is the number of lives saved, because it is not a "person" who is making a choice between killing versus letting die. The artificial intelligence of a self driving car should be seen as an impartial judge and not an emotional human being who is rationally making the decision to kill someone or let someone die. Therefore, artificial intelligence is simply making what is objectively the best decision in regards to saving the maximum number of statistical lives. If all lives are valued to be equal, there is no objective reason not to prioritize maximizing the number of lives saved. The killing versus letting die issue is irrelevant to the self driving car because it is not a human being who is in control of the situation. Self driving cars are not even truly making a "decision" when they are determining what course of action would be best, but rather their "decision" has already been preordained by the artificial intelligence

algorithm. So they are not “choosing” to kill someone, or let someone die, but are simply executing an algorithm. So the only focus should be on how many lives were saved.

Another ethical issue that has arisen in regards to self driving cars is who to hold accountable in event of an accident. In accidents involving cars driven by humans, it is typically the drivers who are held responsible for their actions. However, with self driving cars, there is no clear party to hold responsible in the event of an accident. This is one of the primary reasons that the proper application of this technology requires standardization by a government agency. If a government agency makes clear the rules that are required to be followed by the artificial intelligence of self driving cars, then it becomes easier to apportion blame when appropriate. If engineers design a self driving car that fails to uphold the standards set by the government, then the corporation producing the cars should be held responsible. In addition, if an accident occurs due to a technological malfunction, the corporation again would be held responsible, as it would under normal circumstances. So if a technological malfunction leads to an accident that costs lives, the corporation should be held responsible, however, if upon review it is determined that artificial intelligence was properly designed in accordance to government regulation and it did not malfunction, then no party should be held criminally or socially responsible, as the accident should be regarded as just that, an accident. If the accident was unavoidable and the car functioned as intended then the corporation should not be held responsible. This situation is analogous to the vaccine debate. Theoretically, vaccines are a risk, although a minute one, for the people who elect to have them, however, not only do the vaccines benefit themselves, but they benefit the people who potentially could have contracted the illness from you, had you not received the vaccine. Like vaccines, self driving cars will inevitably be involved in rare deaths, but on average people will be much safer and there will be fewer vehicular related deaths in

general. In fact, the more this technology is integrated into society, the more lives will be saved as the technology improves. Tesla's already have an ability to update their computers as tesla works on and tweaks their algorithms to improve them. As other self-driving cars are introduced, there will be much more data gathered from the vehicles available to their developers. The developers can then use this data that they are accruing to continually update and develop their artificial intelligence to further minimize lives lost, and maximize safety.

The issue of fairness arises a lot in regards to how a self-driving car should act in the event of an accident. Passengers naturally feel that their lives should be given priority. For one, the passengers would typically elect to save themselves if they were the ones in control of the vehicle, and not an artificial intelligence. Another reason is that, if the passenger owned the vehicle, they feel it would not be fair for a product that they paid for, to elect to sacrifice them in order to protect others. On the other side of the conversation, bystanders think it is unfair that they might become the "target" of a self driving car who cares more about its stockholders and passengers then the people walking along the sidewalk. Again this is where the social contract comes into play. Both passengers and bystanders must recognize that although there may come a time when their lives are deemed to be expendable to accommodate the greater good, it is the greater good. The chances of either party dying with an autonomous vehicle is greatly diminished to that of a human driver. So although one might find themselves to be collateral damage, they will find that to be the case less often then it would be with human drivers. However, in the event where it is an equal number of statistical lives saved between the passengers or innocent bystanders, it is my belief that the priority should be given to the passengers. By prioritizing the passengers, the algorithm more closely resembles that of a human driver who would err on the side of self-preservation. I find it to also be more unfair for the

passengers to be sacrificed in the event of a tie, because they paid for the technology. Innocent bystanders have to deal with the reality that they can hit by a car either way, but will find that the chances of being hit by one are much lower with self-driving cars then human driven cars, so even by prioritizing passengers in the event of an accident in which an equal number of people per party are at risk, the innocent bystanders are still better off. It would be unfair to put an increased risk on the party that is paying for a vehicle that makes the lives of not only themselves, but the innocent bystanders, safer.

