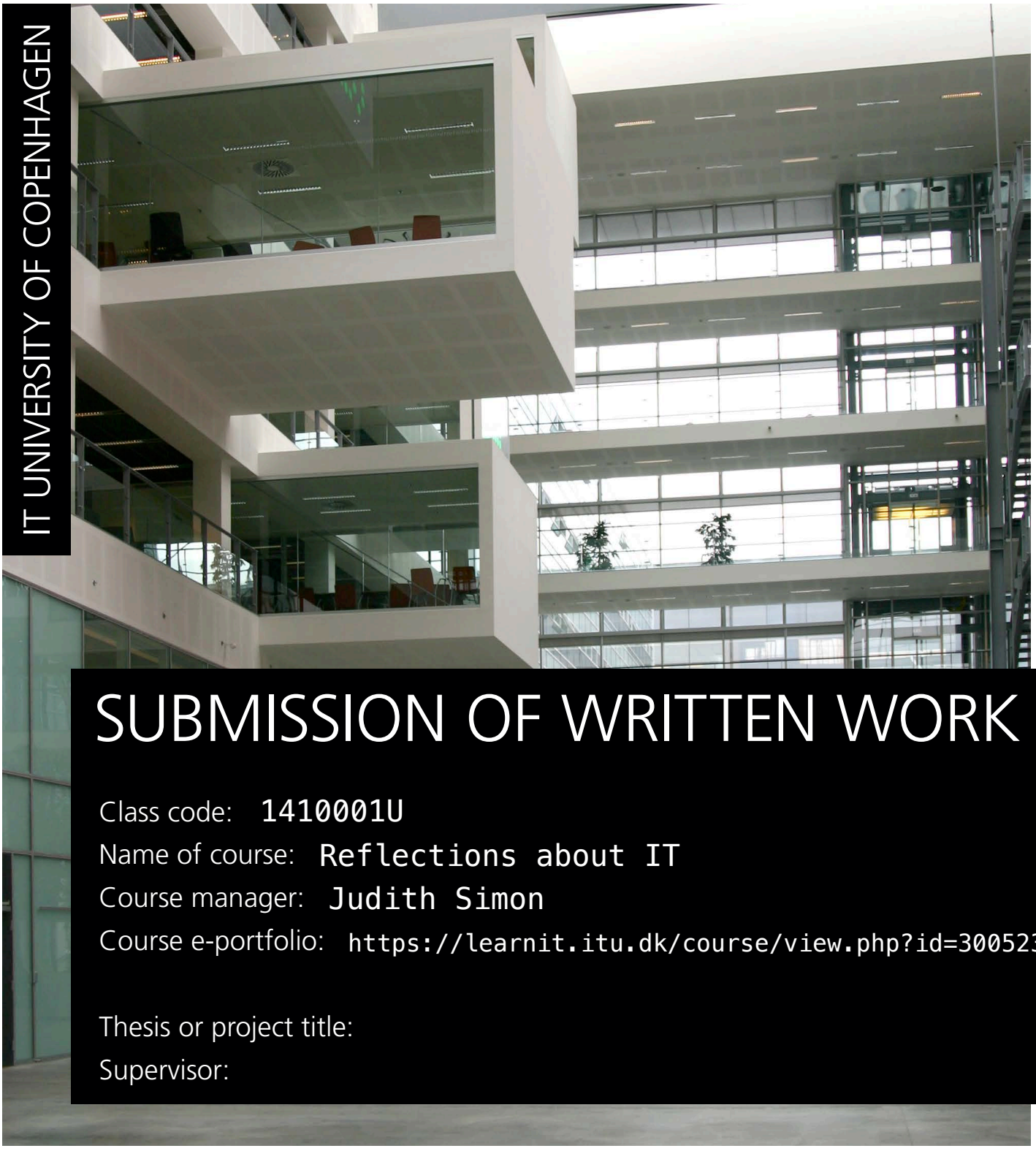
IT UNIVERSITY OF COPENHAGEN



SUBMISSION OF WRITTEN WORK

Class code: 1410001U

Name of course: Reflections about IT

Course manager: Judith Simon

Course e-portfolio: https://learnit.itu.dk/course/view.php?id=3005237

Thesis or project title:

Supervisor:

	Full Name:	Birthdate (dd/mm-yyyy):	E-mail:
1.	Anders Fischer-Nielsen	06/05-1993	afin@itu.dk
2.			@itu.dk
3.			@itu.dk
4.			@itu.dk
5.			@itu.dk
6.			@itu.dk
7.			@itu.dk

Who Should Your Autonomous Vehicle Choose to Crash Into?
Ethics and Autonomous Vehicles

Anders Fischer-Nielsen

May 27th 2016

Contents

Chapter 1

Introduction

The idea of creating mass-produced autonomous vehicle has seen increased interest from researchers and car manufacturers in the last decade.

The idea of inventing an autonomous, self-driving car is not a new thing. Dreaming up future societies where people would focus on other things while their cars drove them from A to B, has been a trope of science-fiction for decades[1]. During the last decade these ideas are finally being realised with research projects such as Google's ongoing *Self Driving Car Project*, previous research projects such as the entries of Carnegie Melon and Stanford in the *DARPA Grand Challenge* [17] and the recent undertakings of most car manufacturers, such as BMW, Audi, Toyota, VW etc. [11]

Certain ethical issues arise with autonomous vehicles, however. Since autonomous vehicles are programmed in advance to handle every possible scenario on the road, autonomous vehicles present ethical dilemmas that human drivers in non-autonomous vehicles do not, since an action has to be decided on *before* the vehicle finds itself in an ethical dilemma.

Human drivers are rarely blamed for acting according to their instinct in life-threatening vehicle collisions, even if this means that they inadvertently make decisions that have fatal consequences for other people.

These decisions have to be programmed into autonomous vehicles in advance in order to make the "right" choice, should they occur. This raises interesting questions, such as; who should be

[1] The popular TV show, *The Jetsons* is one of the more well-known examples of having autonomous vehicles. [19]

blamed for an autonomous vehicle hitting, and possibly killing, one person over another? Should we even program autonomous vehicles to be able to select the "preferable" target of collision, or would it be better to not do any selection at all and make the vehicle choose some random behaviour in a critical situation?

The most interesting question, and therefore the one I will try to answer in this essay is; "Who is to blame for collisions or harm done to people or structures by an autonomous car in a critical situation, and can this blame even be placed on anyone?" by looking into what ethical issues this question presents, and what ethical theories that can help answer this question.

Chapter 2

Autonomous Cars

This essay will focus on *autonomous* cars, and not *automated* cars. Autonomous is, as per Thesaurus, defined as:

> "*an autonomous republic*: self-governing, independent, sovereign, free, self-ruling, self-determining, autarchic; self-sufficient."

Automation implies that cars merely follow artificial hints in the environments, such as early experiments using magnetic strips in the road. [18] *Autonomous* implies that cars react to their environment independently, that is they cannot depend on unnatural artefacts in their surrounding environment in order to follow the road, not crash and avoid obstacles.

Experiments attempting to automate vehicles, mainly cars, have been made since the 1920's with varying degrees of automation and success. Modern research has been focused on autonomous cars, since it is unrealistic to add artificial hints on every road in the world.

Instead, research has been focused on making autonomous cars adapt to environments with uncertainties, so that the same autonomous vehicle can drive both in the inner city and on unpaved mountain roads. Modern vehicles furthermore have an increasing amount of automation built in, which has not brought any ethical dilemmas. On the other hand, the idea of having truly autonomous vehicles brings certain dilemmas with it which can be researched.

Recent research projects use radar or radar-like technology in addition to GPS, odometers and computer vision in and on cars to detect the environment surrounding the car and recognise obstacles, such as people, other cars and structures, which the car will try to avoid.

