Deontological Ethical System For Google's Self-Driving Car

by Edgard Alejandro Arroliga
California Polytechnic State University
Computer Science Department
Advisor: Dr. Clark Turner

May 31, 2016

Abstract

Google's Self-Driving Car is a revolutionary product that is riddled with ethical conundrums. It is able to accurately scan and drive through densely populated roads without much difficulty. However, there are some situations where the car will likely have to make decisions that affect, maybe even take, the lives of those on the road. Issues such as the Trolley Problem and the Rear-End Dilemma describe situations where there seems to be no single ethical answer as to how the car should act. In order to solve these issues, I propose that a Deontological Ethical System should be implemented because it is predictable, consistent, and easy to implement as an algorithm once the rule set has been decided.

Contents

1 Google's Self-Driving Car and The Importance of Thought Experiments

1.1 The Self-Driving Car

The Google Self-Driving Car Project is a leap in innovation and convenience. Google's goal is to create a world where the convenience of car travel is available to anyone, no matter their inability to drive a conventional car.[15] They also wish to reduce the amount of car accidents drastically, seeing that 94% of all car accidents are caused by human error.[15] By removing the human element of driving a car, they may be on the right track.

The self-driving car is able to work accurately by being able to process both map information and sensor information.[15] The sensors are able to pick up information from the nearby vicinity and can accurately classify objects through their size, shape, and movement pattern.[15] After receiving and classifying the information from the sensors, the car's software is able to predict what the object is going to do and adjust its speed and trajectory accordingly.[15] Just like a human driver, the self-driving car keeps in mind where it is, what is around it, what is going to happen, and how should it respond.[15] However, no matter how good the car is at answering these questions, it may run into situations where there is no clear answer.

1.2 Thought Experiments and Their Value

Thought experiments can be defined as "devices of the imagination used to investigate the nature of things."[8] These thought experiments are usually presented as a sort of narrative or diagram that explains the situation. [8] They also have a couple of common features, "we let it run ... ; we see what happens; finally, we draw a conclusion." [8]

Some readers of this paper may believe that these devices have little to no value because they do not rely on empirical data. Many believe that an experiment without empirical data is worthless and cannot tell us anything about the real world.[8] However, "thought experiments help to illustrate and clarify very abstract states of affairs, thereby accelerating the process of understanding." [8] We want to be able to understand the self-driving car and all its nuances as fast as possible. After all, these imaginary situations actually do have a chance of appearing in reality, and we have to be able to answer them up front. We need to provide a solution before we run into a problem, especially when human lives are at risk.

2 Trolley Problem and Rear-End Dilemma

In this section, I will introduce the two thought experiments to be addressed throughout the paper, the infamous Trolley Problem and what I like to call the Rear-End Dilemma. Both of these thought experiments are hypothetical situations that the Google Self-Driving Car may have to deal with. We need to discuss these issues because we have to implement a transparent solution into the car software so that if the situation ever arises, we can properly deal with it. However, the problem is coming up with a solution that is satisfactory and plausible.

2.1 The Trolley Problem

The Trolley Problem was introduced in 1967 by Philippa Foot in her paper *The Problem of Abortion and the Doctrine of the Double Effect.* [13] The original paper, as the title suggests, was about the problem of abortion but in order to highlight an ethical dilemma she proposes a thought experiment. [13] She tells us to imagine a runaway tram with the tracks splitting in two. There are 5 people working on the track the runaway tram is on and a single person on the other. You, the observer, are next to a switch and have full control over which track the runaway tram will go to. Since one "group" is fated to die, it is up to you to decide whether to switch from the group of 5 to the single person. [13] Is the right decision to save the group of 5 because more lives are saved? Is it okay to kill a single person if it meant saving more people? What if the group of 5 included heinous criminals while the single person was a young father of 2? There are many of these variations on this problem, all with the same basis of putting value on human life. This ethical problem has become surprisingly relevant with the dawn of self-driving vehicles.

