UNIVERSITY OF TURKU
Turku School of Economics

Ethical autonomous vehicles: developing a publicly acceptable ethical setting.

Information Systems Science
Master's thesis
Turku School of Economics

Author:
Aleksi Saarinen

Supervisor:
Ph.D. Jani Koskinen

15.3.2024
Turku

Master's thesis

Subject: Information Systems Science

Author: Aleksi Saarinen

Title: Ethical autonomous vehicles: developing a publicly acceptable ethical setting

Supervisor: Ph.D. Jani Koskinen

Number of pages: 96 pages + appendices 10 pages

Date: 15.3.2024

Fully autonomous vehicles (AV) are likely to be a reality within a decade, bringing about numerous benefits from easing congestion to enabling prolonged independence for elderly. But with all these benefits, there are numerous challenges to be solved before this can be a reality. Some of them are purely technical, but some are fundamental ethical issues in the relationship between man and machine. Lately, especially the crash algorithms and harm distribution have been in the centre of AV ethics research. Meaning, how should an autonomous system distribute harm in situations where it can't be avoided? There are numerous variables to answer, from permitting active intervention to what party to prioritize. This thesis set out to explore different decision-making models and ethical theories that could offer a solution for morally acceptable autonomous vehicles. Objective is to gain an understanding of the general values that an AV ethical setting must possess to gain society's trust. This thesis is a qualitative and explorative study, presenting methods of simulating moral dilemmas faced by AVs, as well as distinct moral variables that are present in real-life scenarios. The goal is to see how laypeople, or future users, would prefer an AV to solve these moral dilemmas. The data is gathered from five semi-structured interviews, inspired by the famous trolley problem. The study also utilizes a discourse ethical group interview as a tool to achieve a majority consensus on the scenarios, which are used to propose an ethical policy for a publicly acceptable AV.

The results indicate that a publicly acceptable ethical setting comes down to two rational principles. First, the AV must be able to confine harm according to pre-determined rules, preferably to the responsible party. This is the user by default, but it must have the capability to transfer harm to the party responsible of creating the risky situation. Secondly, an AV must be predictable, both in terms of reactions and actions. It means that an AV must be able to take legal responsibility into account, which will lead to predictable reactions to lawful or unlawful behaviours of other traffic participants. It also means utilizing a rule-based decision-making system, instead of outcome-based optimization. This leads to better system transparency, as other traffic participants can be aware of the AVs decision making process in advance. The study also revealed fierce opposition against prioritization based on qualitative factors such as age. The results present a major divergence from existing research, where harm minimization has been the primary preference. The study suggests that the AV ethics research and development shift away from utilitarian harm minimization and place more focus on harm confinement and predictability. Also, the results suggest that qualitative prioritization may be preferred only in studies where participants do not need to defend this immoral position. The study also provides a methodological contribution by validating discourse ethics as a tool to form a consensus on moral preferences, perhaps also on a wider scale.

Key words: Autonomous vehicle, AI ethics, trolley problem

Pro gradu -tutkielma

Oppiaine: Tietojärjestelmätiede

Tekijä: Aleksi Saarinen

Ohjaaja: FT Jani Koskinen

Sivumäärä: 96 sivua + liitteet 10 sivua

Päivämäärä: 15.3.2024

Täysin autonomiset ajoneuvot (AA) ovat todennäköisesti todellisuutta tämän vuosikymmenen kuluessa, tuoden mukanaan valtavasti hyötyjä ruuhkaantumisen helpottamisesta vanhusten lisääntyneeseen itsenäisyyteen. Mutta hyötyjen mukana tulee myös lukuisia haasteita, jotka tulee ratkaista ennen kaupallistamista. Osa näistä on täysin teknisiä, mutta osa perustavamman laatuisia ongelmia ihmisen ja koneen keskinäisessä suhteessa. Viime aikoina AA-etiikan tutkimuksen pääpaino on ollut erilaisissa törmäysalgoritmeissa ja siinä, miten harmia jaetaan osapuolten kesken. Miten autonomisen ajoneuvon tulisi jakaa harmia tilanteissa, jossa siltä ei voida välttyä? Vastattavana on kysymyksiä aina aktiivisen intervention sallimisetta siihen, mitä osapuolta lähtökohtaisesti priorisoidaan. Tämä tutkimus kartoittaa erilaisia päätöksentekomalleja ja eettisiä teorioita, jotka voisivat tarjota ratkaisun moraalisesti hyväksyttävästä AA:sta. Tavoite on ymmärtää, minkälaisia perusominaisuuksia järjestelmän tulisi omata ollakseen yhteiskunnallisesti hyväksyttävä. Tämä on kvalitatiivinen ja exploratiivinen tutkimus, joka esittää erilaisia metodeja AA:n kohtaamien moraalisten ristiriitojen mallintamiseksi. Tavoite on selvittää, miten tavalliset ihmiset, eli tulevat käyttäjät, preferoivat AA:n ratkaisevan moraalisia ristiriitoja. Nämä preferenssit kerätään viidestä puolistrukturoidusta haastattelusta, joiden inspiraationa toimii kuuluisa ”trolley problem” – ajatuskoe. Tutkimus myös käyttää diskurssieettistä työpajaa työkaluna, jonka tarkoituksena on saavuttaa enemmistökonsensus oikeista ratkaisuista. Tästä konsensuksesta muodostetaan lopuksi etiikkapolitiikka ja eettinen asetus yhteiskunnallisesti hyväksytylle AA:lle.

Tulokset osoittavat, että julkisesti hyväksytty eettinen asetus rakentuu pääasiassa kahden periaatteen varaan. AA:n tulee kyetä rajaamaan harmi vastuussa olevaan osapuoleen ennalta määriteltyjen sääntöjen mukaisesti. Lähtökohtaisesti käyttäjä on vastuussa, mutta AA:n on myös kyettävä siirtämään harmia laillisesti vastuussa olevalle osapuolelle. AA:n tulee myös olla ennustettava, sekä reaktioiltaan että toiminnaltaan. Tämä tarkoittaa sitä, että AA:n on kyettävä ottamaan laillinen vastuu huomioon, joka johtaa ennustettaviin reaktioihin vastapuolen käytöksen mukaan. Se vaatii myös sääntöperusteisen päätöksentekomallin käyttöönottoa, tulosperusteisen optimoinnin sijaan. Tämä johtaa parempaan järjestelmän läpinäkyvyyteen, kun muut liikenneosapuolet ovat tietoisia AA:n päätöksentekologiikasta etukäteen. Tutkimus myös osoittaa vahvaa vastustusta laadullisiin tekijöihin, kuten esimerkiksi ikään, perustuvaa priorisaatiota kohtaan. Tulokset edustavat merkittävää poikkeamaa aikaisemmasta tutkimuksesta, jossa harmin minimointi on ollut pääasiallinen preferenssi. Tutkimus ehdottaa, että AA-etiikan tutkimus ja kehitys siirtyisi utilitarianistisesta harmin minimoinnista kohti harmin rajausta ja ennustettavuutta. Tulokset myös implikoivat, että laadullisiin tekijöihin perustuvaa priorisointia preferoidaan lähtökohtaisesti vain tutkimusasetelmissa, joissa vastaajien ei tarvitse perustella mielipiteitään. Tähän liittyen tutkimus tekee metodologisen kontribuution vahvistamalla diskurssietiikan sopivuuden työkaluna, jolla voidaan saavuttaa konsensus sensitiivisissä ja subjektiivisissä aiheissa laajemmassakin mittakaavassa.

Avainsanat: Autonominen liikenne, tekoälyetiikka, trolley problem

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 An autonomous vehicle

Imagine yourself in the future, travelling in an autonomous vehicle on your way home from work. You are using the time that used to be spent in the monotonous task of driving to finish up your work for the day. From the window you see a man on the sidewalk, waiting for the pedestrian light to turn green. Suddenly, two children run the red lights to the front of your vehicle. The car's computer must now make a split-second decision: will it swerve left to oncoming traffic to avoid hitting the children, stay on its course and try to brake, or swerve right toward the single bystander. There is a plethora of moral dilemmas embedded in this scenario, with no definitive answer.

In the case of the autonomous vehicle, the ethical setting that the AI uses to come to the best solution for this scenario is pre-programmed, which poses one of the biggest obstacles in AV mass market implementation (Keeling, 2018). With so many options, what should the car do? Should it follow Santoni de Sio (2017), who states that a bystander who didn't contribute to creating the risky situation has the right to be prioritized even though it results in two causalities instead of just one? Or should it maybe follow a utilitarian approach and try to minimize the overall harm by prioritizing the two children? Or should the vehicle prioritize passenger safety in all scenarios? What about the fact that children disobeyed traffic rules, does that mean they should bear more of the harm? Figuring out an approach to distribute risk that corresponds best to public moral preferences is imperative, as public trust is a key factor in social acceptance of any technology (Geisslinger et al. 2021).

An autonomous vehicle described in this thought experiment does not exist yet, but manufacturers are nearing on achieving full autonomy in the coming decade. The level of autonomy can be categorized from zero to five, from fully manual operation to full autonomy, which can operate at any conditions without human presence or interference. (SAE International, 2021). When achieved, full autonomy might be one of the most disruptive technologies of the next decades, offering significant benefits both on societal and individual level.

On a societal level, AVs are forecasted to dramatically decrease traffic deaths that are totaling 1,25 million a year globally (WHO, 2018). AVs can decrease accidents by up to 90%, given that their share of the overall fleet is high enough. Autonomy can also

decrease congestion and increase traffic efficiency, as connected autonomous vehicles can cooperate at the levels that humans are unable or unwilling to. This also has a positive environmental impact, as increased efficiency translates to better energy usage. On an individual level, autonomy increases productivity by enabling people to use their time to more productive tasks such as in the scenario in the introduction. Another less-mentioned benefit is the increased independence and inclusion for those who can't access private vehicles due to age or disability. (Bergmann et al. 2018.) With so many benefits, there is a strong incentive to get AVs into mass markets. But along with these benefits, there are a lot of ethical issues that must be answered before mass-market implementation is possible.

1.2 Moral tradeoff between driver protection and utilitarianism

The main ethical problem with autonomous vehicles is that for first time in history, an automated system must be able to distribute harm in a morally acceptable way in scenarios where it is inevitable (Awad et al. 2018). This thesis will focus only on cases that have several options, but all of them result in life-threatening harm for at least one participant in the situation. Expanding the view on how possible injuries and their probabilities that are mentioned in some studies is considered out of scope for this research. According to relevant literature, there are two main concerns surrounding the moral settings of autonomous vehicles. The first one is the relationship between driver and all the other traffic participants (see Kauppinen, 2021; Nyholm, 2018a; Santoni de Sio, 2017). In its simplest form, this can be understood as a scale. On the other end, there is driver protection in all possible scenarios, and driver sacrifice as default in the other. In the middle, there is also the option to minimize the overall harm without considering participant roles. The question is where should an autonomous vehicle position on this scale? According to research by Bonnefon et al. (2016), there seems to be a paradox in what people want and what they are willing to use. Their results on the topic imply that when asked what kind of autonomous vehicles people want to see in traffic, the answer seems to tip toward the utilitarian, harm-minimizing end of the scale. But when these same participants were asked what kind of AV they would use personally, answers were tipped more for driver protection.

The second problem is that when we arrive at a conclusion that what the preferred ethical framework is, should it be universal and regulated, or customizable? (see Contissa et al. 2017; Gogoll & Müller, 2017) There are many views on the topic among researchers, with Gogoll & Müller (2017) being the most vocal about how a mandatory ethical setting is the only morally justified solution to avoid a moral hazard in choosing your ethical settings. Then there are Contissa et al. (2017) who argue that AV's increase traffic safety simply by being in traffic, and we should allow customizable ethical settings to make them more appealing to users, thus accelerating their adoption.

1.3 Framework for AV ethics

In previous literature, ethical problems faced by AVs have mostly been described as trolley problem dilemmas (see Awad et al. 2018; Kopecky et al. 2023). These are psychological thought experiments with different bad options and outcomes, where the agent must choose an action and make a moral trade-off. The name "trolley problem" refers to the original problem presented by Philippa Foot in 1963. This approach has been criticized for being too black and white, as it allows only the specific options with specific outcomes, like sacrificing one passenger and saving two pedestrians, or vice versa (Nyholm, 2018). Excellent motivation for its use as a base of discussion has been made by Bonnefton et al. (2019), who point out that while the trolley problem doesn't lend itself to individual accidents, it's an excellent framework for an industrial point of view. They continue to explain that if all the results of AV accidents are aggregated together, it starts to resemble a traditional trolley problem. Numbers may reveal that for example, for every driver killed, 2,1 pedestrians have been saved on average. This essentially means that the AV ethical problems are a statistical version of the traditional "A or B"- trolley problem. (Bonnefton et al. 2019.) In this thesis, the trolley problem is used as a tool to simulate AV moral dilemmas. The goal is not to solve the trolley problem, but to use it to discuss solutions that could apply also in the real-world scenarios.

According to an article by Geisslinger et al. (2021), there seem to be two main approaches on how trolley cases are solved, outside of simply prioritizing the passenger. These are the rule-based system and the outcome-based system. The rule-based system is based on Kantian machine ethics and revolves around base rules that can't be broken. One could be "don't kill the innocent", which then limits the AV's options to active

participants in the situation and forbids, for example, hitting a bystander to save the driver. The outcome-based system on the other hand doesn't take pre-determined rules on how it should prioritize traffic participants but simply seeks to achieve the least amount of total harm caused. This thesis introduces a range of ethical theories that an ethical setting could be based on, along with moral variables that an AV must consider in its decision making.

1.4 The research questions.

This thesis tries to find what the simplified public preference is for a decision-making framework in autonomous vehicles. To achieve commercial success, the ethical settings must at least somewhat comply with how relevant stakeholders would solve and justify a morally loaded accident scenario (Awad et al. 2018). If the rules are formed at the top by lawmakers without any connection to public preference, this may decrease their adoption and thus slow down all the benefits that we could collectively get from autonomy, regardless of their ethical settings.

The research questions of this thesis are:

1. *How do laypersons prefer an AV to solve moral problems?*
2. *How these preferences could be formed into an ethical setting?*

The objective of the first question is to see how a layperson, a potential future user, would solve ethical problems relevant to AVs in crash scenarios, and justify the chosen option. It doesn't take opinions on technical details but rather tries to conceptualize how a crash algorithm should behave to be accepted publicly. Whether some setting or variable is technically feasible or not is not discussed in this thesis. There have been some studies on this matter, but they have been carried out mostly via quantitative methods, thus lacking any explanation of the moral standpoints taken by subjects. (Awad et al. 2018; Bonnefton et al. 2016.) The research is carried out by a series of modified trolley problems, where the participants must choose how they prefer different moral variables to be weighed against each other. The goal is to see if there are some larger, overarching themes on how AVs should make decisions, which could offer some guidance for the system requirements of a publicly acceptable ethical setting. The second question

addresses if the moral decisions preferred by laypeople are right in the eyes of traditional normative ethics, and which ethical theories best capture public preference. These requirements are also compared to the German Act on Autonomous Driving, currently the only official national legislation on AV ethics governance. The goal is to choose an appropriate theoretical base and decision-making model, and to propose a publicly accepted AV ethical setting.

This thesis tries to widen the scope from researchers to end-user opinions. It's clear that for successful commercialization, regulation can't be separate from public preference. The moral problems are modeled with trolley problem-style dilemmas, but the thesis doesn't assume that the moral decisions AVs face is as simple and deterministic that trolley cases present. This framework is used to draw attention to the ethical reasoning behind certain courses of action, using the simplified reality of trolley dilemmas. The research goal is to formulate a proposal for an ethical policy and associated system requirements, and then see which normative ethical theory could fit the purpose.