The addition of the radar-like *LIDAR* technology (a mix of the words light and radar) on the roof of the cars has provided the cars with a 200 foot-radius "view" of their surroundings, enables the algorithms in the cars to register the world around the cars in great detail. *LIDAR* generates a very precise point cloud of the environment surrounding the car, enabling the on-board software to distinguish a running child from a bicyclist. Over time, this technology gets more precise, informing about the world around the cars in even greater detail.

Google has been researching autonomous cars since 2009 [8]. Over the years the research project has been ongoing, the cars have developed an increasingly detailed view of their surroundings. In addition to the computer vision research Google has developed and implemented in the cars, *LIDAR* helps the car map its surroundings, enabling it to recognise smaller objects such as pedestrians and bicyclists. Google is using the computer vision technology it has been developing and researching to analyse the surrounding environment perceived through stereo cameras mounted on the car to detect possible collisions around the car. The sensors on the car generate 1GB of data per second [3], which is analysed in order to give the most precise depiction of the surrounding environment of the car.

2.1 Motivations for Developing Autonomous Vehicles

The main motivation for developing autonomous vehicles is that human drivers are prone to make mistakes. Human drivers get distracted, have relatively slow reaction times, especially when tired or under the influence of drugs. Driving with a lack of sleep, or while under the influence of drugs, severely inhibits the driver's ability to react in time to avoid collisions.

An autonomous car can register its surrounding environment many times a second and analyse this input to decide on the best possible action to take in a given situation. An autonomous driver does not get distracted unless programmed to do so, and has as fast a reaction time as hardware and software allows, almost guaranteed to be less than that of a human driver. Finally, an autonomous driver does not get drowsy and cannot ingest drugs.

An autonomous driver will therefore, following this reasoning, be a safer driver.

10% of all crashes in the United States were crashes where the driver of the car was identified as distracted *immediately* before the crash. In 2014 these distracted drivers killed 3179 people and left 431000 people injured. [1]. Data on how many accidents have been caused by speeding and driving aggressively or recklessly are hard to find, but account for some percentage of all crashes

as well. Futhermore, 31% of all driving-related crashes were caused by impaired driving in 2014, that is driving while under the influence of alcohol or the like, with 9967 people killed as a result of these crashes. [2]

Eliminating distracted driving and impaired driving would therefore eliminate 41% of all driving-related crashes, sparing the lives of 13146 people every year. Taking human error out of the equation is hoped to save many people from injury or worse.

Furthermore, costs from paying human chauffeurs etc. can also be reduced, traffic jams might be reduced[1], vehicles might be able to park more efficiently, and generally drive faster while still driving safely, saving time spent going from A to B.

In short, a computer driving your car will probably be a better driver, making fewer mistakes than you. An autonomous car can sense its surroundings 200 times per second and make just as many calculations reasoning for its next move based on the input. A human driver simply cannot top that. [4]

This fact presents some interesting questions. Given the amount of input and processing power, an autonomous car should always be able to make the best possible choice in any given situation. An autonomous car would register the child running across the road before a human driver ever could, and should therefore always make the right choice accordingly.

But what *is* the right choice? Most people would say that the car should always try to harm as few people as possible. That would be the "ethically correct thing to do". [5] But what happens in situations where someone *has* to get hurt? If an autonomous car is in a situation where it has the option of hitting two different people, but no option to avoid either person, who should it choose to hit? What if saving both people involves killing the passengers of the car? As Jean-François Bonnefon, after having conducted a major study on autonomous cars and their acceptance into society, aptly puts it:

> "As we are about to endow millions of vehicles with autonomy, taking algorithmic morality seriously has never been more urgent. [5]"

[1] So-called shockwave traffic-jams caused by human error, are researched in [12]

Chapter 3

General Ethics and Computer Ethics

Overall, ethics is the study of morality and moral systems. These moral systems are compromised of many components, with some common features. Bernard Gert has described the four main features of a moral system as; Public, Informal, Rational, Impartial. [14] These can be described accordingly as; The rules of the system should be know to all its members, the rules should be based on logical reason accessible to its members, there is no authority enforcing them, and the system does not treat individuals or groups differently. Furthermore, the rules of the system have connections on different levels.