Lets imagine a situation involving the self-driving car. You approach an intersection when the car spots that there is a runaway driver on course for a head on collision. It assesses the situation and determines it can swerve onto the sidewalk in order to avoid a collision. However, there is a group of pedestrians safely walking that would be either heavily injured or killed as a result of the avoidance. Should the car save the owner of the car from the collision at the cost of the pedestrian lives? Should it take the collision in order to save more lives?

There are multiple variations of this situation for the self-driving car. We can imagine that there are two motorcyclists driving alongside your car, one is wearing a helmet while the other is not.[10] The car in front of you makes a sudden stop and you are on course for a collision unless you swerve to the left or to the right. Should the Google car target the helmet wearing motorcyclist because he has a higher chance of surviving than the other motorcyclist? Something about that decision doesn't feel particularly right, because the motorcyclist wore the helmet in order to be safer, and now he is being targeted for a collision! Perhaps the best option would be to take on the collision and risk the safety of the self-driving passenger. However, didn't the owner of the self-driving car buy the car to be safer? We need to explore the options further with multiple ethical perspectives.

2.2 The Rear-End Dilemma

What I like to call the rear-end dilemma is another thought experiment along the lines of the Trolley Problem with a twist. Lets imagine another situation with the self-driving car. You are driving down a heavily populated city. You approach an intersection with a red light, so you begin to stop. There is a pair of children crossing the intersection safely when you spot that there is a speeding driver on course for a rear-end collision. The Google car observes that it can make a legal and safe right turn maneuver to avoid the collision. However, allowing the car to go through will mean the children crossing the street will be hit and likely killed. Should the self-driving car be able to make the right turn because it is a legal and safe maneuver? Or should it be held responsible for saving the lives of the children and take the hit?

The interesting difference in this situation is that it does not force the car to make a direct decision of who it kills. In the Trolley Problem, the self-driving car makes a direct decision that will end up hurting someone else because there are no other safe alternatives for the car. However, in the Rear-End Dilemma, the car can avoid the collision in a safe and legal fashion. This is what is ideal for the self driving car, since it is safe and does not break any laws. However, should the car consider the lives of children, or any pedestrian for that matter, when it makes decisions? It seems as though if the car is within power of saving lives, it should always do so, even at the inconvenience of the owner. I believe that the the individual owner of the car is the customer, not society as a whole. After all, the owner is the one paying for the convenience and safety of the self-driving car, not society. We need to find a solution to these issues before we face them in real life.

3 Consequentialism vs. Deontology

There are various ethical systems/theories that can help to answer our questions. However, we want a solution that can be implemented as an algorithm in software and that provides a reliable experience to the owner of the car. With this limitation in mind, we will explore two theories, Consequentialism and Deontology.

3.1 Consequentialism

Consequentialism is the "view that morality is all about producing the right kinds of overall consequences."[4] Utilitarianism is a very well known example of Consequentialism, specifically Act Consequentialism. Act Consequentialism states that "an act is morally right if and only if that act maximizes the good." [18]. We can also define the consequence of an action as "everything the action brings about, including the action itself."[4] Let us also define "good" to be "human welfare" or "good fortune, health, happiness, prosperity".[1] So to choose the best possible consequence is to choose the consequence which maximizes human welfare. With this, we can use Consequentialism to come to a decision in our self-driving car thought experiments.

3.1.1 Trolley Problem

Lets recall the situation above, with a self-driving car going into a head-first collision with a runaway car and a group of pedestrians on the sidewalk. On the surface, Consequentialism would say that the best decision to make is to take the collision head on, making a selfless act and saving the lives of the group of pedestrians. This is because more lives would be saved by making that decision. The overall consequence of that action is greater because only one life is lost compared to the group. However, I believe this answer is flawed.