1.5 Structure of this thesis

This thesis is structured as follows. After the introduction, a literary review is conducted consisting of three parts that form the scientific framework for the thesis. Chapter 2 will motivate AV ethics research in general, as well as introduce the theoretical framework of AV decision making. Chapter 2 will first motivate why ethical dilemmas are an integral part of AV development and provide a psychological framework for understanding why these issues need to be at least discussed. After this, it will introduce the theoretical framework of AV decision making, consisting of the trolley problem, moral variables and AV decision making logic. Chapter 3 will offer different viewpoints on solving the problems presented in Chapter 2. It will go through different ethical theories that are discussed in AV- and machine ethics literature, and how each of them could be used to solve moral decision dilemmas. It will also go over the existing legislation on AV ethics, as well as general system requirements for an ethical setting. Chapter 4 will discuss regulation and how a consensus could be formed on this issue trough discourse ethics. The thesis doesn't take a stand on which framework is the most suitable one for AVs. Here arguments for both customisable and regulated AV ethics are presented from previous literature, as well as liability issues arising from them. Chapter 5 introduces the

chosen research methodologies, which were semi-structured individual and discourse ethical group interviews. It also discusses motivation behind using the selected methodologies in this specific domain, as well as the research procedure from design to data analysis. Chapter 6 presents the findings of the empirical research in terms of larger themes and tries to form a consensus on the selected topics. Chapter 7 will use this consensus to propose a general ethical policy for a publicly acceptable AV, as well as present the system requirements for the actual ethical setting. The contribution to AV ethics research as well as possible limitations are also presented in this chapter. Finally, chapter 8 concludes the findings of the entire thesis.

2 Dilemma situations faced by autonomous vehicles.

2.1 Why ethical problems matter for AV industry

There is a lot of debate about ethical dilemmas and their relevance. (see Bergmann et al. 2018; Krügel & Uhl, 2022) In the development community, ethics discussion is seen as a mere distraction from more pressing and tangible issues in technical development. Critics also argue that the AV industry should focus more on solving everyday traffic hazards, rather than on rare dilemma situations presented with trolley problems. (Gill, 2021.) This view is challenged by defendants of ethics research, such as Krügel & Uhl (2022). They argue that everyday traffic demands AVs to constantly weigh and distribute risk among traffic participants, so any reaction beyond staying on course and braking has ethical considerations.

Recent studies have also proven ethical dilemmas to have commercial relevance (Martinho et al. 2021; Nyholm, 2018a). One such study that offers justification for AV ethics research is carried out by Gill (2021), who observed ethical dilemmas to be the primary concern for people considering buying an autonomous vehicle. Gill concludes that this relative overweighting of rare dilemma events is due to the interplay of three psychological heuristics: risk aversion, affect heuristic, and probability neglect.

The prospect theory is the most famous work of Kahneman & Tversky (1979). In short, it explains how in human decision-making, risks are generally weighted more than equally sized benefits. In the AV ethics domain, this could mean that the lack of certainty around accident scenarios overthrows the safety or efficiency benefits. The second heuristic is called the affect heuristic, which states that when risk is present, people tend to weigh emotions much more than logic or statistics (Slovic et al. 2007). Finally, the probability neglect by Sunstein & Llewellyn (2003) finishes the psychological framework explaining why rare accidents are more relevant than one would think. Sunstein states that when an event has strong positive or negative consequences, its true statistical probability matters very little to people. This is why a lot of people are for example, afraid of air travel. Despite being the statistically safest transportation method, an airplane crash has a very low survival rate when one occurs.

The interplay between these three psychological factors provides an understanding of why manufacturers should take AV ethics seriously to achieve commercial success. Although it is true that solving dilemma situations is a minor and

rare part of AVs numerous functions, failing to address them completely may result in lower technology acceptance (Gill 2021). This view is supported also by Kopecky et al. (2023), who describe AV crashes to represent a “tail risk” for the AV industry. A tail risk is an event with low, but foreseeable occurrence, which separates it from a black swan. A good example of a tail risk and the importance of ethical issues is the failure of Uber’s AV program. The company decided to halt its multi-billion-dollar project after a fatal crash in 2018, which they failed to explain credibly to the public. (Topham, 2021.) It seems that failing to give a clear answer to pressing moral issues, however rare, could cause a “moral panic” in people. This might be especially prevalent with families due to the affect heuristic. Without clear communication on if the children are prioritized in an accident scenario, technology acceptance is unlikely. The uncertainty around AV ethics may affect the demand of early adopters, who are a key part of making any technology go mainstream. Ethical dilemmas are then acting as gatekeepers for achieving the safety benefits of having a large AV saturation in traffic. (Kopecky et al. 2023.)

2.2 Original Trolley Problem

In the recent literature, moral decision dilemmas are mostly represented using variations of the trolley problem, a famous psychological thought experiment, first presented by Philippa Foot in 1963 and later names as “the trolley problem” by Judith Thomson. The problem presents a runaway trolley that is on the collision course with 5 people, who are tied to the tracks. A bystander must decide if he pulls a lever that will reroute the trolley to another track, with one single individual standing on the track. The moral trade-off in the original trolley problem is whether or not it is justified to intervene, sacrificing the individual who was not originally in danger. (Thomson, 1984.) Besides describing the concept of the trolley problem, Foot’s article presents a psychological theory called the doctrine of double effect. According to it, causing harm is more justified when it is necessary to achieve a larger benefit. (Foot, 1967.) The doctrine of double effect is relevant for AV ethics research because it can explain the fundamental difference between programming an AV to kill certain traffic participants and programming a military robot to kill humans, an argument that has been made against researching AV crash ethics in the first place. For an autonomous weapon system, causing harm is the primary function. For an AV, the primary function is to make traffic safer and on some rare occasions, cause harm to certain traffic participants when it's unavoidable. (Nyholm, 2018b.)

Third important concept from Foot original work relevant to AVs is her description of positive and negative rights and duties, which can be used when modelling dilemma scenarios. A positive right means that person A is entitled to a certain action from person B, such as protection. Negative right means that A can do an action without B's interference. These rights have their corresponding duties, where a positive duty refers to A's duty to do an action for B, and a negative duty of A not doing an action towards B. (Foot, 1967.) As mentioned, these rights and duties offer a tool to model dynamics with AV and other traffic participants in traffic. For example, the relationship between a pedestrian crossing the street (A), and an autonomous vehicle (B) can be broken down to the following rights and duties:

	Positive	Negative
Right: Pedestrian	Right to be safe from the AV	Right to not get hit by the AV
Duty: AV	Duty of making sure the pedestrian is safe	Duty of not hitting the pedestrian

Table 1 - Rights and duties in a moral dilemma.

2.3 From theory to practice in AV domain

The primary feature that makes trolley problems interesting is the lack of a clear, unequivocal answer. It is highly dependent on the individuals' moral preferences. In the original problem, killing the five persons is wrong from a utilitarian standpoint, as it results in more causalities. But on the other hand, these five were already in danger. Rerouting the trolley results in only one death, but this individual was not originally in danger. So, rerouting the trolley makes the agent take an active role in killing the single individual, which is then wrong from a deontological standpoint. (Encyclopedia Britannica, 2023.) Although the original problem is quite inflexible as it only presents guaranteed outcomes and fixed alternatives, the core idea helps construct a wide variety of scenarios with different variables to prioritize. Because of this it's still widely used and discussed in AV as well as general ethical literature. Paulo (2023) goes as far as saying that trolley problems are so widespread, that they are considered the common language of moral decision-making research.

But as mentioned, it has its limitations. In the original problem, the two options

were determined by the fact that the trolley was fixed on tracks that could only go two ways. The outcome was also fixed in both options. In an AV crash, there are several possible trajectories, that come with different probabilities for harming each participant. Nyholm (2018) therefore argues that trolley problems are simply too abstract to offer any real resemblance to real-world traffic accidents, as there will be too many variables for an AV to make a calculated decision in a split-second. He also criticizes how the probabilities are taken for granted in trolley problems, as almost all AV crashes will be unique on some level.

On the other end, the most convincing argument for using the trolley problem as a framework to model moral dilemmas is given by Bonnefon et al. (2019), who introduce the idea of a *statistical trolley problem*. They argue that although deterministic and simplified scenarios are not optimal for modelling individual AV accidents, they offer an excellent way of looking at the ethical setting of the AV fleet in general. If all AV accidents would be aggregated together, one could for example see that for every passenger, 3,1 pedestrians have been killed on average. This means that the current AV ethical setting is choosing to protect the passenger in an accident scenario. This approach could be logical from an ethics policy perspective, as it's likely not relevant or even feasible to analyse each accident scenario individually. It's the average ethical setting that people want to know. The proposed framework for AV ethics proposed also by this thesis should then be understood as the preferred default response, not a fixed solution for each individual scenario. The focus on the average solution also might solve the problem of uncertainty and uniqueness, a critique by Nyholm (2018) presented earlier.

I presented arguments for and against of using of trolley problems to describe why and how current literature uses it to address moral decision dilemmas. But for the purpose of this thesis, the trolley problem is used merely as a tool to discuss AV ethics. The focus is to explore the reasoning behind what people view as morally acceptable in the AV context. The possible limitations of the trolley problem are therefore not detrimental, because the thesis doesn't even try to explain what an AV should do in a specific dilemma situation. It simply uses the trolley problem as a tool to discuss moral problems. As we need to model distinct variables in AV crash scenarios, the simplified reality of the trolley cases fits well for examining their internal hierarchy. Adopting this view is also supported by the literature. Martinho et al. (2021) write that the trolley problem should not be used in the AV ethics context to solve trolley problems themselves, but to develop ethical guidelines for their decision-making. This is supported by Paulo (2023), who points out

that even the original problem still hasn't been answered unanimously despite a decades-long academic debate. However, it has resulted in thorough research on how different ethical theories fit the dilemma. We should then use the problem not to seek answers for the dilemma itself, but as a tool to present complex scenarios in a simplified form to aid ethical dialogue. Especially for normative ethics focused on the morally correct action (Britannica Online Encyclopedia, 2023), the precise representation of reality is secondary. What matters more is that the variables present in simple trolley dilemmas are the same that AVs will face in real-life scenarios. (Paulo, 2023.)

2.4 Variables in AV moral decision dilemmas

To form an understanding of AV moral decision framework, variables that are present in accident scenarios should be specified. By variables, I mean distinct aspects or agents in the accident scenario, which the AV will in a hierarchical order, resulting in the desired outcome. As the thesis seeks to expand on current research, the chosen variables should be those most present in previous studies. The following five variables are recognised to be most prevalent in relevant literature:

1. Level of Intervention
2. Minimizing overall harm
3. Prioritization between traffic participants
4. Legal responsibility
5. Personal characteristics

Excluding the level of intervention, these variables are used in the most extensive quantitative study on the topic, MIT's Moral Machine Experiment, a gamified website that collected over 40 million answers to AV trolley scenarios (Awad et al. 2018). The results of the study are quoted in almost every other relevant study on the field, such as Bonnefton et al. (2016), who also used much of the same variables listed above. Let's now look at each one of them in more detail.

2.4.1 Level of intervention

Level of intervention means simply if an AV is allowed to make moral decisions in the first place. It means considering inaction as much an ethical choice as any other. A human driver bases his decision on intuition and sometimes on random reflexes, but an AV can produce a calculated decision even in a split-second scenario. If we don't take advantage of this capability because we don't want to address the hierarchy of other variables, that must be an ethical position itself. (Bergmann et al. 2018.) In the original trolley problem, one's views on intervention change the otherwise quite logical utilitarian answer of saving 5 people instead of one. This is because by pulling the lever, one takes an active role in killing the single person, whereas allowing the trolley to continue results in a crash that would have happened anyway. This difference between an active intervention and failing to act is a subtle, but relevant difference in a crash scenario. Also from a legal standpoint, there is a big difference between failing to save a person and actively subjecting him to harm. (Geisslinger et al. 2021.) If it is decided that AVs should not make ethical decisions, then the crash algorithm is simple: brake and stay on course. Only if active intervention is accepted, the other situation-specific variables become relevant. (Bergmann et al. 2018.) There are some arguments for a non-interventionist approach, stating that predictability is a key factor in traffic safety. If other participants know that an AV is programmed strictly to brake and stay on course, it would be easier to them to react. Especially if an AV would only be targeting a certain outcome without any hard-coded rules, the decisions can seem random or unpredictable. (Karnouskos, 2021.)

2.4.2 Minimizing overall harm

Perhaps the most decisive variable in an AV crash scenario is whether the AV should prioritize certain traffic participants or just minimize overall harm. Utilitarian ethics emphasize the maximization of overall well-being, and according to this ethical doctrine, an action is right if it results in the most amount of overall good, in this case collectively for the traffic participants. This means that an AV shouldn't make any distinction to whom the harm would be distributed, and simply calculate how it is minimized in each scenario. A utilitarian AV is a good example of an outcome-based decision-making system, lacking any distinct hard-coded rules on how to prioritize other moral variables than the amount of harm caused. (Faulhaber et al. 2019.) The problem is that people are found to be quite irrational when it comes to preferring harm minimization. A study by Bonnefton et al. (2016) observed that there is a concerning discrepancy in people's

opinions on utilitarian AVs. Objectively, people thought that minimizing overall harm was the correct way to program AVs, and that they would like to see other people buy them. But when the same subjects were asked what they would personally use, most of them favoured driver protection. The preference for utilitarianism from a third perspective is supported by the Moral Machine Experiment by Awad et al. (2018), where respondents strongly favoured "saving more lives" over other variables. The study didn't test them again as users, but the results nevertheless support the view that from an objective perspective, minimizing overall harm is preferred by most people. But from a commercial standpoint, subjective first-person opinions are the ones that make people buy these vehicles, so the results around this area are quite concerning for the industry, especially if a utilitarian ethical setting is demanded by regulators. In an absence of regulation some manufacturers could even create a new form of social discrimination by offering better driver protection and decision making for wealthier customers (Karnouskos, 2021).

2.4.3 Prioritization between traffic participants

The third variable to consider is the prioritization between different traffic participants. As discussed, the trolley cases are used as a simplified presentation of reality to allow for ethical discourse, not to solve the scenarios themselves. It is then enough to have two distinct traffic participants in this thesis: the AV user protected by the vehicle, and a pedestrian. The prioritization variable is linked to harm minimisation, as participant quantity should obliviously affect the decision. But in a vacuum and with one passenger and one pedestrian, which one would deserve more protection and why?

As discussed earlier, Bonnefton et al. (2016) suggest that people prefer buying an AV that prioritizes passengers, which incentivizes car manufacturers to have a more egoistic ethical setting. This is especially true if there's no universal regulation. Consider an example of two companies, with an otherwise identical AV product, but with different prioritization settings. Company A's algorithm splits the harm 50-50 between participants, and company B's favours the passenger in all cases. If there is no regulation, customer demand would obviously be higher for company B. (Bonnefon et al. 2019.) This will lead to more dangerous overall traffic, an effect opened more in chapter 4. People are also found to be willing to pay more for an AV with passenger prioritization, further complicating the market dynamics of a utilitarian AV in the absence of regulation (Liu & Liu, 2021). It is also an understandable view that no matter how well self-sacrifice is argued for in theory, people can't be demanded to sacrifice themselves in an accident

scenario by default (Kriebitz et al. 2022). Opting for self-preservation has been observed in several studies where the decision is made from a passenger's point of view (Bergmann et al. 2018; Bonnefton et al. 2016). Also, the industry has already taken a moral position in favour of the passengers as most automotive safety innovations from airbags to chassis design have been focusing on their safety (Gerdes & Thornton, 2016). This validates the concerns of Karnouskos (2021), who argued that manufacturers could be inclined to invest heavily on passenger prioritization if made possible.

Opposing views are presented by Bergmann et al. (2018), who favour pedestrian prioritization. Their main argument is that the whole idea of a side- or a crosswalk is to guarantee people some level of protection. An AV, at least in the beginning, is a novel technology. People must understand that using it may pose some risks. It would be unethical to ask for people using a sidewalk, that is meant for pedestrians, to share some of the harm in an accident. However, this argument is not supported by their own study, where most respondents did not see a sidewalk to be a relevant factor in how harm should be distributed. Although it must be noted that in this study, the subjects made the decisions in the first-person and with time constraints using a VR headset. So, the same answers from third-person observers in this study might differ from these results.