The first level of the system, which is the one that we interact with, is the rules of conduct that the system sets for us. These are rules that all members agree on, and that everyone is subjected to equally. These can either guide the actions of people or help establish social policies.

Rules of conduct are derived from a set of "basic moral values", a subset of the core values of a society, which are important to its thriving and survival. The rules are subject to principals of evaluation that will justify the rules. These principals are either Religion, Law or Ethics. [13]

The moral system of computer ethics is based upon the systems of general ethics, and uses the same concepts and categories as base values. The values in the field of computer ethics changes alongside the evolution of the technology, however, and actions occurring brings a policy vacuum that can end in new values being formed and changes happening to existing policies. As said by Moor:

> "Computers provides us with new capabilities and these in turn give us new choices for action. [...] A central task of computer ethics is to determine what we should do in such cases, i.e., to formulate policies to guide our actions." [10]

Even though actions made by computers are the same as tasks previously performed by humans, tasks which already have certain morals, the way the action of a computer has been implemented can change the way the it is viewed.

A problem of this can be described as "the invisibility factor", which is that operations of a computer cannot be seen directly. It is therefore difficult to know if the operations are unethical. [10] Invisibility of actions in computer systems makes it difficult to evaluate with ethical principles, but it is possible to examine the design of a system, and how this design makes people act. It can be examined if the system has been designed in a way as to promotes unethical behaviour. It is possible to evaluate the morals of actions someone has taken in many ways. Therefore a perspective to look at these actions has to be chosen.

3.1 Consequentialism

In the study of normative ethics, consequentialism holds that the consequences of one's conduct are the basis of which to determine the rightness or wrongfulness of one's actions. That is, the means to which you achieve your goal pose no ethical relevance, rather the end result is what has ethical relevance. It is the idea that the end justifies the means. [9]

3.1.1 Utilitarianism

Utilitarianism is a form of consequentialism, founded by Jeremy Bentham, holding that the best moral action is the one that maximises *utility*, that is the well being of sentient beings:

> "The ethical theory that holds that the action that is morally right is the one that results in the greatest possible utility (or greatest possible happiness) for the greatest number of people." [7]

The action that maximises the well being and minimises the suffering of humans is the one most ethically correct.

This theory is based on the assumption that all human beings strive for maximising the utility of their actions. The action that brings as much well-being as possible to as many people as

possible is the morally right action. The intention of the action is not significant in utilitarianism, rather the quality of the action is "determined solely on the basis of its outcome.", which puts utilitarianism under the class of consequentialist ethics. [7]

Different takes on what constitutes the utility of an action has been defined. [15] Universal Consequentialism considers the happiness of all entities, and the non-well-being of a single individual therefore drowns in the well-being of all entities. Therefore a man can be killed to save others, without this being deemed "wrong". As per Mill:

> ""The individual's concept of happiness can only be accepted insofar as it is not harmful to others."" [7]

3.2 Deontological Ethics

To contrast utilitarianism, deontological theories of ethics tell us what action we ought to take. Kant states, that the only unqualifiedly good is a good will, that is the will behind an action, and therefore not the consequence of the action. [16]
Deontology argues that actions are inherently good or bad. No matter how morally good their consequences, some choices are morally forbidden. [16]

Entities cannot make morally wrong choices, even if by making these choices the morally wrong choices of other entities will be minimised. For deontologists, conformity with a moral norm makes a choice right. Norms are to be obeyed by each entity, but norm-keepings are not to be maximised.

Therefore, the Right is said to have priority over the Good. If an action does not obey the Right, it may not be undertaken, no matter the Good that might come out of it. [16]

3.3 The Non-identity Problem

The non-identity problem describes a situation where the action of bringing a person with a so-called *flawed existence* into existence presents a dilemma when trying to determine whether the action is right or wrong. Bringing the person into existence usually brings a significant amount of good with it, but because the existence is flawed, the existence itself necessarily also brings some bad with it.