Risk Distribution

The distribution of risk shows why everyone should consent to the social contract of allowing self-driving cars to maximize statistical lives saved. If engineers programmed self-driving cars to follow a utilitarian standard of saving the maximum number of statistical lives, then the distribution of risk is applied equally to everyone in the society. There is no party who has a disproportionate amount of risk due to third party factors such as age or gender. So by ignoring factors such as age and gender, everyone has an equal risk of being harmed, but everyone in society also is probabilistically safer due to self-driving car technology. A study done by the National Highway Transportation Safety Administration proves just how much safer in general self-driving car technology would be for all parties. Across a time span of two and a half years, the National Highway Traffic Safety Administration investigated 5,470 accidents that were a weighted-sample that was representative of over 2,100,000 accidents across the country. The National Highway Traffic Safety Administration found that of all those accidents, 94 percent of the time, the accident could be contributed to driver error or misconduct. The next closest

cause of the accident was vehicle malfunction at just above two percent. Theoretically, if properly programmed, over two million accidents could have been avoided nationwide, if we removed the human element from driving and replaced it with an artificial intelligence. This is not only saving lives of both drivers or passengers and innocent bystanders, but it is also saving countless dollars worth of damages for taxpayers. Artificial intelligence in self driving cars will never find itself distracted by a phone or their passengers, nor will it ever disobey traffic laws or take unnecessary risks. A self-driving car's behavior is infinitely more predictable than that of a human and it's that predictability that makes them that much safer.

Part of the reason that the standards that govern the rules of self-driving cars need to be centralized by the government, is because people can rarely come to a consensus on anything in this world, and in those cases, it should be up to the government to act in people's best interests. In the case of the ethics governing self-driving cars, each demographic would have a different perspective on the issue, and this is not limited to demographics such as age or gender, but culture as well. The young would believe they should be prioritized over the old, but in certain cultures, they believe the old should be prioritized over the young. Women may think that they should be prioritized over men, but men may feel that this would be unfair. The simplest way to create a fair environment is to remove all qualifiers other than saving the maximum number of statistical lives saved, as would be the utilitarian method. The clearest consensus of the moral machine survey, whose goal was to explore peoples preferred ethics for self driving cars, was to save the greatest number of lives when possible, and the fairest way to do so would be to ignore any of the other qualifiers such as age or gender. If saving the maximum number of people is what is most agreed upon, then the only question that remains, is what should the self driving car do in the event where there is an even statistical probability of saving an even number of lives? If

there are three passengers and three pedestrians in jeopardy, whose lives should the artificial intelligence prioritize? In my opinion, in these scenarios the car should prioritize its passengers. If we acknowledge that the risk distribution of the self driving car technology is equal among all members of society if we only prioritize saving the maximum number of lives, and that the total risk being distributed is lower due to the technology, then out of fairness, the passengers should be given priority. One of the reasons it is more fair to prioritize the passengers is because they are paying for this technology that the pedestrians are benefitting from. The overall cost incurred on society as a result of vehicular accidents is being diminished by this technology, and this cost is being subsidized by the people who are paying for the technology. Pedestrians pay no additional costs for the additional safety they are receiving due to the use of the self driving car technology. It would not be fair to the people who are paying for society to benefit from the technology to be granted a slightly larger distribution of the risk when they are already reducing the overall risk substantially due to their investment.

The Value Of a Life

The reason the Utilitarian approach of saving the maximum number of statistical lives is the best approach to self driving car technology boils down to fairness. The utilitarian approach does not leave any demographic feeling disenfranchised as it does not value any particular parties lives over another. Everyone's life is given an even amount of value in regards to their right to life. So if everyone's life is valued equally, then the only thing that matters is saving the maximum number of lives. Coupling Utilitarian ethics with self driving car technology leads to an equal distribution of risk across everyone, but the sum total amount of risk being distributed is

being severely reduced due to the increased safety provided by the self driving car technology. No one should feel as if they are being "targeted" as no age or gender demographic is being considered less important than others, as it should be. Objectively, one life is equal to one life, and it is not fair to say that anyone's life should be valued over anyone else's. This is necessary if there is to be a social contract amongst society to accept that although everyone's life is safer, there is a chance that they could find themselves on the tragic end of an unfortunate accident, but that the likelihood of this happening is far lower than if there were a human driver behind the wheel. However, if people knew that a certain demographic was being given priority over another, those people given a lower priority might be loath to accept this social contract as they would feel that they are unfairly being undervalued in their society.

Conclusion

Self-driving car technology is clearly beneficial to society if applied properly. When the vast majority of car accidents are attributed to human error, then the vast majority of car accidents are avoidable with this self driving car technology. There should be no reason to fear this technology if we distribute a lower total amount of risk evenly across society. Not only can this technology increase the safety of everyone in society, but there are multiple other benefits as well. Decreased traffic can improve everyone in society's productivity, adding to the utility provided to society. Not only can people more efficiently allocate their time, but the resources being wasted on damages due to car accidents will also be minimized. However, it is up to the government to standardize the rules governing the artificial intelligence of self driving cars in order to guarantee consistency across these cars, as well as provide people with the reassurance

that their lives will be given equal consideration in the event of an accident. This reassurance will give people the confidence to opt into another social contract to the entire society's benefit.

Sources

https://www.nhtsa.gov/risky-driving/drunk-driving
https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115
https://towardsdatascience.com/the-ethics-of-self-driving-cars-efaaaaf9e320
https://www.moralmachine.net/
https://www.landmarkdividend.com/self-driving-car/