Luckily for constructing a universal ethics setting, there seems to be a middle ground. In a study by Karnouskos (2021), there were a high number of subjects willing to make compromise as AV passenger. Their survey results advocate for an algorithm that assigns moderate harm to all parties, instead of one being saved and one severely injured. This approach is also supported by numerous law cases of traffic accidents, where verdicts usually place some responsibility for the driver, even if the fault might be entirely on the pedestrian. This comes down to the fact that the car has the potential to cause severe harm to the pedestrian, but not vice versa. The driver thereby has an added level of responsibility, even if the pedestrian might have caused the accident. (Santoni de Sio, 2017.) But these cases have been with human drivers and human decisions, and in the case of an AV, the passenger can’t control the situation. It is then up to debate if the passenger can be assigned the same level of responsibility as a driver of a manual vehicle. It could be morally questionable to automatically assign harm to the passenger in a crash situation because a passenger whose AV crashed didn’t do anything different from a person whose AV didn’t. (Kauppinen, 2021.)

2.4.4 Legal responsibility

The fourth variable is legal responsibility, which is closely related to the relationship of different traffic participants. As the passenger in an autonomous vehicle is not in control of the situation, he can't obey or disobey any rules. The question is then, should the protection of the other traffic participants, such as manual cars, pedestrians, and cyclists vary depending on their own contribution to the accident scenario? Key research on this topic has been done by Kauppinen (2021). He is concerned that even though autonomy is guaranteed to save lives, some careful consideration is needed to ensure it saves the right lives. If people voluntarily take or create a risk in traffic, they shouldn't enjoy the same level of protection as someone who obeys traffic rules. He argues especially against the utilitarianist approach of minimizing total harm when someone has clearly placed himself in a risky situation voluntarily. (Kauppinen, 2021.) Going back to the duties and rights present in the original trolley problem, one could argue that Kauppinen sees that by breaking the traffic rules, an individual should be more exposed to harm from an AV.

Coca-Vila (2018) extends on this using current legislation on self-defence as an analogy. According to him, it is generally allowed to transfer harm to the party that is responsible for the dangerous situation, even if that party would be more vulnerable. So, an AV could be justified to transfer harm to a jaywalker, even though the passenger is more protected. A question also arises if an AV wouldn't consider legal responsibility, will it form a moral hazard for other traffic participants to act unlawfully? Meaning that they might start to act more carelessly if they know an AV prioritizes them regardless of their involvement. Opposing views to Kauppinen's ideas have been presented by Nyholm, (2018b). He argues that even if the passenger didn't have any way of affecting the course of events, he accepted the responsibility by getting in the AV. If an AV crashes due to a technical malfunction, the driver is responsible just as a dog owner is responsible for the damage caused by the dog.

2.4.5 Personal characteristics

In previous literature, discussion around personal characteristics has mostly revolved around whether children should be prioritized in an accident scenario. This is predicted to be the variable with the widest discrepancy between public opinion and legislation. The Moral Machine Experiment featured a scenario where pedestrian age was the moral trade-off, and all other things being equal, an overwhelming majority prioritized the children in this scenario. The authors of the experiment argue that even though it is

impossible for legislators to weigh one life higher than another, forcing equality by regulation may severely hinder the adoption of AVs. (Awad et al. 2018.) Their findings are supported by the results of Faulhaber et al. (2019), who saw a direct correlation between age and chances of being saved in a trolley dilemma test scenario.

It must be understood that if prioritizing children would be allowed, we must give the AV the capability to rank people in real time, which as a mere idea is deeply concerning. This view is supported by Kriebitz et al. (2022), who are concerned that a dysfunctional algorithm or maleficent data input could result in discrimination against minorities. Egalitarian ethics can also be used to prohibit any discrimination based on personal characteristics. Egalitarianism calls for equality and strictly prohibits any difference in treatment based on personal attributes. (Bergmann et al. 2018.) Although personal characteristics could be unfit as an actual variable in an ethical setting, I still find it valuable to be included in this thesis. This topic is discussed and studied extensively in existing research, with results strongly favouring prioritization of children (Awad et al. 2018; Bergmann et al. 2018). By finding out the preference and reasoning for it, we can see just how big of an outrage is awaiting legislators if this is simply forbidden. It should also be pointed out that although somehow deemed immoral in the AV context, prioritizing by age is standard practice in some fields. For example, medical resources are prioritized by "remaining quality-adjusted life-years", which basically means prioritizing young and healthy patients if it's not possible to treat everybody in need (Anderson & Anderson, 2011). So if the legal system already allows for prioritization in the medical context, this topic should at least be open for debate in the AV ethics domain.

2.5 Dimensions of moral decision making

Now that we have introduced the trolley problem as a basic theory, expanded it to the AV domain and introduced the moral variables, we can construct a framework for AV moral decision-making. There have been several studies on AV moral decision-making using modified trolley problems (Awad et al. 2018; J.-F. Bonnefon et al. 2016; Coca-Vila, 2018). As the primary objective of this thesis is to provide a deeper rationale for the findings of previous studies, it's best not to copy any specific study, but to implement best practices from the literature and see if a universal model can be constructed.

We'll use the main 5 variables from the Moral Machine Experiment by Awad et

al. (2018), but divide them into gatekeeper, main, and secondary variables. First, there is intervention as the gatekeeper. As discussed, only if active intervention is accepted in the first place, the other variables are addressed. After we decide that active intervention is indeed permitted, we can assess the main variables present in the scenario: minimising overall harm and prioritization between traffic participants. They are chosen as the main variables because they are used consistently in most empirical studies, so it's safe to assume that they are considered most relevant by most of the researchers. The secondary variables are legal responsibility and personal characteristics. They are more minor, situation-specific factors that will alter the initial hierarchy of the main variables if present in a scenario. For example, an AV could have a setting that prioritizes pedestrians by default, if they act lawfully. If they don't, the passenger is prioritized.

AVs are divided into three categories according to their preferred action type, an approach presented by Kopecky et al. (2023).

1. The altruist AV. Minimizes overall harm and prioritizes other traffic participants.
2. The egoist AV. Doesn't prioritize overall harm and prioritizes the passenger.
3. The conservative AV. Doesn't allow active intervention.

Now we have the variables and the different approaches to decide on their hierarchy. Next, we'll consider the dimensions of the moral decision itself. Excellent work has been done by Schäffner (2021), who recognizes three main dimensions which demonstrate how AVs decisions differ from ones taken by a human driver:

1. Type of decision
2. Decision perspective
3. Point in time.

First, let's consider the type of decision. A human driver's decision in a specific accident scenario is a combination of intuition and reflexes and is made by the driver alone. The rationale behind an AV decision is instead affected by multiple actors, such as public opinion, legislators, and manufacturers. The decision is therefore a collective one, especially if the ethical setting would be universal through regulation. The second dimension is the decision perspective. A human driver naturally makes the decision from a first-person perspective. The decision by an AV is instead made from a third-person

perspective by developers and indirectly by other stakeholders. The third dimension, point in time, is the reason why AV ethics discussion is needed in the first place. A human driver makes the decision in real-time, in a split second. But as outlined in the introduction of this thesis, the AV decision is pre-programmed. The decision is therefore made before the accident and with no time constraints. The decision is carried out in a split-second by the AV, but the rationale might have been years in the making. (Schäffner, 2021.)

2.6 Framework for AV moral decision-making

In this first chapter, we have discussed 1) why ethics matter in the first place 2) introduced the trolley problems as a theoretical base 3) expanded the trolley problem to the AV domain 4) identified the constructs of a moral decision taken by an AV as an answer to a trolley problem scenario. Now I present my simplified model for the moral decision process by an AV based on the best practices found in relevant literature. In a simplified representation, AV will make a decision in the following way:

1. Is active intervention allowed? If no, brake and stay on course.
2. If yes, what are the main variables present in the situation? What is their hierarchy based on the chosen action type?
3. Are there some secondary variables present? How do they affect the initial hierarchy?
4. An action based on the final hierarchy and available action type(s).

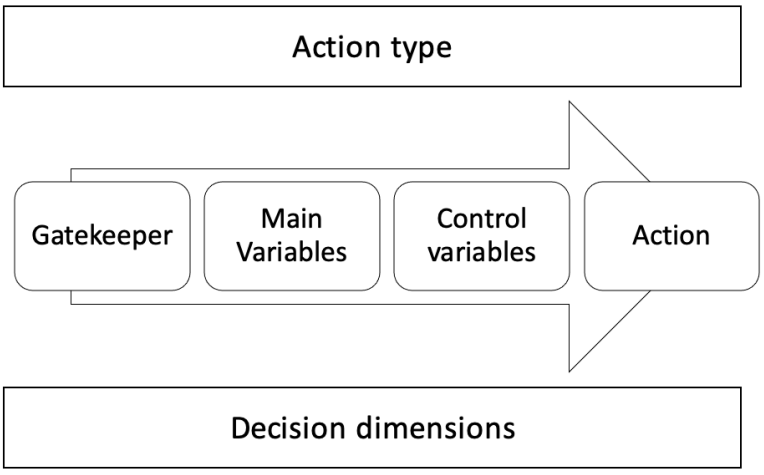


Figure 1 - A simplified moral decision framework.

The above figure presents the process in a visual format for better understanding. The decision-making process is in the middle, and in the background, there are the pre-determined action type and decision dimensions that give the process constraints. One might argue that the decision dimensions are not an active part of the decision-making process, but I disagree. Consider if for some reason we implement a self-learning approach for decision-making instead of a pre-programmed one. If the moral decisions were made purely by artificial intelligence without any pre-coded constraints, the decision dimensions would resemble more of those experienced by a human driver: individual, first-person, and real-time. There hasn't been any serious discussion about this approach in the AV ethics literature, perhaps due to technological feasibility. But it's still relevant to consider that these dimensions might not be fixed. This rationale gets some support from existing empirical research, as there have been studies using different dimensions. Some have used third-person decision-making with no time constraints using written and picture format scenarios (Awad et al. 2018; J.-F. Bonnefon et al. 2016). Some on the other hand, have managed to implement first-person decision-making with time limitations using VR technology (Faulhaber et al. 2019). I assume that the proposed ethical setting will be formed collectively, from a third-person perspective and without time constraints. Despite this, it's still important to acknowledge that this might not be the only way to approach AV ethics research or implementation.

3 Constructing an ethical setting

3.1 Existing guidelines for an ethical setting

The German Act on Autonomous Driving (GAAD) is currently the only binding regulation on AV ethics. It consists of 22 ethical rules that AVs must comply with, most of them concerning the relationship between traffic participants and an AV. By analysing GAADs position, we can observe how current regulation takes the chosen moral variables into consideration. Later in the results chapter, GAAD is also compared to empirical results to see how well current regulation matches public preference.

GAAD's documentation doesn't explicitly take a stand if active intervention should be allowed or not. However, it does discuss the moral variables that are only relevant after active intervention is accepted as an option. Therefore, we assume that GAAD allows for AV to make moral decisions and focus on the following variables. The paper contains mixed statements on harm minimisation. It does introduce some principles that would favour harm minimisation as the primary variable in an AV crash, although this is left for interpretation. On the other hand, it condemns any sacrifice of innocent people for another. But considering the following statement, we'll assume that harm minimisation is preferred, albeit not explicitly:

> "A different decision may have to be taken if several lives are already imminently threatened and the only thing that matters is saving as many innocent people as possible. In situations of this kind, it would appear reasonable to demand that the course of action to be chosen is that which costs as few human lives as possible."

GAAD also states that an AV should minimize harm in a consistent manner, being equal to all parties. This is in line with the Rawlsian ethics that we'll discuss later in this chapter. And even if GAAD is not entirely clear on harm minimization, it establishes a clear stance on what is considered harm. It states that animal safety and possible material damage are not relevant if there is human life at risk. The definition of harm is not well defined in academic literature, so we'll adopt GAAD's view in our framework.

The most comprehensive rule of GAAD is the 7th, which establishes the regulators' position on all remaining moral variables: participant prioritization, legal responsibility, and personal characteristics. The states that:

> "In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical or mental constitution) is strictly prohibited. It is also prohibited to offset victims against one another. General programming to reduce the number of personal injuries may be justifiable. Those parties involved in the generation of mobility risks must not sacrifice non-involved parties."

At first, it seems that the stance on prioritization is unequivocal. But in another chapter, it is also stated that users can't be demanded to sacrifice themselves in an accident scenario. Because of this statement and the fact that pedestrian priorisation is not mentioned in the paper, I conclude that GAAD is slightly biased toward passenger prioritization. It is possible that because of the significance of the auto industry for Germany, the ethical commission's views could be biased toward solutions that aid AV commercialization.

On the secondary variables, interpretation is more straightforward. GAAD documentation states that "*Those parties involved in the generation of mobility risks must not sacrifice non-involved parties",* suggesting that legal responsibility should affect harm distribution. This is important for public acceptance, as legal responsibility is a deciding factor for assigning guilt in human driver accidents. It's likely to be relevant to public opinion on AV accidents too, as proposed by Kauppinen (2021). The act is also clear on AVs not being allowed to decide based on personal characteristics. This view is understandable from a legislator's point of view, as equal treatment is the premise of a just legal system. But from a public acceptance perspective, this is where GAAD misaligns itself from the moral preferences of laypeople. Its views are opposite to most empirical results, such as MME, where a large majority regards prioritizing children as a top priority (Awad et al. 2018). This might again be a classic case of the affect heuristic, where strong feelings towards children getting hurt by AVs make people forget how fundamental equality is to a just legal system. But even if the public preference is irrational, it still needs to be considered as it may severely hinder AV adoption, especially in families (Kopecky et al. 2023). To conclude, current AV legislation is not taking any firm positions to either end. The GAAD documentation contains several conflicting statements, and at its current state, seems too ambiguous to be used in programming an ethical setting. One of the main takeaways is that it recognises the absurdity of demanding driver self-sacrifice by default, something that might severely hinder AV adoption.

3.2 Basing an ethical setting on normative ethics

AV ethics literature has presented countless theories as a base for a functional ethical setting. For this thesis, I have chosen to focus on three major theories: utilitarianism, deontology, and Rawlsian ethics. These three are most prevalent in previous studies and they also have sufficient overlap with smaller theories. In this chapter, I will explore how they would fit as an ethical setting, and in the empirical part their relation to public preference is examined. To aid discussion, we'll introduce 5 requirements that an ethical setting must fulfil to be functional:

1. Technical feasibility. Can an ethical theory be transferred into an ethical setting?
2. Transparency. How easy it is to see why a certain decision was made?
3. Representation of reality. Can an ethical theory address all five moral variables?
4. Universality. Can an ethical theory deliver an adequate response in every possible scenario?
5. Social acceptance. How does the ethical theory comply with public preference?

   (Geisslinger et al. 2021)

It must be noted that when we are talking about technical feasibility, we are talking on a theoretical level. Meaning, how easily an ethical theory could be communicated to a computer if we had perfect programming knowledge. In reality, it will be immensely challenging to program an ethical setting to make split-second decisions. Now we'll go through the fundamentals and existing research on all three theories and see how they align with these five requirements.

3.3 The utilitarian approach

Utilitarianism is the most prominent form of consequentialism, a class of normative ethics that considers the rightfulness of an act based on its consequences. Utilitarianism was first conceptualised by Jeremy Bentham in the late 1700s and considers an act to be justified if it results in the greatest amount of good for the participants, or the least amount of harm in the context of traffic accidents. (Driver, 2022; Faulhaber et al. 2019.) By definition, a utilitarian AV would prioritize minimizing overall harm. A utilitarian AV

would base its decision most likely on a utility function that calculates harm probabilities for each possible trajectory and chooses the one with the least amount of overall harm, without a default setting for priorisation (Faulhaber et al. 2019).

3.3.1 Feasibility and transparency

A utilitarian ethical setting is an outcome-based decision-making model. This means that it lacks any hard-coded rules, and instead tries to optimise the outcome of a crash based on a utility function. At its core, utilitarianism as an ethical theory tries to seek an optimal result based on the risks and benefits available. This optimisation is also a common practice in programming and therefore, a utilitarian ethical setting scores high on technical feasibility. (Geisslinger et al. 2021.)