Three intuitions are at stake in the nonidentity problem:

- The first intuition is the person-affecting, or person-based, intuition itself, that an act can only be wrong if that act makes things worse for some existing or future person.
- The second intuition is the intuition is that if the act gives a person an existence that is unavoidably flawed, that is the person would not exist if the act had not taken place, then the act does not make things worse for that person, because the person would otherwise not have existed.
- The third intuition is that some existence-inducing act are wrong, even if they do not make things worse for either the person that they bring into existence, and therefore make suffer, or any other future or existing person.

Chapter 4

Discussion of Ethical Views and their Application on Autonomous Vehicles

Using the definitions from the previous section, I will in this section analyse the ethical dilemmas that might occur with autonomous vehicles, and set these definitions up against each-other. On the first look, autonomous vehicles being part of our everyday life might not present many ethical dilemmas, but as Moor says:

> "Although a problem in computer ethics may seem clear initially, a little reflection reveals a conceptual muddle. What is needed in such cases is an analysis that provides a coherent conceptual framework within which to formulate a policy for action." [10]

Defining and designing algorithms that will guide Autonomous Vehicles (AVs) when put in moral dilemmas is very challenging. So-called moral algorithms should accomplish and be evaluated against potentially incompatible objectives [6]:

- **Be consistent:** Take predictable measures in critical situations.
- **Not cause public outrage:** Act according to the moral principles of society.
- **Not discourage buyers:** Make the public comfortable buying and using AVs thereby allowing the widespread use of AVs.

Accepting autonomous vehicles into society presents some interesting ethical dilemmas. Because completely autonomous vehicles have only been on the road for a few years, no policy for action has been formulated regarding these vehicles. An example of an ethical dilemma would be yourself sitting in your autonomous car, going to work. In front of you is a flatbed truck with a heavy load tied to the truck. On your immediate right is a motorcyclist and on your left is an SUV with a family of four in it. Your autonomous car is keeping a safe distance to the truck, so that if it breaks suddenly, you could stop.

Then the cable on the truck snaps, and the heavy load falls onto the road immediately in front of you. Your car could not predict this, it merely sees a new obstacle in front of you. Avoiding it without injuring you is impossible. Your car could choose to swerve, hitting either the motorcyclist or the SUV. Hitting the SUV might injure both you and the family inside, but hitting the motorcyclist would not injure you, instead it would injure the motorcyclist severely.

I will use this ethical dilemma for the following discussion of how to view the ethical issues involved.

These situations *can* occur, even though they may happen extremely rarely. The mere possibility of this happening forces us to reason about the decisions autonomous vehicles must make, and who should be made responsible for any accidents that occur, if it is even possible to hold anyone responsible.

4.1 An Utilitarian View

According to utilitarianists, no matter the ethical dilemma involved in possible critical situations involving autonomous cars, simply just having autonomous cars will always be ethically sound, since autonomous cars will save lives. Provided that an autonomous vehicle must choose to hit a person, the end result will still drastically lower the current amount of people killed in accidents, and therefore maximise the well-being of humans.

Provided the ethical dilemma described, a car will have to choose whether to injure a person. The utilitarianist view gives that the action increasing utility - saving the highest amount of people - is the right action to take.

Utilitarianism does not provide a sound answer to who is to blame for the possible death of the person hit. An extreme view is that no-one is to blame, since the action was not wrong. By saving the minivan with the family of four, and therefore either sacrificing the passenger of the car, the maximum utility has been reached.

Why not hit the motorcyclist? *Ducking harm*, the act of transferring harm to another person by stepping out of the dangerous situation, when the person is left behind, is generally considered more right, than explicitly sacrificing the other person by placing him/her in harms way instead of oneself [6]. By hitting the motorcyclist, harm is explicitly transferred to this person, thereby being a worse act. On the other hand, this would seem to change the act of a car avoiding hitting five people and instead hitting one person on the sidewalk, into being a bad act. Suddenly the single person has had harm transferred explicitly to him from the passengers of the car, which would be deemed worse than simply "doing nothing" and hitting the five people. This conflict with the utilitarian view, and a new ethical dilemma is thereby introduced.

