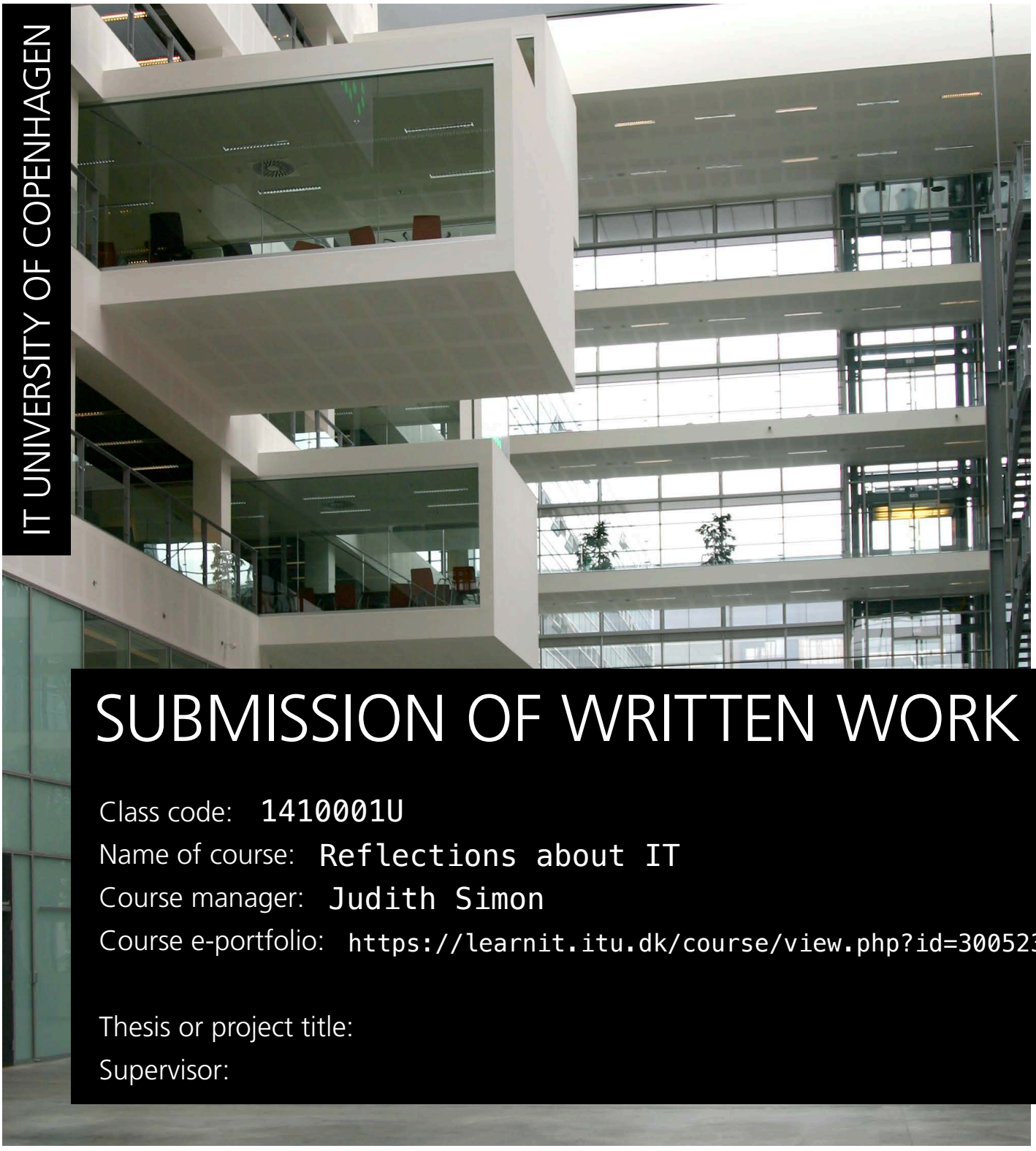
IT UNIVERSITY OF COPENHAGEN



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

Should the Designer of Your Car Determine Who It Kills? Ethics and Autonomous Vehicles

Anders Fischer-Nielsen

May 27th 2016

Contents

Chapter 1

Introduction

> "As we are about to endow millions of vehicles with autonomy, taking algorithmic morality seriously has never been more urgent. [1]"

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

The idea of inventing an autonomous 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[1], has been a trope of science-fiction for decades. 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 Stanfords in the DARPA Grand Challenge [8] and the recent undertakings of most car manufacturers, such as BMW, Audi, Toyota, VW etc. [6]

Certain ethical issues arise with autonomous vehicles, however. Since autonomous vehicles are programmed to handle preferably every possible scenario on the road in advance, autonomous vehicles present ethical dilemmas that human drivers in non-autonomous vehicles do not.

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 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 main question I will try to answer in this essay is; "Should the designer or programmer of your autonomous car determine who the car will hit in an emergency situation?" by looking into what ethical issues this question presents, and what theories can help answer this question.

[1] The popular TV show, *The Jetsons* is one of the more well-known examples of having autonomous vehicles. [9]

Chapter 2

Autonomous Cars

Experiments attempting to automate vehicles, mainly cars, have been made since the 1920's with varying degrees of automation and success.