The drawback of the outcome-based model is that it may be less transparent to users. As it doesn't follow any logical set of rules, one can't predict beforehand what a utilitarian AV would do. It might also be difficult to communicate the decision-making logic of a sophisticated utility function to the user. Goodall (2014) has expressed some additional concern related to this, stating that because of no visible rules, it would be easier for manufacturers to secretly make their AVs prioritize the driver, as it's harder for regulators to audit outcome-based systems.

3.3.2 Representation of reality and universality

A utilitarian ethical setting scores high on the representation of reality, as it can account for four out of five variables present in our AV decision-making model: Intervention, minimizing overall harm, prioritization, and personal characteristics. Starting with intervention, a utility function could be able to address the need for intervention for each case individually. Of course, assuming that intervention is still accepted as an option. A utilitarian AV would just be flexible in assessing the need for active intervention. For example, at low speeds, braking and staying on the course might be the harm-minimizing solution, and the non-interventionist decision is taken naturally.

The primary variable for a utilitarian ethical setting is of course minimising overall harm, the basic premise of utilitarianism. It might be up for debate how harm is calculated and weighted for each traffic participant, but the solution would still always strive to minimize harm under the given parameters, which are defined by people.

On priorisation, a utilitarian setting is less explicit. According to utilitarian doctrine, the priorisation of traffic participants would be directly correlated with quantity

and injury probability. Any participant wouldn't be preferred by default, but instead, the AV would calculate probabilities for each one and choose the action with the least total harm. The problem is that although a decision can be produced under a utilitarian setting, it is specific to the situation and thus can't be predicted. It is also seen as problematic from a system transparency perspective by Gerdes & Thornton (2016). They state that to the user, an outcome-based decision may seem random as one can't say beforehand who gets prioritized. Also realistically, a utilitarian setting might bias pedestrian protection, as the passenger is protected by the car.

Of the two secondary variables, accounting for legal responsibility could the shortcoming of a utilitarian setting. Goodall (2014) presents an excellent analogy for how possible problems may emerge. Consider if a utilitarian AV had two options: to hit a motorcyclist with a helmet and protective gear, or a motorcyclist without. An AV purely focused on minimizing harm would choose to hit the one with a helmet, as it minimizes overall harm. But subjecting someone to more harm because he is more responsible is likely to be shunned by the public. It could also create a moral hazard to act more carelessly, as AV is known to prioritize the most vulnerable participant regardless of personal responsibility. (Goodall 2014.) Supporting views are presented by Gerdes & Thornton (2016), who argue that although technically challenging, a utilitarian AV should be able to assess participant accountability in creating the situation. This would create an incentive for certain behaviours such as helmets, resulting in safer overall traffic. Not being able to account for legal responsibility could be a problem if the public preference is strongly favouring it. Of course, one could argue that legal responsibility could be implemented into the utility function. But then if some boundary condition like this would result in not minimising overall harm, it wouldn't be a pure utilitarian system anymore. To simplify things and to match the current literature, we assume that only injury severity and probability are concerned.

On personal characteristics, a utilitarian setting might offer a loophole to implementing child protection in a legally acceptable way. This is because some individuals might have an increased injury probability due to personal characteristics, such as children. However, this logic would also apply to older people. A utilitarian AV could be the most democratic when deciding whom to assign harm. Because the only thing affecting the decision is the expected harm for an individual, a utilitarian AV could not be made to target minorities or other vulnerable groups intentionally, a fear expressed by Kriebitz et al. (2022). It might also be the flexible in case regulation gives in to public

preference and child prioritization is allowed, as the utility function could be modified to give even more weight to children.

In summary, the representation of reality is high for a utilitarian AV. Even if it can't provide a preliminary answer on priorisation and personal characteristics, it will still produce a decision based on it. This makes it also score high on universality. Geisslinger et al. (2021) argue that universality is dependent on the technological sophistication of the utility function, not on the adequacy of utilitarianism for an ethical setting itself. Universality is also enhanced by the fact that an optimization task will always produce a decision. Even If this decision might be suboptimal due to programming mistakes, the AV would still be able to react somehow in all scenarios.

3.3.3 Social acceptance

Taking a utilitarian approach has strong empirical support from existing studies. In a study by Bergmann et al. (2018) , 95% of respondents chose to sacrifice a single person. It didn't matter if the single person was a driver or pedestrian if the other option was to hit multiple people. These results are supported by Bonnefon et al. (2016), who saw a clear preference to sacrifice AV passengers if that would save a greater number of lives. The utilitarian approach of prioritizing the number of people saved was also one of the few variables that were preferred by people globally by similar percentages in the Moral Machine Experiment (Awad et al. 2018). On the contrary, one might think it's not justified to demand self-sacrifice as a prerequisite for adopting a new technology like an AV. This view is recognized by GAAD which was discussed earlier, where utilitarianism was preferred, but demand for self-sacrifice was also deemed as excessive. But as most empirical studies have proven utilitarian principles to be preferred, there is good ground to base an ethical setting on it.

3.4 The deontological approach

The second widely discussed ethical theory in the AV ethics literature is deontology. Whereas utilitarianism defines an act as rightful by its consequences, deontology focuses on the act itself. Deontology considers something to be rightful if it aligns with a core principle that is universally regarded as "right", such as not killing an innocent person. A German philosopher Immanuel Kant defined moral rightfulness as a categorical

imperative, a rule that shouldn't be broken in any condition. The categorial imperative has three forms:

1) "*Act only on that maxim through which you can at the same time will that it should become a universal law*."

2) *"Act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means, but always at the same time as an end."*

3) *"An act is morally right if and only if the agent, in performing it, follows the law autonomously."*

Especially the second reading of the categorial imperative is especially relevant for AV ethics, as it directly sets boundaries for what an AV can and can't do when distributing harm. (Britannica - Deontological Ethics, 2023; Feldman, 1978; Kant, 1981.)

It's clear to see how a deontology would be suitable as an acceptable moral setting. Although often compromises, laws represent an agreement on what people have deemed righteous. Under a deontological model, an AV will therefore act morally if it acts based on rules that are commonly agreed on.

3.4.1 Feasibility and transparency

Whereas the utilitarian model was outcome-based, a deontological model is a rule-based. Instead of a utility function, it uses hard-coded rules and relies on their correct hierarchical order to produce a decision. A good analogy for a rule-based model is the *three laws of robotics*, a famous set of machine ethic rules described by Asimov (1940). In the original book *I, robot*, the laws describe the relationship between a human and a robot as follows:

1. *A robot may not injure a human being or, through inaction, allow a human being to come to harm.*

2. *A robot must obey orders given it by human beings except where such orders would conflict with the First Law.*

3. *A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.*

It would be possible to build a deontological setting based on this kind of approach. For example, the rules concerning participant priorisation could be presented in a hierarchical order similar to Asimov's rules:

1. An AV should strive to minimize the amount of harm in an accident.
2. An AV should protect the passenger except if it conflicts with the first law.
3. An AV should avoid material damages except if it conflicts with the first and second laws.

From a technical perspective, this kind of rule-based decision-making is also common in programming. The technical feasibility of the deontological setting is therefore high. (Geisslinger et al. 2021.) Of course, this approach is highly simplified, but this demonstrates how deontology also offers more transparency than a utilitarian setting. Because the decision-making is based on absolute rules, it might be easier to communicate to the AV user. It would also be easier to see why a certain decision was made after an accident has happened. Goodall (2014), supports this view and argues that the deontological setting allows for better post-accident evaluation, as the decision logic could be presented as a decision tree, for example.

3.4.2 Representation of reality and universality

Because a deontological setting would act based on the rules assigned to it, it is not predisposed to some decisions like a utilitarian model is. It therefore can provide an answer to all five variables in our decision-making model, given that they are transferred into appropriate rules and society is able to arrive to a consensus regarding them.

On intervention, the deontological setting is also flexible in choosing whether or not to intervene. If it is not allowed, then "No active intervention" would be the only rule in the Asimovian hierarchy presented earlier. If it is, then it could be the last rule, stating," No active intervention unless it's necessary to fill all the rules above". Intervention is therefore dependent on our decisions where to place it in the rule hierarchy.

Then let's consider minimizing harm and prioritization under a deontological setting. As with intervention, the decision would depend on the rules assigned to the setting. Even though deontology is based on core principles such as "don't kill", this approach is not suitable because as are discussing non-avoidable accidents. It therefore must be specified with additional rules that how quantity and vulnerability should affect

a decision. Compared to utilitarianism, deontology offers more freedom in describing harm and who to prioritize. But it does present some boundary conditions. According to deontology's second reading of the categorical imperative, it is strictly forbidden to use people as a tool in achieving a desired result. (Britannica - Deontological Ethics," 2023.) This means that although not defaulting to one type of participant, a deontological setting would forbid transferring harm to outside parties. This would narrow the possible decisions to only those involved in the accident, whereas utilitarianism would allow hitting bystanders if it would minimise overall harm. The literature doesn't seem to offer any suggestions how a deontological setting would address legal responsibility. I'll argue that it can account for it and actually regards it as an important variable. This is based on two factors. Firstly, it could be easier for a rule-based model to take legal responsibility into consideration from a technical perspective. An example case could be that an AV will need to prioritize pedestrians if they act lawfully. From AVs perspective, the decision tree is simple: if the pedestrian is not breaking the law, he is prioritized. If he is breaking the law, the passenger is prioritized. Secondly, a person breaking traffic laws is also acting wrong on deontological grounds, as laws are our collective agreement on what is "right". Similar logic based on collective agreement could be applied to personal characteristics.

A rule-based approach would erase the perceived randomness that we concluded as one of the shortcomings of a utilitarian setting, as it would be easier to know AVs decision logic beforehand. So, to conclude, the deontological setting beats utilitarianism in the representation of reality but loses in universality. Whereas a utilitarian setting can make a suboptimal decision, stricter deontological system could fail to act completely if the situation is too complex. Careful planning is demanded from the developers that a rule-based system doesn't freeze in an accident because the situation doesn't match its strict rules. (Geisslinger et al. 2021.)

3.4.3 Social acceptance

Deontology solves utilitarianism's inability to address legal responsibility and perceived randomness of decisions. To gain even more social acceptance, rules could be categorized into forbidden, permissible, and obligatory actions. By having a hierarchy for the rules, it would perhaps be easier to communicate the AV ethical setting to the user. (Geisslinger et al. 2021.) It could also be beneficial from a regulatory standpoint, as each of these categories could be addressed separately. Especially the division of forbidden and permissible would allow for non-absolute prioritization rules. It could be for example,

forbidden to sacrifice the passenger by default, but permissible if it would save more lives. Compared to the utilitarian setting, the deontological approach is better at taking secondary variables into consideration, as it could permit even age-based prioritization if it is accepted by the jurisdiction.

3.5 The Rawlsian approach

The third ethical theory chosen for discussion is a version of contractualism, more precisely the version of John Rawls, presented in his book *A theory of justice* in 1971. Rawls's ideas are widely discussed in the relevant AV literature as a promising approach to a functional ethical setting. Rawlsian ethics regards an act lawful, if it could be agreed by all people from a theoretical viewpoint called "*the original position".* This is a decision-making perspective where participants know the outcomes of each action, but not their role in the situation. A decision made on the premises of Rawlsian ethics is then rightful if all parties agree to it not knowing which side of the dilemma they are. Rawlsian ethical setting would also introduce a minmax- approach to decision-making. In minmax, the goal would be to achieve the best possible outcome for the party that is initially in the worst position. In AV context, this would mean minimising harm for the participant whose probability and severity to harm is the highest. (Leben, 2017; Rawls, 1971.)

3.5.1 Feasibility and transparency

From a technical perspective, a Rawlsian setting would be a hybrid model between outcome-based utilitarian and rule-based deontological models. There is no exact description of a Rawlsian algorithm in the literature, but Leben (2017) provides some guidelines on how it could be implemented. It is similar to a utilitarian setting but with some key differences. Instead of an unrestricted goal for harm minimisation, a Rawlsian setting would introduce individual payoff calculations and utility functions. This is the first key difference to a utilitarian setting, which tries to minimise the overall harm in the situation. A Rawlsian algorithm would instead use the minmax approach to select who to prioritie. It would calculate the payoffs for each participant, and then maximize the utility function of that individual. Like with two previous ethical theories, this approach is also natural for computing.

On transparency, the Rawlsian setting suffers from the same problems as the utilitarian one. Without a clear set of rules, a decision may seem random to the user. It is

also a concern that the algorithm behind calculating the individual payoffs is man-made and might contain some bias toward certain participants (Leben, 2017). Another shared problem is that these possible biases are hard to locate from the code and would only be revealed by long-term accident data (Goodall, 2014). So to conclude, a Rawlsian ethics has good technical feasibility, and places between deontology and utilitarianism on transparency. It is not as clear as a rule-based system, but I'll argue that it's more transparent than pure utilitarianism, as some preconditions are given for the decision.

3.5.2 Representation of reality and universality

A Rawlsian setting provides a similar representation of reality to utilitarianism. It can account for four out of five variables present in our AV decision-making model: Intervention, minimizing overall harm, prioritization, and personal characteristics. On intervention, it can also decide if intervention is needed based on its utility function, and only intervenes if the optimal solution is achieved through it. So, we can conclude that intervention isn't a problematic variable as all three theories can address it. Where a Rawlsian setting differs from a utilitarian is that it doesn't regard harm quantity as a deciding factor. It therefore doesn't try to minimise *overall* harm but tries to minimise it for the party that it calculated to have the lowest odds of survival. It's clear to see that this approach is not without its problems. According to Leben (2017), it is theoretically true that a Rawlsian machine would prefer to injure an infinite amount of people if the other option would be to kill one.

On passenger prioritization, a Rawlsian setting doesn't provide a direct answer, but it's quite easy to conclude from theory. If the goal is to prioritize the worst-off participant, it almost certainly means prioritizing pedestrians and other more vulnerable road users, as participant quantity was not relevant. But it's important to recognise that these biases would are based on realistic injury probabilities. On a theoretical level, a Rawlsian setting would see all traffic participants as equals. This means that in a case with equal chances for all participants, the AV would randomise the decision. If everybody has the same chances of survival and one's position is unknown, randomisation is a fair decision from the original position. (Leben, 2017.)

Like utilitarianism, legal responsibility is Rawlsian setting's weak point. It has the same shortcoming presented by Goodall (2014), where a Rawlsian AV would prioritize the most vulnerable party, even if the vulnerability was caused by one's own actions. On personal characteristics, a Rawlsian algorithm would not prioritize children directly, but

would likely be heavily biased towards it. This is because children would probably be the most vulnerable party in traffic, at least in a non-passenger role. It must be noted that Rawlsian setting would, in theory, also allow for direct age-based prioritization. Savulescu et al. (2019) proved this in their study, where answers were given from the original position. If the age difference was small, people preferred to randomize the decision. If it was larger, they preferred to save the younger participant, choosing to self-sacrifice in the chance that they would be in the position of the older participant. Judging from these results, prioritizing children is socially acceptable if the age difference is substantial. But as the basic premise was to prioritize the worst off, this would have to be implemented as some form of override rule.

So the Rawlsian setting performs similarly to utilitarian in the representation of reality, although the bias towards pedestrians and children is clearer due to perceived vulnerability being more relevant than harm quantity. It is also highly universal and can provide a decision in all cases except when there are completely identical minimum payoffs present. In these cases, a Rawlsian algorithm would prefer to randomise its decision, which is optimal for equality but not for public acceptance (Leben, 2017).

3.5.3 Social acceptance

On a theoretical level, Rawlsian ethics would have high social acceptance, as it calls for decisions that all participants would agree on being objectively right. But as Rawlsian application in AV ethics would be based mostly on the minmax- principle, the setting suffers from the same problems as utilitarianism. These were the perceived randomness of decisions and the inability to account for legal responsibility. A little help is gained through the possibility of allowing age-based prioritization. But then it would have to be implemented through an overriding rule, which would be a solution already possible with the deontological setting. It therefore is more important how other variables are preferred to be solved, as deontology and Rawlsian ethics both allow for prioritization of children.