4.2 A Deontologist View

According to deontologicalists, an autonomous vehicle should be programmed to never hit any people. Deliberately choosing to hitting another person is a morally wrong act, no matter the outcome. Letting the passengers of the car crash in order to not explicitly target people, would be the right action.

On the other hand, if hitting a person is seen as an accident that naturally occurs when driving, then it is unavoidable, and would happen either way, no matter whether the autonomous car is involved or not. With this view, the action of hitting a person can not be bad, since it would happen regardless.

According to this logic there *is* no victim, since the person hit is just one of many accidents that happen every year, and there is therefore no-one to blame.

4.3 The Non-identity Problem

The non-identity problem comes into play when discussing the issue of bringing autonomous vehicles into the world that possibly have to choose to hit, and possibly kill, people. The existence of this car would be flawed, but choosing to bring a non-autonomous car into this world would not result in the same existence.

Furthermore, not bringing the car into this world would make life worse for future or existing people, since these people would be involved in possibly lethal accidents they otherwise would not have.

It could therefore be argued that bringing autonomous vehicles that have to hit, and possibly kill, people into this world is the right thing to do. With this view, no-one is to blame, since the situation is flawed. Not bringing possibly killing autonomous cars into the world would make current and future people worse of, and is therefore not an alternative. Blame can therefore not be placed on anyone in particular, since the ethical issues involved prohibit this.

Furthermore, giving designers and auto-manufacturers the responsibility of determining whether to hit or not hit a person is unrealistic, due to the nature of these situations. A designer or programmer cannot design an algorithm that will be satisfying in every situation, due to the complicated nature of these situations. Choosing the same course of action will not always make sense. Responsibility for the actions taken by autonomous vehicles can therefore not be placed here, either.

4.4 Are We Ready for Utilitarianist Cars?

Looking at ethical dilemmas and trying to answer them using ethical theories might not provide a satisfactory answer to how the requirements of the algorithms of autonomous vehicles should be fulfilled, as demonstrated by researchers Jean-François Bonnefon and Azim Shariff and Iyad Rahwan, when looking at the presenting the idea of an utilitarian car for 300 people. [5] Their research provided rather interesting results. Most people thought that an autonomous vehicle should make the utilitarian decision, and always harm as few people as possible. Furthermore, they also thought that the car should injure the passenger of the car instead of a random passer-by while saving more people from being hit, satisfying the deontological view that no bad action can be taken by the car.

Surprisingly, the people asked then said they would not purchase such car. An algorithm making actions according to the opinion of the people is therefore not satisfactory either, illustrating that a single answer to the presented ethical dilemmas is impossible to find.

Automated cars promise great benefits and unforeseen effects that are hard to predict. Autonomous vehicles seem to be coming either way, with progress happening in big increments yearly. Change is unavoidable and not a bad thing, but new harms should be avoided where possible. This is the role of ethics: it can help us get a nicer future life, or it could wreck us if we do not watch out.

As Moor fittingly said:

> "On my view, computer ethics is a dynamic and complex field of study which considers the relationships among facts, conceptualizations, policies and values with regard to constantly changing computer technology. Computer ethics is not a fixed set of rules which one shellacs and hangs on the wall."

The ethical dilemmas presented by the introduction of autonomous vehicles do not allow for clear-cut answers, telling us how to design algorithms controlling the vehicles. It is impossible to give a clear preemptive answer as to who should take the blame for actions autonomous vehicles take.

Chapter 5

Conclusion

In this essay I have discussed the ethical implications of autonomous vehicles transporting passengers from A to B. I have described the technology behind recent research projects developing autonomous cars, and shown an example of some recent projects, namely the Google Self-Driving Car Project. Furthermore, I have described motivations for developing autonomous cars, and the benefits they could bring.