The average person does not wish to be in a situation where a group of people will die for their sake. Nevertheless, the goal of the Google Self-Driving Car is for "everyone [to] get around easily and safely, regardless of their ability to drive."[15] The product goal is to provide a safe alternative to the owner of the self driving car but in the situation above the car chooses to disregard this principle. If a person is willing to spend money on a car that is supposed to be a safer and more reliable alternative to driving, they should be saved from avoidable accidents. The problem here is that having a Consequentialist self-driving car will lead to **unreliable and unsafe** outcomes for those involved in an accident with the self-driving car.

Lets look at the example with two motorcyclists and the impending crash as stated above. This situation does not involve choosing the life of the user over the life of many people but rather focuses on choosing between two people of different worth[1]. The Consequentialist solution would be to choose the person with the lowest "worth" and hit them. A big problem with the situation is that the self-driving car can not put any value on the two motorcyclists! The self-driving car only has sensors that can distinguish shapes and can categorize those shapes into accurate predictions.[15] From looking at those shapes it can accurately tell that the two motorcyclists are in fact motorcyclists but it cannot tell which one is worth hitting! Making a decision that has the best consequence would be **difficult to implement**. This is because the Google Self-driving car only uses 2 visual cameras.[15] These cameras are used as stereo cameras meant to judge the distance as well as to visually see if there is an obstacle.[19] The lidar, or light detection and ranging, system is what does most of the sensor work.[19] The cameras are not meant to visually analyze a specific object, which is what it would need to do in order to determine the worth of the motorcyclist.

However, this leads to another problem where the car won't be able to make a decision because of technological limitations. In order to make a proper Consequentialist decision it has to determine which consequence is the best. If the self-driving car did hypothetically have the proper technology to determine every factor from visual recognition, it would still need to run scenarios that would determine which outcome is the best and the number of scenarios increase drastically with more information. For example, with better technology perhaps the car could estimate that the car behind him is moving slower and can take the impact without much damage. So now the car would have to compare this outcome with the outcome of swerving into the helmeted motorcyclist. There could be many more of these estimations, all adding significantly more computation. For example, in order to process a video it would need more processing speed and storage.[11] A solution to this could be to link to clusters of computers over a network but "adding com-

[1] Worth is defined as "a quantity of something of a specified value."[1] The value referred to here is a metaphysical value such as how much a person contributes positively to society.[17]

puters involves considerable data transfer over a network, which can be bound by input-output restrictions, further limiting processing speed."[11] Secondly, in order to calculate each scenario correctly it would need to know the worth of the two motorcyclists, as stated above. This would mean it would have to accurately assess things like "this motorcyclist has a helmet on", and "this motorcyclist has committed more crimes". According to the CTO of Dynamic Ventures, Itzak Ehrlich, most computer vision is done through machine learning and "learning by example" so it would be **unfeasible and impractical** to implement.[12] This is because to calculate the value/worth of each individual would require creating a database of all possible objects and people, an impossible task.[12]

3.1.2 Rear-End Dilemma

We can tackle the Rear-End Dilemma described above with Consequentialism. The best possible consequence in this situation would be to take the rear-end and possible give up your life in order to save the life of the children crossing the street. This situation shares the same issue with the Trolley Problem, where the self-driving car is purposely choosing to sacrifice its owner to save the lives of others. However, in this situation it could have made a safe and legal maneuver for its owner. The Google Self-Driving car was designed to be a safe alternative to driving a vehicle. By deliberately choosing to injure or kill the owner of the self-driving car it has become a "self sacrificing super hero" of sorts but not a safe alternative to driving, especially when there was a legal maneuver that it could have executed. If there were no children crossing the street, the car would have chosen to make the legal maneuver. This makes the car too unreliable for the user, because in one situation the car would save them and in the other, it would doom them.