4 Implementing an ethical setting

4.1 Role of public acceptance and ethical policy

It's unlikely that people will accept regulation that doesn't align with public opinion, especially when personal safety is concerned. If no ethical theory provides a clear answer on what's objectively right, then public opinion should have the final say. (Awad et al. 2020.) To begin assessing public acceptance, we must be sure that the public would accept an AV to make moral decisions in the first place. Judging from the results of Karnouskos (2021), there seems to be public acceptance of active intervention. In their results, 63% of participants trusted a computer to make better-calculated decisions than a human driver. Considering these findings, discourse should be focused on the other 4 moral variables as those seem to have more incoherent public preferences.

To better communicate AV ethics to the public, it's important that we distinguish the difference between an ethical setting and ethical policy. An ethical policy is our collective agreement on which ethical principles an AV should base its decisions. An ethical policy sits on top of the ethical setting and acts as a boundary for developers on how their AVs can act. The ethical setting can be highly technical and ambiguous, especially in the case of the outcome-based model. It is therefore up to the policy to inform the public how an AV is required to act in an accident scenario. (Liu & Liu, 2021.) Also, a lot of thought must go into these policies to eliminate possible loopholes, as it was concluded that in the absence of regulation, AV manufacturers have a financial incentive to prioritize passenger safety. In addition to possible ambiguity of an ethical setting, it is also possible that some users are not familiar with the ethical theories behind the ethical setting. Most people seem to evaluate AV actions not through specific ethical lens, but through the standards of the culture they are a part of. (Faulhaber et al. 2019.) This begs the question of how do different ethical theories fit different cultures? A global ethical setting can prove difficult to implement if all development has been made from Western point of view.

Although agreement is usually the preferred goal, legislation can also act independently from public morality. A good example is euthanasia, which is widely banned in Western societies, even though evidence suggests that it's widely accepted by the public. On the other end, there are no legal reasons why a couple couldn't use technology to affect their child's gender, but it's forbidden due to fierce public opposition.

These examples showcase that regulation can be irrational when it comes to listening to public preference, and people should expect AV ethics regulation to not align perfectly with public moral standards. (Savulescu et al. 2019.) The authors of the Moral Machine Experiment also argue that although public opinion should translate directly to policy, it should still be researched and considered (Awad et al. 2018). Also, public preferences don't always align with what's considered moral, and public perception of right and wrong is dynamic and evolves over time. There have been numerous times when public preference has been in favour of morally questionable practices, such as apartheid in South Africa. (Savulescu et al. 2019.)

Nevertheless, the role of public acceptance must be considered when we start to form ethical policies to govern ethical settings. It's clear to see that now, regulators haven't been researching public opinion, as many views of the German Act on Autonomous driving differ drastically from public opinion concluded from empirical studies. This thesis tries to correct this by acknowledging the role of public acceptance and comparing empirical results to existing regulation, to highlight disparities that need addressing.

4.2 Universal ethics setting and the prisoners dilemma.

Most notable work for the universal regulation of an ethical setting has been carried out by Gogoll & Müller (2017), who are also heavily advocating against driver protection. They present their argument through a game theoretical thought experiment called "The Prisoners Dilemma" and transfer it to the AV domain. The prisoner's dilemma refers to a psychological problem described by Flood (1958) where prisoners A and B are interrogated separately, and the police don't have enough evidence to jail them for longer than one year if they don't get further evidence. So, if they both stay silent, they face a year in prison. If both testify against each other, they both face two years. The police also present a deal that if they testify against their partner, the other one goes free while the other faces three years. Let's see this problem unfold in a visual form for more clarity:

	A stays silent	A testifies
B stays silent	A = 1 year B = 1 year	A = 0 years B = 3 years
B testifies	A = 3 years B = 0 years	A = 2 years B = 2 years

Table 2 - Payoff table of the original prisoner's dilemma.

The problem in the prisoner's dilemma is that the individuals maximise their own benefit by testifying, regardless of what the other person chooses. It's clear to see that there exists a pareto-optimal solution, which is achieved if both stay silent. The problem is that the prisoners don't know what the other one is choosing, and don't want to be taken advantage of. So, they both try to minimise their sentences under uncertainty and testify, resulting in the longest combined amount of jail time.

According to Gogoll & Müller (2017), allowing for a customizable ethical setting results in more dangerous traffic and a higher overall probability of injury for traffic participants, similar to the prisoner's dilemma. Consider choosing a utilitarian setting as staying silent, and passenger prioritization as confessing. A utilitarian AV distributes 1 harm when crashing and takes 1 when crashed into, whereas an egoist AV distributes 2 and takes 0. Let's see how this scenario of A crashing into B would look like as a game theoretical representation where 1 harm injures and 2 kills the participant:

	A altruist	A egoist
B altruist	A 1 = Injury B 1 = Injury	A 1 = Injury B 2 = Killed
B egoist	A 2 = Killed B 1 = Injury	A 2 = Killed B 2 = Killed

Table 3 - Payoff table of the prisoner's dilemma in AV context.

Once again, there exists a Pareto-optimal solution of choosing a utilitarian AV for both. But traffic participants can't know what other people have chosen and therefore, they'll maximize their chances by opting for the egoist AV, resulting in the worst overall outcome in the lower right quadrant. This game-theoretical model proves a good case of how a utilitarian ethical setting results in the safest overall outcome, which would be also in the best interest of those users who prefer to be prioritized as a passenger

Avoiding the prisoner's dilemma is widely accepted in the literature as the main reason why an ethics setting must be mandatory. Bergmann et al. (2018) stress the need for regulation as people are unlikely to realise how allowing self-sacrifice would result in safest overall traffic. This draws us back to the "why ethics matter" chapter and the three psychological heuristics. People are not acknowledging the pareto-optimum of choosing an altruist AV when their own lives are at risk. Instead, they base their decision more on emotions and risk aversion. The result of the game theoretical model has also been validated quantitatively by Mordue et al. (2020), who simulated crashes in a mixed fleet of altruist and egoist AVs to see how the setting choices affect traffic safety. In their results, the passenger in the altruist AV was 37% more likely to die when there are mixed settings, but the lowest number of casualties is achieved by having 100% altruist AVs. These results are in line with the game's theoretical model and add to the argument that the chosen setting should be universal and mandatory. But although this may sound like a claim for an utilitarian setting, the model of Gogoll & Müller (2017) or the results by Mordue et al. (2020) don't advocate for certain ethical theory. They simply advocate for an universal ethical setting and consider harm minimisation as the most important moral variable.

4.3 In favour of a customisable ethical setting

Although most of the recent research seems favour regulation, there are some viable arguments made for allowing users to decide on their own ethical settings. A customisable ethical setting could also be a solution for a responsibility gap formed by AV use. This refers to people's need to have a clear party to blame in an accident scenario. The passenger can't be blamed as he lacks any control, and blaming the AI controlling a vehicle would not likely satisfy this need. (Kumfer & Burgess, 2015.)

If the user would have the possibility to make the AV bias certain decisions, it could be easier to assign some responsibility to him. This would clear the responsibility

gap of an AV crash, as with a regulated algorithm as the user lacks control and thus can't be held responsible. (Sandberg & Bradshaw-Martin, 2013.) Also, the proponents of regulation assume that given a choice, everybody would automatically favour passenger protection. In reality, there might be a lot of variation between how altruistic people are. An old couple might think that they have lived long enough, thus wanting their car to prioritize others. A new father on the other hand could reasonably want his car to prioritize him. (Gogoll & Müller 2017.) A customizable ethical setting would allow the AV to base its decisions on the moral stance of its owner. This would make the AV a "moral proxy" instead of a moral agent itself, which would give some validity for placing more responsibility on the user (Millar, 2015).

A middle-ground approach for the ethical setting is proposed by Soltanzadeh et al. (2020). In their model, the users would be able to customize minor settings, such as environmental values. Meaning, that in a hurry they could allow the vehicle to take a different route at a faster pace, and on different occasions use the longer, most economical route. The moral settings that would affect other traffic participants are regulated. Or taking it a bit further, a driver could allow for passenger prioritization, but regulation would forbid not minimizing overall harm. The passenger's preferred settings would be used only when there is an equal amount of different traffic participants. This kind of approach could possibly solve the dilemma found by J.-F. Bonnefon et al. (2016), where people preferred utilitarian AVs for others but wanted driver protection for themselves. Now utilitarianism wouldn't be demanded by default, but only when there are more lives on the other side. I argue that most people who prefer driver protection would see this as the morally correct thing to do. This is backed up by empirical evidence by Bergmann et al. (2018) , who observed an inverse correlation between the preference for protection and number of pedestrians in danger.

4.4 Quest for consensus: discourse ethics

Although there are a lot of different theories being proposed by AV ethics literature, there are no real efforts to arrive at a consensus. Sure, the three ethical theories described in this thesis do overlap with each other, but they still disagree heavily on, for example, harm minimisation. In this final chapter, I'll try to address this problem and provide a framework for my empirical research. For this purpose, a fourth ethical theory is

introduced: discourse ethics based on the thoughts of Jurgen Habermas. Unlike the other three, discourse ethics doesn't take any stand on what is morally correct. It is more of a tool to achieve agreement on ethical issues. Whereas Rawlsian ethics regarded an act as rightful if it could be agreed on from the *original position,* discourse ethics sees an act as rightful if all participants would agree on it in a *rational discourse* (Habermas, J. (1996).

A rational discourse that provides validation for a moral standpoint is characterized by two things: The ideal speech situation and universalization. The ideal speech situation refers to an environment that fosters good argumentative practices, from equal participation to the right to say or question everything. Though no discussion with actual humans with emotions can match the exact criteria laid out by Habermas, it should act as the objective. (Habermas, J. (1996.) Universalisation means that discourse ethical assessment should be able to consider what is morally right for a wide range of stakeholders, not just for individuals or subgroups. Haberman's thoughts have some commonality with Kantian deontology, both providing a procedural approach for assessing morality. The key difference is that in discourse ethics, universalization is achieved by discourse, rather than a monologue and categorial imperatives. (Habermas, 1996; Feldman, 1978; Mingers et al. 2010.) The concept of universalisation is what could make discourse ethics a good approach to arrive at the preferred ethical setting, as AVs decisions will concern all traffic participants in the future. Discourse ethics also offers more flexibility than other normative theories. The problem with utilitarianism, deontology and Rawlsian ethics was that they were flexible on some topics but had rigid views on others (see Knaapi-Junnila et al. 2022). Discourse ethics could offer a way to implement the public preference *as is* to an ethical setting, of course in the boundaries of law.

The discourse ethical approach of directly considering people's collective agreement seems to have strong theoretical support. The EU Commission's *Ethics guidelines for trustworthy AI* establish human agency and governance as a requirement in AI system development. Human agency, as described by Gunn (2009), is differentiated from being a mere subject by the ability to cause change. This means that to be an agent in AV domain, people should have the ability to affect how the systems are developed and governed. This is supported by the commission's paper, stating that a trustworthy AI system should consider the opinions of all stakeholders affected by its decisions, and this is achieved by stakeholder discourse (Ethics Guidelines For Trustworthy AI, 2019).

A lot of Habermas' work discusses juridical topics and the role of the state, and

interesting parallels can be drawn between Habermas' view of a functional government and role of discourse ethics in AV development. According to him, a functional juridical system has three separate branches: the legislative (senate), applicative (courts) and administrative (government). These three systems should be kept separate to avoid asymmetrical power distribution, where one actor could implement and enforce laws arbitrarily. (Habermas, 1996.) This separation of powers is also critical in systems development, where special attention should be paid to separating the system requirement (legislative) and development branches (administrative) (Ross & Chiasson, 2011). In AV domain, this would mean exactly what Ethics Guidelines For Trustworthy AI (2019) proposed. Public consensus achieved by discourse ethics would dictate the system requirements, and these would be implemented by the AV manufacturers. It would then leave the state as the applicative branch to govern the process. The result would be a publicly accepted AV ethical setting, with functional separation of powers. The empirical part of this thesis goes on a quest to explore these system requirements, by mapping out individual-level preferences as well as utilising discourse ethics to move towards a consensus.

5 Methodology

5.1 Research objective and philosophy

The research objective was to find out how do laypersons prefer to solve moral problems faced by autonomous vehicles and to compare these views to normative ethics and current legislation. The idea is that an ethical setting should not be dictated by regulation, but instead be based on an established ethical theory that best aligns with public morality. The thesis claims that technological acceptance is unlikely if AVs don't align with the moral standards of their users. A clear research gap was identified in the literature, as AV ethics research has previously been focused on quantitative methods, such as studying averages on the importance of different moral variables. The subjects have not been offered the change, or on the other hand, required to motivate the moral preferences expressed in these studies. This might have resulted in a situation where a certain position is adopted based on subjective opinions or intuitions, as its morality does not need to be defended. The goal of the empirical research was then to extend the research to the underlaying reasoning behind people's moral standpoints, to give them credibility beyond simple questionnaire answers. Also, this study uses discourse ethics to triangulate the individual answers and to test their credibility, something that hasn't been used previously in AV ethics research.

When choosing a research methodology, it's important to understand what is meant by it. Eriksson & Kovalainen (2008) describe research methodology as a tool for uncovering and solving the research problem. To Tan (2017), a research methodology is a way to map out the research process from the research question to insightful conclusions. In this model, choices that a researcher makes during the research process are categorised into three parts: research philosophy, design, and method. Research philosophy dictates the suitable approaches for design and method, so one must place careful consideration in choosing it. To be able to choose it correctly, one must first understand what is meant by research philosophy in boarder terms. Although there are numerous distinct research philosophies available, a rough division can be made to causal and interpretive science. Causal science, as the name suggests, studies causality. The focus is on giving objective explanations and discovering causal relationships of the world around us. Research problems are solved either statistically with mathematical methods, or by identifying mechanisms through case studies. But the problem with causal

approaches, especially with complex problems, is that it's not uncommon for some things to seem falsely correlated. An example often used in scientific literature is the seeming correlation between ice cream sales and drowning deaths, which both peak during summer months. Although this correlation can be observed statistically, common sense tells that its actually caused by a third external variable, temperature. (Porpora, 2009.) As causal science focuses on presenting a hypothesis and testing it, results can be falsely validated if the correlation is caused by some external variable. (Tan, 2017.)

The second main branch of research philosophy is interpretive science. To interpretivists, world can't be explained objectively through a unified theory, as it is a construct of subjective views and individual experiences. Interpretive science seeks to find something novel, regardless of statistical realities. Often, interpretive research does not present a hypothesis. Instead, it uses an exploratory framework to describe the phenomena, background, and issues within. The answers obtained through interpretive research designs will be analysed using the explorative framework. The overarching goal is to find something new about a topic, not to present a unified theory. Whereas causal science suffered from false correlation and statistical biases, interpretivism suffers from the irregularity of personal realities. An interpretive scientist must understand that everyone will base their answers on personal views and motives, which are necessarily not clear to the researcher. Numerical representation of findings is not common. (Tan, 2017.) As AV ethics literacy doesn't have any commonly agreed-on theories and the topics discussed are highly subjective, an interpretive philosophical stance is an appropriate choice for this thesis. Both philosophies have their downsides, but false correlations seem more harmful to the research objectives than the irregularity of personal realities, which is a natural assumption when discussing personal moral preferences.

5.2 Research design

As Tan (2017) describes, interpretive research designs usually produce a narrative instead of numerical data. The analysis of this data is then focusing more on reasoning, rather than causal relationships. This description aligns with the basic premise of this thesis, as it aims to provide a deeper understanding of the moral preferences of laypeople as future AV users. When choosing a research design, it must be remembered that the chosen philosophical stance provides some boundaries on what kind of research design should

be used. According to Ghauri et al. (2020), a good research design will allow us to get the answers we are looking for efficiently while acknowledging possible resource limitations such as time, budget, or skill level. They also stress the importance of aligning the design structure and the research problem. A common mistake is that novel problems are paired with a strict and structured research design. Or vice versa, unstructured designs are applied to well-researched problems that would benefit more from causal methods.

There are no well-established theories on AV ethics and not much prior research has been made on the moral reasoning of users. The field is also scattered in terms of suitable ethical theories and governance models. It's clear that the research problem is unstructured, and the chosen research design approaches should reflect this. On AI research in general, technological innovation is the forerunner, while regulation and research try their best to keep up. It must therefore be assumed that there are no absolute answers available. To avoid shackling to certain views, an exploratory research design is adopted. This view is supported by Ghauri et al. (2020), who recommend explorative research with problems that are badly understood, or when theoretical consensus has not been achieved.