I have described one of many possible ethical dilemmas, and described possible way to view this ethical dilemma, depending on which ethical view is used, more precisely the utilitarianist and deontologist views. Furthermore, I have described the non-identity problem and how the topic of autonomous cars come into play with this problem.

I shown that it is not possible to determine who is to blame for autonomous cars given either an utilitarianist or deontologist view because the problems involved are not clear-cut. Furthermore I have detailed that the opinion of people regarding autonomous vehicles of other people differs from their own wishes for their own vehicles, and that this fact in combination with the issues presented earlier, show that it is difficult to find a clear-cut answer to these ethical issues.

Finally I have concluded that the introduction of autonomous vehicles into society presents ethical dilemmas that we cannot imagine yet, and that this fact makes it very difficult to say what is right or wrong when it comes to developing the algorithms that drive the cars.

Bibliography

[1] National Highway Traffic Safety Administration. Distracted driving 2014. 2016.

[2] National Highway Traffic Safety Administration. Distracted driving 2014. 2016.

[3] Amara D. Angelica. Google's self-driving car gathers nearly 1 gb/sec. `http://www.kurzweilai.net/googles-self-driving-car-gathers-nearly-1-gbsec`, 2013. [Online; accessed 26-May-2016].

[4] A. J. Baime. Can an autonomous audi beat a pro-driver on a race track? `http://www.roadandtrack.com/car-culture/a27200`, 2015. [Online; accessed 26-May-2016].

[5] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars? *CoRR*, abs/1510.03346, 2015.

[6] Roy A. Sorensen Christopher Boorse. Ducking harm. *The Journal of Philosophy*, 85(3):115–134, 1988.

[7] A.B. Holm. *Philosophy of Science: An Introduction for Future Knowledge Workers*. International Specialized Book Service Incorporated, 2013.

[8] Google Inc. Google Self-Driving Car Project. `https://www.google.com/selfdrivingcar`, 2015. [Online; accessed 26-May-2016].

[9] J. Mizzoni. *Ethics: The Basics*. John Wiley & Sons, 2009.

[10] James H Moor. What is computer ethics?*. *Metaphilosophy*, 16(4):266–275, 1985.

[11] Joann Muller. The road to self-driving cars: A timeline. `http://www.forbes.com/sites/joannmuller/2015/10/15/the-road-to-self-driving-cars-a-timeline`, 2015. [Online; accessed 26-May-2016].

[12] Yuki Sugiyama, Minoru Fukui, Macoto Kikuchi, Katsuya Hasebe, Akihiro Nakayama, Katsuhiro Nishinari, Shin ichi Tadaki, and Satoshi Yukawa. Traffic jams without bottlenecks—experimental evidence for the physical mechanism of the formation of a jam. *New Journal of Physics*, 10(3):033001, 2008.

[13] H.T. Tavani. *Ethics and Technology: Controversies, Questions, and Strategies for Ethical Computing*. Wiley, 2011.

[14] Stanford University. The definition of morality. `http://plato.stanford.edu/entries/morality-definition/`, 2002 (rev. 2016). [Online; accessed 26-May-2016].

[15] Stanford University. Consequentialism. `http://plato.stanford.edu/entries/consequentialism`, 2003 (rev. 2015). [Online; accessed 26-May-2016].

[16] Stanford University. Deontological ethics. `http://plato.stanford.edu/entries/ethics-deontological`, 2007 (rev. 2012). [Online; accessed 26-May-2016].

[17] Wikipedia. Darpa grand challenge. `https://en.wikipedia.org/wiki/DARPA_Grand_Challenge`, 2016. [Online; accessed 26-May-2016].

[18] Wikipedia. History of autonomous cars. `https://en.wikipedia.org/wiki/History_of_autonomous_cars`, 2016. [Online; accessed 26-May-2016].

[19] Wikipedia. The jetsons — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/wiki/The_Jetsons`, 2016. [Online; accessed 26-May-2016].