Consequentialism proves to be **too unreliable** and conflicting with the goal of the Google self-driving car. In order for the self-driving car to be a good product, it needs to "guarantee"[2] the safety of those using it. Most people would choose to save the lives of others, but they would like to know when that decision was being made. In addition to being unreliable, **is not easy to implement**. The car would have to have computer vision software that has machine learned every possible object and person and can then assess the safety or worth of that person.[12] We need to find a solution that is easy to implement, reliable, and predictable.

3.2 Deontology

Deontology is derived from the Greek words *deos* meaning "duty" and *logos* meaning "the study of".[7] Duty-based ethics, in direct contrast to Consequentialism, does not look at the " states of affairs ... choices bring about" but rather states that "some choices cannot be justified by their effects."[7] According to Kant, a primary proponent in Deontology, "the sole feature that gives an action moral worth is not the outcome that is achieved by the action, but the motive that is behind the action."[6] People have the duty of doing the right thing, even if the result is bad. In order

[2] However, this is impossible to guarantee

to know what is “right” a rule set is usually set in place such as “It is wrong to kill innocent people” or “It is wrong to tell a lie.”[5] So in order to discuss the Trolley Problem and the Rear-End Dilemma, we need to come up with a Rule Set for the self-driving car.

3.2.1 Deontological Rule Set

The following Rule Set is my own derivation and takes inspiration from Asimov’s Three Laws of Robotics.[2]

1. **Minimize the harm to the driver and passengers of the self-driving car.**
2. **Minimize the harm to any human outside of the self-driving car, as long as it does not conflict with the first rule.**
3. **Do not destroy the property of others, as long as it does not conflict with the previous two rules.**
4. **Follow all of the rules of the road that apply to your current location, as long as it does not conflict with the previous 3 rules.**

   We can define “minimize” as “to reduce to the smallest possible amount or degree.”[1]

3.2.2 Rule Set Rationale

This is my proposed rule set for the self-driving car. It offers guidance in the most essential area of driving, the safety of human beings. The safety of the people is one of the biggest reasons driving laws (and laws in general) are in place.[9] Along with driving laws, traffic signals are in place in order to protect the people driving, because driving is inherently a very high-risk activity. With this in mind, I chose to put the safety of the human beings as a higher priority than the destruction of property and the breaking of traffic laws. This is because I believe that human life, or a human being that has experiences, has intrinsic value.[17] People have the intuitive view that “the more people that exist, the better.”[17] Human consciousness may be a factor but human value is hard to define and is a discussion for another paper.[3][17] The goal of the Google Self-Driving Car is to “transform mobility by making it easier, **safer** and more enjoyable to get around”, not reduce the destruction of property.[16]

The reason the driver/passenger is placed before other humans is primarily because the self-driving car is a product. A survey of 900 participants showed that 75% of people believed that the self-driving car should swerve and kill the driver in order to save even one pedestrian. [14] Another similar survey was given to hundreds of Amazon’s Mechanical Turk workers and the results were similar, “In general, people are comfortable with the idea that self-driving vehicles should be programmed to minimize the death toll.” [3] However, the participants were not confident that the car would be programmed in such a fashion because “they actually wished others to cruise in utilitarian autonomous vehicles, more than they wanted to buy utili-

[3] The topic of intrinsic value and whether human life has intrinsic value is a topic hotly debated among ethic philosophers.

tarian autonomous vehicles themselves."[3] This is one of the biggest issues with the utilitarian self-driving car and the reason I put the driver's life before the pedestrian's. Since the Google Self-Driving car is a product with the goal of "making it easier, safer and more enjoyable to get around", the owner of the car should be entitled to their safety.[16]