Explorative research can be defined as chasing systematic and purposeful research to produce generalizations and a thorough understanding of a social phenomenon that has previously been badly understood. The result of an explorative study should be an inductive generalization of the researched group or phenomena. Explorative research is a useful approach when scientific knowledge available on a topic is still limited, but it is nevertheless assumed studying it could lead to important discoveries. (Stebbins, 2001.) This is indeed the case in AV ethics, where the technology to implement a sophisticated ethical setting doesn't even exist yet. But this doesn't mean AV ethics research is not relevant, as theoretical research and dialogue can and possibly should be anticipatory to affect the actual development of AVs. As mentioned, technological development is proceeding at such a rapid pace that research can't get stuck behind.

Explorative research places very distinct skill requirements for the researcher. Instead of mathematical skills, more focus is on the ability to collect information by observation and social interaction. The researcher must also be able to compile the information into an explanation, which is called theorizing. (Ghauri et al. 2020.) Stebbins (2001), provides two principles that'll allow for effective exploration: flexibility and open-mindedness. A researcher must be flexible in how empirical material is searched and collected, as well as open-minded when analysing it. In this thesis, flexibility will be

demonstrated by adopting a novel research method, discourse ethics. Open-mindedness on the other hand is achieved by establishing the researcher's position purely as an observer and facilitator, who allows the subjects to present their ethical opinions without any prejudice.

5.3 Method

The word research method refers to an organised way of collecting and analysing data to solve the research question (Ghauri et al. 2020). These are divided broadly into two categories, quantitative and qualitative research methods. In quantitative research, the focus is on testing and explaining a hypothesis using mostly mathematical methods. In quantitative research, the goal is to expand or create new knowledge about the research problem. It also acknowledges that there is not one correct reality that can be presented statistically. (Eriksson & Kovalainen, 2008.) When choosing a suitable research method, we need to reflect to the chosen research design. Luckily, explorative research doesn't limit us to either qualitative or quantitative methods, as both can be utilised during exploration. However, these methods are not equally useful in the same research stages, as the maturity of knowledge and the research problem affect the correct method choice. When the research problem is new and unstructured, qualitative research methods are generally preferred. (Ghauri et al. 2020.) Quantitative methods can be utilized after exploration when there is a need to confirm the initial explorative findings (Stebbins, 2001).

For this study, qualitative methods are a natural choice for two reasons. Firstly, as the research problem tries to seek out *individual* preferences, there is a fundamental mismatch between the objective reality of quantitative methods and the research problem. Secondly, the research objective is to observe *why* a certain moral decision is preferred, so the obtained material will be predominantly linguistic, something that is hard to analyse with quantitative methods. This choice is supported by Ghauri et al. (2020), who emphasise the qualitative method's ability to generate holistic descriptions and understanding of novel topics. They state that qualitative methods are an appropriate choice when there is a need for in-depth understanding rather than qualitative confirmation. Also, Eriksson & Kovalainen (2008) mention that qualitative methods provide flexibility especially suitable for unstructured problems in novel domains, where an explorative approach is implemented.

5.4 Data collection

This study will utilize two qualitative methods with different objectives. First, individual interviews are conducted to discover a variety of individual preferences and motivations. The second phase will gather all the individual subjects into a discourse ethical group discussion, where the objective is to achieve at least a majority consensus on the same topics as in the individual interviews. The group interview will give us some insight into just how challenging it is to arrive at a consensus on these sensitive topics, and how socially acceptable the different arguments for each approach are. By having two methods, we also utilise triangulation as a method to further validate our results and create a more well-rounded picture of the research problem. Triangulation is the use of multiple methods to study the same problem. It can offer us more insights into the phenomena, as different methods can have different strong points in extracting information. (Ghauri et al. 2020.)

5.4.1 Interview

Interviews are one of the most used qualitative methods and are used extensively in both commercial and academic contexts. Interviews offer a flexible framework for addressing all sorts of research problems, and there are a lot of easily available resources on how to conduct them. (Eriksson & Kovalainen, 2008.) But interviews are by no means a simple method. They put a heavy emphasis on the interaction between the subject and the researcher, demanding more in form of soft skills like observation and interpretation. (Ghauri et al. 2020.) To conduct a qualitative interview, one must first know what kind of interview structures there are, and which one to choose.

Qualitative interviews are often categorized into structured, semi-structured and unstructured. A structured interview uses standardized questions, and usually predefined answers, although this is not a requirement. The answers are usually processed with statistical tools. This approach allows little in terms of flexibility, as all the interviews must have the same structure and variables to allow for statistical analysis. On the other end, an unstructured interview relies entirely on informal and open discussion, although it can have some guiding questions to keep the discussion on the desired track. This interview structure places the highest demand on the soft skills of the researcher. In the middle, there is the semi-structured interview, which has elements from both the structured and unstructured interviews. It consists of pre-determined questions and has the same basic structure for all participants, but also allows for open and informal

discussion about the interview topics. The main benefit of this method is the ability to have a standardized interview setting, but still have open discussion to gain new insights as possible answers are not fixed beforehand. (Eriksson & Kovalainen, 2008.)

The second consideration for the interview setting is whether we want to have a positivist, emotionalist or constructionist perspective, which then dictates what type of questions we should utilize. Positivist interviews are after facts and often structure the questions to extract as much information as possible. Emotionalist interviews are the opposite of this, focusing more on the participant's subjective experiences and viewpoints. The third option, constructionism, then shifts the focus on the interview situation itself, and how answers are derived from the narrative between the participant and the researcher. (Eriksson & Kovalainen, 2008.) Positivism and emotionalism are mostly focusing on the "what", whereas constructionism is after the "how". Although not necessary, the best results are often achieved if the researcher can include both types of questions. (Holstein, 1995.)

For this study, semi-structured interviews with a mix of emotionalist and constructionist perspective are chosen. As mentioned, we need the interview situation to include elements from both structured and unstructured interviews. To be able to generalize, the subjects must have the same questions with the same options, as we want all answers to be given from the same viewpoint. But to gain novel insights, we can't have fixed answers. So, the moral reasoning behind the answers needs the open discussion component, which will also benefit from using both emotionalism and constructionism. The interview questions will predominantly be what-questions, but the interviewee is also able to guide the conversation with additional "how" questions.

In the sampling of the participants, two things had to be kept in mind. Firstly, as the research problem was to see *How do laypersons prefer an AV to solve moral problems,* the participants naturally had to be regular people, not technical experts. But the sampling also needed to recognize that motivating one's moral stances is indeed as a challenging task, especially on a novel topic like this. As the number of subjects is low, there was no room for interviews that would get stuck on the surface level. The participants are therefore people who have at least an average understanding of ethics and good argumentative skills. As there were no pre-interview testing done, this mostly meant having at least a bachelor's degree. Although it is not claimed that a university degree and these skills are inherently correlated, this was the seen as the best criteria given the limited resources. The interview itself consisted of five trolley-problem scenarios with

fixed answers and trajectories. It was made clear that no other trajectories can be chosen even if there would be some other options in a real-world scenario. As mentioned in chapter 2, the trolley scenarios are a simplified representation of reality, and only need to include the same tradeoff between moral variables as in real life (Paulo, 2023). Their purpose is to facilitate discussion, not to solve the scenario itself. None of the available answers were inherently wrong, but they represented a different hierarchy between the variables in the scenario. After choosing an answer to a scenario, the open discussion started. Here, the participant had to argue why a certain option was preferred, and the interviewer used additional scenario-specific questions to guide the conversation wherever there might be interesting insights.

The interviews were conducted in Finnish, as it was seen necessary that complex topics must be discussed in the participant's native language. All quotations presented in the results chapter are therefore translated by the researcher but with a rigorous focus on maintaining their original tone and meaning. To further enhance the participant's confidence in the interview, a short 2-page summary of AV technology and general ethical guidelines was sent out before the interview, along with the translated interview form and a consent form. The interview form can be found in the appendix 1. The form also included some basic visualization of the scenarios to make sure the answers weren't based on a mental image of the scenario, which could differ among participants. All interviews were conducted face-to face at the participants residence. The interviews were recorded as it was mentioned in the consent form, and these recordings were then transcribed within 24 hours of the interview. This allowed the researcher to reflect on the interview situation while it was still fresh in memory. These transcriptions were then used for the exploration of AV ethics, and to answer the first research question.

5.4.2 Discourse ethics

The second method used in this thesis is discourse ethics. Its fundamentals were discussed in detail in the end of chapter 4. As mentioned, discourse ethics does not take any stance on moral issues, but rather provides us the means to reach an agreement on them (Mingers et al. 2010). It is adopted as a second qualitative method for three reasons. Firstly, it is unlikely that all individual moral preferences are accepted by a larger audience. Deciding on the objective morality of individual statements is a dubious task to be left to the researcher alone. As stated by Habermas, J. (1996), an act is seen as rightful if all participants would agree on it in a rational discourse. The discourse ethical group

interview (DEGI) allows us to see which of these individual positions are also accepted by other people, and just how much variation there is between people's opinions. It is also interesting to see whether some participant's opinions change when the choices have to be communicated to other people. This type of situation has been observed by Kopecky et al. (2023), where egoist AVs were preferred over altruistic ones only if this choice was not visible to other people. It is also worth examining how the motivations for people's moral preferences strengthen or weaken when confronted by people with possibly opposing views. Going back to the research gap, it was mentioned problematic that in many prevalent studies there has been no requirement to motivate the taken positions, which may have made it easier to express intuitive, but immoral preferences. Secondly, DEGI can be seen as the miniature version of the societal discourse which will eventually result in wider AV legislation. By observing which topics are easily agreed on and which are not, we can draw conclusions to possible sticking points in achieving a national or even global consensus. Thirdly, by utilizing discourse ethics as a form of triangulation, we can conclude what individual preferences are the most common and agreed upon in a group setting, a moral average so to speak. It might be impossible or at least impractical to compare all the individual preferences against ethical theories and GAAD. Instead, we'll use the *average* preference to answer the second research question: *How these preferences could be formed into an ethical setting?*

Some basic guidelines on how to conduct a DEGI are listed in an article by Westerstrand et al. (2023) , originally presented by Koskinen et al. (2023). Three things are emphasized to make the most out of this method: the attitude of the participants, the five rules of rational discourse, and the role of the facilitator. To have an effective rational discourse, the basic attitude of the participants must be open and without interpersonal conflicts. This is only possible if all participants are aware and comply with the five rules of rational discourse. These rules are outlined more in appendix 2, but they include such basic principles as open-mindedness, clarity, and sincerity in the discourse. The facilitator's role is to stay neutral, enforce these rules and to guide the discussion into the desired direction. In practice, the group interview will go over the same five scenarios as the individual interviews, but the topics are discussing within the group rather than with the interviewer. On each scenario, the facilitator has summaries all individual answers from each participant, and from there they have to motivate it using rational discourse. The more exact procedure is also found in the full documentation. The facilitator's role is to guide the discussion so that we'd end up with one commonly agreed position for

each scenario. As its unlikely that total agreement on all topics can be achieved within the time constraints of the interview, a majority vote will suffice. This approach is supported by Habermas himself and Ross & Chiasson, (2011), who both acknowledge full agreement as an ideal, not a realistic target. To them, a majority vote offers a sufficient base to make further decisions, as it is already common practice in most democratic processes, such as elections and legislation. But unlike in these examples, they emphasise that in a rational discourse, the majority vote is a viable solution only if there is genuine agreement that this position can be given up immediately if a better solution is found. (Ross & Chiasson, 2011.) The majority rule can be thus described as the best solution *for now,* and that is also how one should understand the results of the DEGI conducted in this study. A majority consensus, although imperfect, should still be more credible than individual preferences. It also offers a solution combat the irregularity of individual personal realities, which was mentioned as a flaw of interpretive science by Tan, (2017). Even an incomplete agreement in a small group is a step in right direction, from solely individual level preferences to a societal agreement on AV ethics. (Westerstrand et al. 2023.)

5.5 Data analysis and empirical data

5.5.1 Empirical data

For the individual interviews, empirical data consists of a recording and a written transcript for each participant. The transcribed material consists of 42 pages and 10 135 words of written text and is the primary material for the analytic process. For the DEGI, no transcribed material was seen necessary to produce. Instead, the data gathered in the group interview was a structured spreadsheet with the “moral average” solution for each scenario, which was then compared with the normative ethics and existing regulation. As motivations for each individual participant were already gathered during the first phase, transcribing the discussion in the group session was seen as unnecessary work. The interview was recorded, so the discussion could be revised if seen necessary. As no personal information was collected, the backgrounds of the subjects are not specified. But as mentioned, they all were people at least an average understanding of ethics and good argumentative skills. All individual interviews were conducted live at the participant’s homes, to enhance their confidence to speak about sensitive subjects. The group interview

was held online via Microsoft Teams, as it proved too difficult to align both schedules and locations of the subjects for a live session. Interview metadata can be seen from the table below, and all the necessary aspects of data collection, data items and privacy are collected in a data management plan provided by the University of Turku

Type	Execution	Participant	Age	Gender	Date	Length
Individual	Live	P1	56	Female	17.12.2023	42min 43s
Individual	Live	P2	23	Male	17.12.2023	31min 45s
Individual	Live	P3	54	Male	19.12.2023	32min 45s
Individual	Live	P4	25	Female	21.12.2023	36min 10s
Individual	Live	P5	25	Male	22.12.2023	32min 28s
DEGI	MS Teams	P1-P5	-	-	15.1.2024	1h 12min 29s

Table 4 - Interview metadata.

5.5.2 Data analysis

There are a lot of different methods for analysing research data. The choice must be based on the chosen methodology, the researchers' experience level and how the results are presented. As mentioned, qualitative methods are suitable for a more explorative research design, especially with novel, unstructured problems. (Eriksson & Kovalainen, 2008.) For many of the same reasons, a thematic analysis was the method of choice, as it matches all the criteria listed above. According to Braun & Clarke (2006), qualitative data analysis methods come in roughly two different shapes. First, some theories are bounded by a certain theoretical or epistemological position, which makes them perfect for problems fitting those positions but offers little flexibility if not. Thematic analysis is in the second group, where methods are not constrained by such positions. Because of the novel research field with no clear hypothesis, flexibility was seen as the key requirement for the analytic process of this thesis. Also, from the literature review it was evident that although individual, most of the moral preferences are based on some wider themes, such as altruism or egoism. Thematic analysis is also mentioned to be a suitable method for researchers with limited experience with qualitative methods, as it doesn't require extensive knowledge of any underlying theory.

To implement thematic analysis, we first need to define a theme. According to Braun & Clarke (2006), a theme is an overarching concept that is a foundation for a lot of different answers. There are no requirements on how prevalent something must be to be promoted from a code to a theme. It is the researcher's responsibility to spot the relevant factors that a lot of data items have in common, that could be regarded as a theme in the final analysis. Braun & Clarke (2006) also state that simply choosing themes on a quantitative basis might be tempting, but incorrect. A theme must be something vital for understanding and answering the research problem, that offers a way to structure the research data. In the analytic process of this thesis, themes were some concepts that formed a foundational basis for human moral judgment. These are not necessarily what was said most frequently, but things that can be thought of as immutable concepts, such as self-preservation. Besides choosing the right themes, the researcher must choose the focal width, as it greatly affects how the results are presented. Braun & Clarke (2006) argue that there are two basic ways to approach this that each have their strengths and weaknesses. The researcher can either choose to provide a holistic thematic description of all the data, with all the themes that the readers might see as relevant. In this approach, the coding process has to focus on creating an accurate representation of the entire data set. The drawback of this approach is that even if there would be some more dominant themes that would need much deeper analysis, it is not necessarily possible due to time or space constraints. In this case, the researcher could choose to focus only on one or a few key themes and provide a detailed view of those. No clear drawbacks for this method are mentioned, but one could see problems, especially in the neutral representation of the data, as the choice for the "most relevant themes" is left to the researcher to decide. There might be some vital information in the data that gets completely ignored. Because the fundamental idea of the interviews was that there are no right or wrong answers to the moral dilemmas, the analysis covers the whole data set, and the interpretation of what is the most important is left to the reader to decide.