This essay will focus on *autonomous* cars, not just 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. *Autonomous* implies that cars react to their environment independently, that is they cannot depend on unnatural artefacts in their surrounding environment in order to drive properly.

Modern research has been focused on the latter, since it is unrealistic to add artificial hints on every road in the world. Rather, research has been focused on making autonomous cars adapt to environments with uncertainties, so that the same autonomous car can drive in the inner city and on mountain roads. I will therefore not look at automatic vehicles in this essay, but instead autonomous vehicles.

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, enabling them to sense the world around them in great detail.

2.1 Google's Autonomous Car Project

Google has been researching autonomous cars since 2009 [3]. Over the years the project has been ongoing, the cars have developed a detailed view of their surroundings. In addition to the computer vision research the company 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.

2.2 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, do not always behave logically such as when angry or tired. Furthermore, human drivers sometimes drink while under the influence of drugs, which severely inhibits the driver's ability to react in time to avoid collisions.

Labor costs of human drivers can also be eliminated, traffic jams might be reduced[1], and vehicles might be able to park more efficiently, drive faster, and occupants of the car might have time to be more work while driving.

In short, a computer driving your car will probably be a better driver 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 will never top that. Your computer will also never get tired or drunk and make mistakes because of that.

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 choice possible. An autonomous car would register the child running across the road before a human driver ever would, 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". 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 one, who should it choose? Who should be to blame for any collisions or pay for any harm done to people or structures?

Answering these questions requires us to look at other ethical arguments described in the following sections.

[1] So-called shockwave traffic-jams caused by human error, are researched in [7]

Chapter 3

General Ethics and Computer Ethics

Computer ethics, and generally ethics of technology, evolves with the corresponding technologies of the time. This is one of the main problems of computer ethics, and also why it is important to discuss. Moor says:

"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."[1]

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. [4]

"Every advantage in the past is judged in the light of the final issue. — Demosthenes"

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

"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. (Beck Holm: 207)"

3.2 Deontological Ethics

"Kant's moral law. The point of this law is that we must always act in such a way that we can accept the consequences that would occur if all others were to act in the same way. (Beck Holm:

[1] Quoted from [5]

213)" Deontology argues that actions are inherently good or bad. If the action performed is bad, then the entire action, no matter the outcome is bad. An action should adhere to certain moral rules, and if it does not, then it must be bad.

3.3 The Non-identity Problem

The nonidentity problem describes a situation where bringing a person with a so-called flawed existence into existence has to be determined as being good or bad. Bringing the person into existence brings a usually significant amount of good with it, but since the existence is flawed, necessarily also some bad with it. Three intuitions are at stake in the nonidentity problem. The first 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 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

Ethics of Autonomous Vehicles

Initially, autonomous vehicles 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." [5]

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 [2]:

- **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 illustrates where debate can be held regarding autonomous vehicles is the example where you are 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 possibility of this happening forces us to reason about the validity of having autonomous vehicles, and who should be made responsible for any accidents that occur, if we can even hold anyone responsible.

According to utilitarianists, no matter the ethical issues involved in possible critical situations involving autonomous cars, having autonomous cars will always be ethically sound, since autonomous cars, even with many critical situations, will save more lives overall. Even if the act of selecting individuals to hit if it is unavoidable, the end result - drastically lowering the current amount of people killed in accidents - will maximise the well-being of humans. Therefore developing autonomous cars, as long as this saves more lives, will be a good thing. In the end, an autonomous driver will outperform a human driver, and therefore save more lives. No matter the implications, having autonomous drivers will be better.

If viewing possible ethical dilemmas involving autonomous cars in accidents as an utilitarianist, placing the responsibility either of who should be held responsible for accidents or determining whether to hit the motorcyclist, injure yourself or hit the SUV, would simply mean analysing the consequences of the hitting or blaming all possible actors, and choosing the one that brings the least human harm.

According to deontologicalists, the car should never choose to hit anyone, since deliberately hitting another person is wrong, and the entire act would therefore be wrong. Though, if the hitting of a person is seen as an accident then it is unavoidable, and would happen either way. Therefore it is hard to argue that the action is bad, since it is unavoidable.

Another argument for allowing the selection of a person to hit with the car, would be that since accidents happen, and that some number of people every year would be hit, whether it's this person or another person deos not matter. The person hit is just hit one of many accidents that happen every year, and the person is therefore not a victim of specific targeting, and therefore not a victim at all.

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 better than explicitly sacrificing the other person, by placing him/her in harms way instead of one self [1]. This would seem to change the act of a car avoiding hitting five people, and instead hitting one person into being bad. Now the single person has been explicitly sacrificed, thereby transferring harm from the passengers of the car, which would be deemed worse than simply "doing nothing" and hitting the five people.

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. 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.

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. Choosing to always injure the driver of the car or always hitting the oldest person registered by the car is not really an option, since the situations in which these choices have to be made change.

[1] [2]

Chapter 5

Discussion

"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. - Moor"

Chapter 6

Conclusion

Bibliography

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

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

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

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

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

[6] Joann Muller. The road to self-driving cars: A timeline, 2015. [Online; accessed 26-May-2016].

[7] 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.

[8] Wikipedia. Darpa grand challenge, 2016. [Online; accessed 26-May-2016].

[9] Wikipedia. The jetsons — Wikipedia, the free encyclopedia, 2016. [Online; accessed 26-May-2016].