3.2.3 Trolley Problem

Now we can see how the Deontological Self-Driving Car would react should it face the Trolley Problem. We can look at the situation with a self-driving car going into a head-first collision with a runaway car and a group of pedestrians on the sidewalk. In this situation the Deontological Self-Driving Car will look immediately at its 4 Rules. It needs to avoid the head-first collision, since that would break the very first rule, minimize the driver/passenger of the self-driving car. The second rule states that the car should minimize harm to any other people outside the car **unless** it conflicts with the first rule. Since we devised this situation to have no other solution other than running into the pedestrians, it would choose to run into the pedestrians because otherwise the first rule would be broken. However, if there was a hypothetical situation where the only result were to have damaged property or a broken traffic law, such as swerving into a parked car or speeding to avoid the collision, the self-driving car would choose those. Choosing to kill an innocent pedestrian seems like an awful decision to make, and it is, but its a decision that needs to be made. More importantly, it a decision that, while hopefully rare, is **consistent and implementable**.

Lets take a look at the second trolley situation. In this situation we have the two motorcyclists, one with helmet and one without, and the impending crash in front of us. Again, in this situation there is no way to "break the law" or "destroy property" in order to save lives so we will have to crash into one of the motorcycles in order to save the life of the driver/passenger. Now, here is where we find the biggest problem with a Deontological Rule Set: it is not flexible enough to make a decision on which motorcyclist to hit. One of the biggest weaknesses of Deontology is the fact that it can't deal with a situation where two duties conflict, the two duties being to not hurt either motorcyclist.[5] This can be solved by implementing a sort of system that takes into account the "approximate safety" of other pedestrians. The "approximate safety" here can be defined as the likelihood of a person's survival. However, this "approximate safety" value has to be simple enough to calculate using just the regular car scanners. For example, it can determine bigger objects as having a higher chance of survival and objects that arent human shape as having a higher chance of survival (for example a person in a car as opposed to a motorcyle). The car can store the information on the "approximate safety" of those surrounding the car and have it ready in case something goes wrong. It is easy to **implement** this sort of solution because "approximate safety" can be calculated as soon as something enters the proximity of the car and discarded as soon as it leaves. It does not have to be computed in the split-second the crash occurs and it does not have to compute many different variations of a situation like in the Consequentialist car. With this solution and the rule set, the Deontological Self-Driving Car is **easier to implement and produces consistent results.** The results may be up for debate according to your ethics, but it will always make the same decision, no matter what situation.

3.2.4 Rear-End Dilemma

The Rear-End Dilemma sees us in an impending rear-end collision unless we make a legal maneuver to turn to the right. However in us making the legal maneuver, we kill indirectly kill pedestrians crossing the road by letting them take the hit. Using our Deontological rule set, the decision seems to be to make the legal maneuver and let the car hit the pedestrians. This would directly violate the second rule, **only** if we were aware of the pedestrians and their danger. Lets assume that we were aware that the pedestrians were in danger of being hit, then we cannot make the legal maneuver, we must minimize the harm to any other person outside the car.

This highlights the big issue with a Deontological system, it is inflexible when it comes to these ethical situations.[5][7] If there was no way to minimize the harm of the pedestrians without also minimizing the harm of the driver/passenger, the pedestrians would be left to die. If there was a way to minimize the harm to both the driver and the pedestrian it would make that decision no matter what, but that decision is impossibly hard to reach. As we saw in the consequentialist car, calculating how to minimize the damage would be impossible to implement, especially within the short time frame of a car crash.[12] Thus, given our technology and rule set, the pedestrians crossing would be left to die. While the solution may not sound pretty or appealing, it will be **consistent.** The Deontological Self-Driving Car will always make the same decision, to let the pedestrians die, because it does not have the proper computation power to see how to minimize harm.

4 Conclusion

Having a Deontological Self-Driving Car sounds like the least optimal solution in terms of saving lives. While this may be true, there is more value in the Deontological Self-Driving Car because it is feasible given our technology and it provides consistent outcomes no matter the situation. A Consequentialist Self-Driving Car sounds more appealing, because who wouldn't want to save more lives, but it is impractical and inconsistent. Having to make decisions based on "the best possible consequence" in the limited time that a crash is recognized as happening would not work. Even if it did work, it would result in crashes with varying outcomes, sometimes it chooses to kill the driver and other times it chooses to kill the pedestrian. This impracticality and inconsistency is not good for a product that is being sold to "increase safety and reduce deaths."