Third choice that Braun & Clarke (2006) see as a relevant is whether a deductive, inductive, or abductive approach is implemented. These approaches describe if the analytic process and researchers' claims are based on existing theory or the data itself. In deduction, the existing theory acts as the base for making a hypothesis and testing it through empirical methods. The theory provides constraints for the research problem, and ultimately the analysis of themes in the data set. Induction takes the opposite approach and sees data as the primary source of information. In inductive analysis, themes are not

constrained by the research questions, interview topics, or theoretical framework. (Braun & Clarke, 2006; Eriksson & Kovalainen, 2008.) The third approach, abduction, is a combination of the two. In abductive analysis, it is recognized that often, neither deduction nor induction is suitable as a standalone approach. Abduction allows the researcher iteratively to use both approaches, in different stages of the analytic process. This may happen in the form of approaching the data with no thematic constraints but comparing the findings with the existing theory to either validate or challenge it. In this thesis, the analytic approach was abductive. Induction was used with the individual interviews, and theme formulation was not constrained by theory or interview topics. The discourse ethical group interview was coded with a heavier emphasis on theory and the moral variables as themes, as this data was reported as a comparison to existing theories and legislation.

The procedure to conduct thematic analysis was implemented from Braun & Clarke (2006), who describe a six-step protocol and possible pitfalls to be avoided. This approach was followed rigorously in the analytic process, and we'll go over it to increase the dependability factor of this study. In the first step, the researcher familiarizes himself with the data set. This meant several rounds of listening to the audio files, transcribing them into a written format, and then reading the written transcripts before starting any coding work. This process was aided by not using any digital transcription tools, as manual writing and listening forced the researcher to focus deeply on the interview materials. In the second step, data is analysed, and all relevant or interesting topics are given a code. In the third step, preliminary themes are formed from the most relevant or prevalent codes that share some common factor. In the fourth step, the researcher evaluates if all the themes, codes, and code extracts (citations) create a coherent narrative and represent the data accurately at the chosen width. In this step, some themes may be demoted to subthemes or excluded as irrelevant. In the fifth step, the final themes are refined in terms of content, quantity, and naming. A thematic map is updated from step 2, and by now the researcher should have a map that accurately represents the data with a logical structure. In the sixth step, the thematic map is used to report the findings. The researcher must acknowledge that he is not simply describing the data to the reader but must produce a narrative that presents an argument based on the research question. Here, sufficient focus must go into presenting enough evidence in the form of data extracts so that the reader can evaluate the credibility of the study. (Braun & Clarke, 2006.)

5.6 Research ethics and Evaluation

5.6.1 Research ethics

Ethical considerations are a vital part of conducting any scientific study and should be considered during the whole research process. According to Ghauri et al. (2020), ethical considerations and self-evaluation are needed because readers might take the researcher as an utmost authority, and not be able to evaluate the results objectively. It is therefore the ethical responsibility of the researcher to also point out the possible limitations of one's study and results. The researcher must also describe the research process and methods used to allow for external evaluation of the results. Ethical considerations are naturally relevant for this study due to the research problem. Not only because it discusses personal moral preferences, but because of the novelty of the subject. Eriksson & Kovalainen (2008) state that the novelty of the subject and the need for careful ethical considerations grow proportionally. This is because there might be new ethical dimensions in novel fields that have yet to be established. Researchers entering new fields must understand that it might be up to them to figure out the ethical considerations needed.

The objectivity of a study may be hindered for several reasons. Ghauri et al. (2020)consider third-party interests, researchers' own interests, and peer pressure from the scientific community as most likely to bias the representation of results. In this study, there are no third parties, so there is no mutual interest to achieve certain results. Also, no peer pressure is recognized. To account for researchers' interests and biases, careful consideration is needed in the relationship between the researcher-participant relationship. The researcher must assume a neutral role in the interviews, making sure that the respondents are the sole data source, and that the discussion is not steered towards a "desired" direction.

The three pillars of research ethics in this specific study are informed consent, voluntary participation, and confidentiality. Informed consent means that the participants are not only asked for consent but also informed about the research objective, methods, and data use beforehand. If they consent to partake after being provided with extensive background information, informed consent is granted. (Eriksson & Kovalainen, 2008.) To achieve this, all participants were approached with a written letter of consent, which provided basic information about the study that they would need to consider. The letter also described the rights of the subjects on topics like data usage and communication

during the research process. All respondents were also encouraged to ask anything, anytime, something that is thought to inspire trust among participants (Eriksson & Kovalainen, 2008). Due to the sensitivity of the subject, voluntary participation is perhaps the most important ethical consideration. This is partly achieved by informed consent, but also by designing the empirical process in such a way that the subjects can limit or end their participation if they choose so. In the consent form and before the interview, the subjects are informed that they can refuse to answer any question without further questions if a certain moral topic feels uncomfortable. This option was seen as crucial, especially in the group interview, as the presence of other people might make some participants more reluctant to open their moral preferences. To boost the confidence of the participants and to make sure they all have sufficient understanding of the topic, a summary of AV technology and ethics was sent to them before the interview. Confidentiality is assured by providing the subjects clear information about how and what data is gathered, how it is processed and how it is managed after the study. In the consent form, participants were given the chance to have their data erased after the study was completed. They all used this option, and all research material was deleted after the study was complete. Also, it was stressed that no personal information was gathered, and all anonymity was maintained throughout the research process.

Because of the lack of third-party interests, the neutral position of the researcher, and the rigorous focus on the three ethical pillars, there are no ethical limitations in this study that should be addressed specifically. The results are also reported as individual preferences and compared to existing ethics and legislation. This means that there is no researchers own interest and biases should not affect the reporting, as no benefit is achieved for reporting the data in a way that would align with e.g. existing literature.

5.6.2 Research Evaluation

When choosing evaluation criteria for research, the methodological choices must align with the evaluation criteria, as quantitative and qualitative methods require different approaches to evaluation. Failing to do so could decrease the quality of the research, as evaluation criteria guide the researchers' choices. It is therefore important to establish the right criteria is used before any empirical research is conducted. (Eriksson & Kovalainen, 2008.) Traditionally, quantitative research has been evaluated based on its reliability and validity. Reliability refers to how accurately a study can be replicated and yield the same results, and validity how well the results of a study can be deemed as "true". It's clear to

see that these criteria don't necessarily do well with qualitative research. (Eriksson & Kovalainen, 2008.) That is why an alternative set of evaluation criteria has been proposed by Lincoln & Guba (1985), who argued that the traditional approach is not suitable with research that acknowledges the existence of multiple, subjective realities. Instead of reliability and validity, this type of scientific research should be evaluated based on its *trustworthiness*, which is divided into four components: credibility, transferability, dependability and confirmability.

Credibility is the bedrock of good qualitative research. It refers to the accuracy of the findings from the data, and if these findings offer a sufficient representation of reality. It's important to recognize that despite similar tone, the quantitative validity and qualitative credibility are not the same. Whereas validity was described to evaluate the accuracy of results, credibility evaluates the methods and processes with which the results are produced. (Lincoln & Guba, 1985, as cited in Eriksson & Kovalainen, 2008.) The credibility of the research can be increased with prolonged engagement with the subjects and by utilizing different types of triangulation (Guba, 1981). In this thesis, the credibility component is pursued by having no time limits for the interviews and making them long enough to find meaningful insights. The interviews were also transcribed on within a couple of days from the interview, without any digital tools. This required careful listening of the whole recordings, which familiarized the researcher with the data before the analytic process. The data was analysed using a well-designed thematic model, so the findings could be found, analysed and reported logically. Both theoretical and method triangulation were used, which meant that the results were compared against multiple ethical theories, and the data was gathered using two distinct methods.

Transferability addresses if the findings could be used or compared with in another study or if they are dependent on this specific sample and research setting. It also considers if some connection can be made between the results and existing research. (Eriksson & Kovalainen, 2008; Lincoln & Guba, 1985.) Yin (2002) explains that generalization in quantitative and qualitative studies follow a different logic. Quantitative methods try to achieve *statistical generalization*, meaning that results should be replicable with similar sampling. Qualitative methods on the other hand, search for *analytic generalization,* which means that results are compared to existing theory, and generalization is achieved with several studies supporting the same theory. This helps to understand why small sample sizes in qualitative research don't automatically mean poor generalization, which is more dependent on the depth of information extracted, regardless

of the samplesize (Eriksson & Kovalainen, 2008). According to Guba (1981), transferability can be increased by providing a "thick description" of the research setting and choices, which allows others to conclude if the findings would be applicable in another setting. In this thesis, transferability is pursued by providing a thorough description of the collection, composition, and analysis of the empirical data. Also, the subjects are described as accurately as possible, given the privacy requirements. A possible shortcoming in the transferability of this study is the fact that moral preferences could change over time and through populations, so the results might have some dependence on the research setting.

Dependability addresses if the research process is described and documented accurately, and whether the researcher acknowledges the possible shortcomings of his own research choices. Dependability and credibility are closely intertwined, as the evaluator needs an accurate description of the research choices to assess the credibility of the findings. (Eriksson & Kovalainen, 2008; Lincoln & Guba, 1985.) Dependability could also be seen as relevant for transferability, as an accurate description of the research setting is needed to conclude if there are such irregular variables in the research setting that make the study difficult to repeat or compare results to. In this thesis, dependability is pursued by the methodology chapter, which documents and describes all parts of the research process, from design to empirical data.

Confirmability addresses if the findings are truly based on the data through neutral interpretation, and not affected by personal or third-party interests or biases. In another way, could somebody else with the same amount of theoretical knowledge arrive at similar conclusions? Conformability and dependability components are closely linked together, as thorough documentation of the analytic process helps to evaluate if the findings offer an accurate representation of the research data. (Lincoln & Guba, 1985.) In this thesis, confirmability was pursued by using both theoretical and method triangulation. The results are interpreted and compared with multiple ethical theories, which means that there is no distinct theory that the researcher would be biased to guide the interpretation towards, to achieve analytic generalization discussed earlier.

6 Results

6.1 Overview

To sufficiently understand the results, the reader is recommended to be familiar with the scenarios presented in the individual interview form, which is found at appendix 1. Below is a summary of each scenario and the underlying moral problem, which provides sufficient background knowledge to interpret the results presented in this chapter.

Scenario 1: Whether to hit three persons in the AV's original trajectory or to swerve towards a single bystander. Is the AV allowed to sacrifice innocent third parties to minimize overall harm?

Scenario 2: Whether to sacrifice three AV passengers or to hit an innocent pedestrian. What is the hierarchy between harm minimisation and prioritization variables?

Scenario 3: Whether to sacrifice a single AV passenger, or to hit a single innocent pedestrian? What is the hierarchy between the passenger and other parties in a one-to-one scenario?

Scenario 4: Whether to sacrifice a single AV passenger or to hit three pedestrians jaywalking. What is the hierarchy between legal responsibility, priorisation and harm minimisation?

Scenario 5: Whether to hit a child in the original trajectory or to swerve and hit an elderly person. Is the AV allowed to make active prioritization based on personal characteristics?

Overall, there was a lot of variation between individual answers, and the motivations behind them. All participants were surprisingly rational in their choices, which mattered more for obtaining meaningful insights. Despite seemingly different positions, the motivations were generally based one of three distinct, over-arching themes: harm confinement, harm predictability and moral acceptance. *Harm confinement* means that the harm caused by the AV should be confined to only certain participants, either based on quality or quantity. The second main theme, *harm predictability*, where respect for rules and system transparency were seen as the backbone for safe AV traffic. These two themes are interlinked, as confinement enhances the systems predictability and vice versa. The third theme, *moral acceptance,* covered more subjective reasoning either at an

individual or societal level. These three themes were not scenario-specific, but larger concepts that offered justification for making certain choices in many different scenarios. As the study set out to research the underlying motivations behind moral choices, the results are presented based on this thematic map, not scenario by scenario.

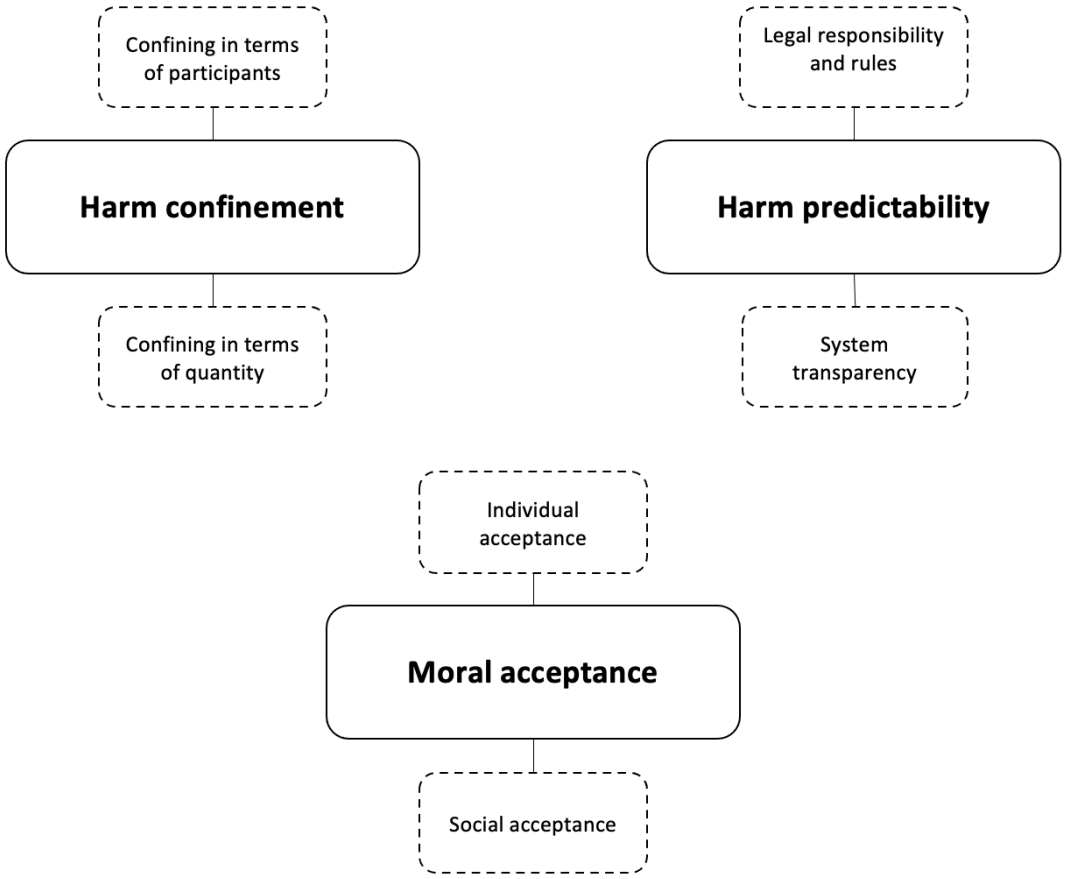


Figure 2 - Thematic map of the findings.

6.2 Harm confinement

The first major theme is harm confinement. It was the overarching concept in most of the answers to scenarios 1-3, where the trade-off revolved around active intervention, harm minimisation and prioritization. The participants expressed two kinds of approaches and interpretations to harm confinement: confinement in terms of participants, and confinement in terms of quantity. These approaches were concluded to be mutually exclusive, as participant and quantity confinement require different hierarchical orders to level of intervention, harm minimisation and prioritization. A good example of this is

scenario 2 in the interview, in which it's not possible to confine the harm exclusively to the AV passengers and achieve the least amount of harm at the same time.