References

[1] "Dictionary.com." [Online]. Available: http://dictionary.reference.com

Dictionary to define many of the ambiguous terms

[2] "Isaac asimov's "three laws of robotics"," 2001. [Online]. Available: http://www.auburn.edu/~vestmon/robotics.html

Link to Asimov's Three Laws of Robotics

[3] "Why self-driving cars must be programmed to kill," October 2015. [Online]. Available: https://www.technologyreview.com/s/542626/why-self-driving-cars-must-be-programmed-to-kill/

another utilitarian car survey, with conflict of wanting to own one

[4] "Consequentialism," February 2016. [Online]. Available: http://www.iep.utm.edu/conseque/

Peer reviewed academic resource on consequentialism

[5] "Duty-based ethics," February 2016. [Online]. Available: http://www.bbc.co.uk/ethics/introduction/duty_1.shtml

Introduction and explanation of duty-based ethics

[6] "Immanuel kant: Metaphysics," February 2016. [Online]. Available: http://www.iep.utm.edu/kantmeta/#H8

Overview of Immanuel Kant with section about his ethic beliefs

[7] L. Alexander and M. Moore, "Deontological ethics," in The Stanford Encyclopedia of Philosophy, spring 2015 ed., E. N. Zalta, Ed., 2015.

[8] J. R. Brown and Y. Fehige, "Thought experiments," in The Stanford Encyclopedia of Philosophy, spring 2016 ed., E. N. Zalta, Ed., 2016.

[9] J. L. Center, "Law and the rule of law," March 2016. [Online]. Available: http://judiciallearningcenter.org/law-and-the-rule-of-law/

Why are laws important and what is their goal

[10] I. Chipman, "Exploring the ethics behind self-driving cars," August 2015. [Online]. Available: https://www.gsb.stanford.edu/insights/exploring-ethics-behind-self-driving-cars

Explores the ethics behind the self driving car, including the trolley problem

[11] J. Edgell, “4 limitations of facial recognition technology,” November 2013. [Online]. Available: http://www.fedtechmagazine.com/article/2013/11/4-limitations-facial-recognition-technology

Limitations on facial recognition and visual processing using a camera

[12] I. Ehrlich, private communication, 2016.

[13] P. Foot, “The problem of abortion and the doctrine of the double effect,” 1967. [Online]. Available: http://pitt.edu/~mthompso/readings/foot.pdf

Introduction of the Trolley Problem

[14] O. Goldhill, “Should driverless cars kill their own passengers to save a pedestrian?” November 2015. [Online]. Available: http://qz.com/536738/should-driverless-cars-kill-their-own-passengers-to-save-a-pedestrian/

source about survey for utilitarian car

[15] Google, “Google self-driving car project,” 2016. [Online]. Available: https://www.google.com/selfdrivingcar

Google’s official page on how the google car works

[16] ——, “Google self-driving car project,” 2016. [Online]. Available: https://www.google.com/selfdrivingcar/faq/#q2

Link to the goal behind the Google self-driving car

[17] J. Gray, “Does human life have value?” October 2010. [Online]. Available: https://ethicalrealism.wordpress.com/2010/10/14/does-human-life-have-value/

Article on human life and its value

[18] W. Sinnott-Armstrong, “Consequentialism,” in The Stanford Encyclopedia of Philosophy, winter 2015 ed., E. N. Zalta, Ed., 2015.

[19] R. Whitwam, “How google’s self-driving cars detect and avoid obstacles,” September 2014. [Online]. Available: http://www.extremetech.com/extreme/189486-how-googles-self-driving-cars-detect-and-avoid-obstacles

Specifics on the sensor systems behind the Google Car, LIDAR, and their stereo camera