6.2.1 Confinement in terms of participants

All participants agreed that the AV should be given the capability to harm people if it couldn't be avoided. It was generally reported as an immoral option to forbid active intervention altogether, if a more optimal solution could be achieved within the boundaries of other moral variables. However, they did not agree that all participants can be equally subjected to harm. It was generally agreed that the harm must be confined to a predefined and limited amount of traffic participants. Only the scope of this confinement differed among participants. In this context, confinement is defined as limiting AVs options for harm distribution to only certain, pre-defined parties. According to P1, this is needed to avoid a future where anybody, regardless of involvement, can get hit by an AV. From the individual answers, two different scopes of confinement can be identified. At the wider scope approach, harm could only be assigned to the active participants in the original trajectory or scenario. This was the prominent motivation for moral decisions in scenario 1, where the optimal overall solution in terms of harm quantity required the AV to steer towards a pedestrian outside the original situation. Three out of five participants chose to hit the three persons in the original trajectory, arguing that active intervention should only be allowed within AVs intended domain, the road. Participants feared that not limiting the AVs option to its immediate surroundings would lead to a scenario where anybody could become a victim of an AV in the sake of optimisation. This could happen if unrestricted harm minimisation was set as the primary variable to make decisions in a crash scenario. Especially P1 expressed deep concern about traffic turning chaotic if some form of harm confinement is not implemented. She claimed that pre-determined and clearly communicated harm confinement was a prerequisite of predictability, which she saw as the backbone for overall traffic safety with AVs present.

> "Traffic as a whole would become totally unpredictable, if you can get hit by an AV in any situation, in any role, regardless of your actions. -- I think in the beginning it would be a good idea to mark the AVs somehow so that other participants can better prepare for possible uncertain behaviour" P1

The notion of the importance of predictability was agreed by P2 and P5, where P5 presented the second, stricter approach for harm confinement. In scenario 2, he opted for

sacrificing the 3 AV passengers instead of a single pedestrian, motivating this choice with the following argument:

> "I think the risk should always be limited to the AV user, rather than other traffic participants. From a societal standpoint, it makes more sense for the AV driver to accept the risk, than for every pedestrian to think that they might get hit by an AV "missile" at any moment". P5

A supporting statement was presented by P1 when discussing whether an AV user should accept more harm in exchange for the realized benefits.

> "If you grab an AV from your yard, you know that you are stepping in a vehicle that acts and optimises in a certain way. But if you are a bystander, you can't know that you are encountering an AV. That is why I thought that involvement should be a consideration. Again, I am talking about predictability. You have made a conscious decision when you board an AV, but not when you are crossing a road." P1

The main idea behind these arguments was that the AV passenger is the only participant able to make an active decision to subject himself to using an AV, and the risks that come along. All the other traffic participants do not make an active choice to be in a traffic scenario with an AV, nor are they obtaining any benefits from them. It is therefore morally unacceptable to subject them to harm and risks should be confined to only those parties that are getting the benefits, *the AV users*. This approach was favoured especially in scenario 3, where there was no need to regard for harm quantity. Four out of five participants chose to sacrifice the AV passenger in this case. Only P2 opted for an egoistic solution, and he failed to provide any credible reasoning behind this preference, and later gave up on this position in the DEGI. When asked why AV passengers should bear more responsibility despite the lack of control, all participants disagreed with the notion that the lack of a steering wheel somehow would make the passenger not accountable. It was jointly expressed by P1, P3, and P5 that although active decisions can't be made while in the AV, the moral responsibility is obtained when choosing to use the AV in the first place.

> "Moral responsibility is transferred to the passenger when he buys the AV. But this requires that he is aware of the moral code behind it. So if a person uses an AV that he knows to sacrifice other people for him, I think he actually has quite a big moral responsibility" P1

6.2.2 Confinement in terms of harm quantity

The second way to interpret harm confinement was in terms of total harm quantity, prioritizing the harm minimisation variable. This solution was mostly offered as a solution for scenario 2, where three out of five participants decided based on total harm quantity, while two favoured confining the harm to the AV users. As in this scenario, the pedestrians were not outside the original situation, it was seen as more morally justified to save a greater amount of people:

> "How I see it, if there are no other variables present in a situation, they are all in the same line here. And I think it's a smarter solution to choose one injury over three" P3

This position was opposed strongly by P4 and P5, who thought that harm confinement to AV users was imperative and did not depend on the amount of total harm. With some participants, the discussion went towards splitting the harm between participants, as proposed by Karnouskos (2021). Both P3 and P4 saw it as a favourable solution, on the condition that the AV would have the technical capability to know for sure that the result would be two mild injuries. This topic was also discussed with P5, who stood firmly with the argument that risks should always be assigned only to AV users, and any kind of optimisation that required harm to be distributed to outside parties should not be allowed.

6.3 Harm predictability

The second major theme identified from the data was harm predictability. Although many preferences were motivated by harm confinement, to many participants it was only to achieve predictability. Based on its direct and indirect prevalence in the interview answers, predictability is perhaps the single most preferred attribute of a functioning AV ethical setting. Based on the interviews, a predictable AV must 1) be able to enforce and distinguish lawful behaviour and 2) have a transparent decision-making logic. These two sub-themes can be described also as external and internal predictability. Meaning that an AV must be predictable in terms of reactions, as well as actions.

6.3.1 External predictability: Enforcing rules and personal responsibility.

External predictability means that an AV should be predictable in a way that it reacts to other participants actions. This makes it easier to adapt one's own behaviour to the AVs around, as their probable response to your own traffic behaviour is known. One could compare this to knowing that most probably, cars will stop for you when crossing the street with a green light. Related to this, perhaps the most interesting finding of the study was the heavy emphasis on legal responsibility in a situation. It was regarded as the most influential moral variable by four out of five participants, and. Although discussed more further on, it should be mentioned that also P4, who initially disagreed, changed her position in the group interview. Legal responsibility was the only variable with a strong consensus among participants before the group interview. Based on the answers to scenario 4, being able to account for legal responsibility was seen as the bedrock of safe and predictable AV traffic. Even though many participants had been strict on confining harm to AV users, there was a clear consensus that if you are responsible for the situation, you lose your position as an innocent third party. Even P5, who was the most vocal for the stricter scope of confinement, regarded personal responsibility as the ultimate variable:

> "I see that your role as a bystander is taken away when you actively decide to break the law, when personal liability steps in" P5

Although P5 changed his views, his answers were still rational. On scenarios 2 and 3, he argued that confining harm to an AV user is morally justified because he willingly subjects himself to risk. By this logic, a person breaking the law is also an active decision-maker subjecting to risk and could therefore be assigned harm just as justifiably as the AV user.

> "I think that people have the responsibility of their own traffic behaviour. If you break them (laws), you subject yourself vulnerable to an accident, just like at this moment (with manual drivers)." P1

Besides the unlawful pedestrians losing their protection, accounting for legal responsibility was seen as paramount for maintaining overall predictability in traffic. The programming of AVs was seen as impossible if we couldn't assign them clear rules of what other parties can and can't do in traffic. As discussed with P5, it would create a situation where the possible outcomes for each scenario will be unlimited, if the AV

would have to account for all lawful and unlawful behavior. This is could especially be a programmatical problem for a rule-based, deontological system.

> "When we are talking about programming, there are variables that are so qualitative, that we can't program them. If you try to build an AV system, at least to my knowledge, the base must be a set of rules. If they are just moral situations, there are too many variables" P5

There was also an interesting notion from P3, who stated that it would be easier to define system requirements if AVs could only react to actions within traffic rules, that are limited and describable to the system.

> "Let's think it in this way. We have a system that is in any case deficient and highly vulnerable. If we add in that other parties do not follow any sort of rules, it will not work even that much. I think from the perspective of a functioning system, it would be even more important to follow the laws than it is now" P3

P1 and P5 also thought that it is fundamentally flawed that AVs wouldn't address legal responsibility. AV saturation was seen to have a positive effect on how we view rules, if we acknowledge that by adhering to them, we can ensure that we are on the right side of their decision-making logic. P5 felt that AV prioritization rules would dictate how all the other participants would behave in traffic.

> "Let's assume that AVs get more numerous and that at some point, they are the norm. Of course, then their prioritization (settings) will determine a lot of out traffic rules and behaviour." P5

Similar arguments were presented by P1 from the opposing viewpoint, but with the same fundamental idea: if lawful action does not affect your safety level, it would create a moral hazard for people to act unlawfully, resulting in less predictable traffic.

> "Traffic would become a total wild west if you can't think that you can be in a safer position by adhering to traffic rules. It decreases predictability and erodes the meaning of traffic rules." P1

P1 and P5 also both noted that we perhaps haven't realised that a lot of issues covering AV ethics are also addressed by current traffic laws. P5 used an example of legal responsibility, which is the basis on which we assess who to punish in an accident with manual drivers. P1 presents a supporting argument, hinting that AV ethics research is perhaps trying to invent the wheel again on many occasions:

> "Current legislation is, or at least I hope it is, formed through the moral views of people and ethics. Why would we drastically change this ethical code just because AVs are introduced? I think we should not do this. We get to solve new moral problems, but we should not forget old, already agreed-on principles" P1

Accounting for legal responsibility thus seems to already have wide societal support, and there seems to be little justification for why this shouldn't be a requirement for an AV ethical setting.

Based on the results, harm confinement and predictability are closely linked together. The legal responsibility component of predictability adds a third layer to confinement, that sits between the confining to scenario participants or to the driver alone. This middle ground confines the harm not to all parties who can make active decisions to subject themselves to risk. More predictability would then arise from the fact that to be subjected to harm by an AV, you'll have to make an active decision by breaking the traffic rules. If you do not, risks will be confined to the AV user by default.

6.3.2 Internal predictability: system transparency

The second type of transparency was internal transparency, or predictability in terms of actions. It is achieved by enforcing system transparency, meaning that the AVs decision-making model is visible to all other stakeholders. Two types of transparency were identified to play a role in moral decisions: transparency to third parties, and transparency to users. The first one, transparency to third parties, also relates to external predictability. It simply means that to safe in traffic with AVs present, other traffic participants need to be aware of the decision-making principles that are used. The second identified dimension of system transparency was transparency for the user, which was recognized to solve a lot of the issues regarding the possible responsibility gap during AV crashes. As it was discussed regarding harm confinement, even though the AV user is not in control, all participants saw that he can be assigned moral responsibility as he has made a conscious decision. But important notes were given by P1 and P3 on the conditions that are needed to hold the AV user accountable. Essentially, the driver can only make a moral decision when getting into AV, if he is provided a chance to understand the decision-making logic behind the system. Transparency is therefore a prerequisite for confining harm to the AV user. Also, having a clear description of the decision-making rules was seen by P3 to prevent a situation where a person can "hide" behind the complexity of the AV, and claim that he is not morally responsible as the system can't be interpreted by a normal person.

> "I think that the user should be informed on what kind of system he is using. Otherwise, the responsibility, or at least the power, is retained by the one making the program. Then, the user can't escape the responsibility by claiming that he does not understand how the system works." P3

Transparency was also seen as a requirement for effective regulation of AV technology. An outcome-based system was generally shunned by the participants, with the fear that it would make the AV a black box to everyone except technical experts.

> "Regulation will always be avoided if there are profits to be made. What is preventing the (AV) companies from increasing their profits by cutting a few corners inside a black box that the end customer will never see." P5

6.4 Moral acceptance

The third theme was moral acceptance. Although many answers could be assigned to confinement and predictability, which represented more rational stances on how an AV should make decisions, many participants still based their views on more subjective moral standpoints. Not surprisingly, this theme was mostly present in scenarios that included any type of prioritization. From the discussions, two dimensions of moral acceptance were recognised. Certain attributes of AVs result either in technological acceptance or rejection, and this is referred as individual moral acceptance. On a wider scale, there is societal moral acceptance, which in the interviews covered mostly regulatory topics.

6.4.1 Individual moral acceptance

Certain features of an ethical setting can result in either moral acceptance or downright technology rejection. Based on the interviews, transparency and regulation were seen as positive attributes that would enhance individual moral acceptance. It was stated by many participants that the credibility of an AVs moral setting would increase if they knew that it was thoroughly planned and represented the collective social agreement. When asked if she would accept an ethical setting that is given from above, P1 gave the following opinion:

> "If I would see that when coming from above, this (the ethical setting) is considered broadly from a juridical, ethical, and philosophical standpoints, and it is a commonly agreed policy that the AVs would use, then yes, I would accept it. It doesn't necessarily matter if it comes from above or what, but that it (AV behaviour) is based on commonly agreed ethical code, which I could also personally agree on." P1

The other thing that was greatly seen to enhance trust in AVs, was system transparency. This was because all participants placed a strong importance on the manufacturer's ability to demonstrate that moral dilemmas can be accounted for. To further motivate this thesis and AV ethics research in general, the participants were asked if they viewed ethical issues as mundane or relevant to their technological adoption. Surprisingly, all participants expressed that answering to rare, but morally loaded dilemmas was seen as mandatory if they would allow themselves or people close to them to use an AV. P3 also recognized the psychological heuristics that make moral problems punch above their weight in people's safety assessments.

> "They (moral problems) are very important in my opinion. If we think about people's sense of safety, it doesn't matter too much what is the probability of something. What matters is if it is possible" P3

As was anticipated based on existing literature, people based their motivation primarily on subjective reasons when assessing whether personal characteristics should be accounted for. So not surprisingly, scenario 5 had the widest discrepancy between individual moral preferences. The first three participants chose the perhaps intuitive answer of prioritizing the child, stating that for them subjectively, it was morally unacceptable to consciously place harm on a child. This was motivated mostly in similar lines with Anderson & Anderson, (2011) , stating that the participant with the longer expected life expectancy should be prioritized. P3 argued that this preference might be intuitive to a lot of people, even though it's not admitted:

> "I think there are a lot of things that are quite widely accepted, but a few people want to say it out loud. I think that most people would choose to hit the elderly if you had to hit somebody. But saying it out loud is quite difficult. But as a scenario, this is quite clear. If a person is 8, he most likely has more life years left than someone who is 80, thus being more "valuable". P3

In addition to this, P2 also presented similar ideas to Gogoll & Müller (2017) , reckoning that the elderly person might also want the child to be prioritized. It's important to note that all the participants agreed that granting an AV the ability to rank people is deeply problematic, but it did not make them change their answers. To P4 and P5 however, it was clear that to actively intervene towards the elderly, it requires the AV to distinguish between elderly and a child in the first place. P4 expressed that any attempt to prioritize should be based on participant quantity, not quality. Both expressed a sense of

technological rejection about an AV that can rank people in real-time, and to P5, the reasoning behind this was much deeper than just AVs:

> "I think that If that sort of technological possibility exists (to rank in real-time), it raises a lot of questions about surveillance and other things. If you can put people in a hierarchical order this fast, then there has to exist a database that raises a lot of questions about privacy. I think this is not possible, if there is not a totalitarian surveillance system in place" P5.

It was concluded that when there are emotional considerations in a scenario, people place much more emphasis on individual moral acceptance than on the more rational reasons described in this chapter.

In the interview form, Rawlsian ethics were included in two ways. First, for each scenario, the participants were asked if they would agree on their answer also from the original position. Also on one occasion, they were asked if they would agree on the harm minimisation to the most vulnerable party, which was the primary position of Rawls regarding harm distribution. All the participants did agree on their views from the original position, but that is just an indicator of their sincerity, not a merit for Rawlsian ethics. When it comes to prioritizing the worst off, no participant agreed. As it is evident by now, the primary values of a good AV system, harm confinement, and predictability, are not allowed by a Rawlsian ethical setting. Even the ideas of Karnouskos (2021) on splitting the harm were agreed only by two participants, others disagreeing either because they sought harm confinement to the AV user alone, or because they distrusted the system's ability to distribute the harm evenly.

6.4.2 Societal moral acceptance

Societal moral acceptance means that to achieve commercial success, AVs must align with the average moral preferences of a larger public. As discussed in the background chapters, this will mostly be a matter of regulation, as both existing research and the interview results suggest that a customisable ethical setting is not a feasible solution. Three out of five participants claimed customizable settings to be immoral, one hoped for minor settings to be customisable. Only P2 argued for customisability, as he felt that it's unlikely that a quick solution is available for a commonly agreed ethical setting. To realise the safety benefits of high AV saturation in traffic, some leeway must be granted to the users. He also presented a valid argument that "customisable" ethical settings are essentially in place and permitted by current legislation, as people make decisions based

on intuition and reflexes.

In the middle, there was P4 who aligned with the thoughts of Soltanzadeh et al. (2020) , saying that mild customisation options could be permitted to increase societal acceptance. The other participants expressed strong disapproval towards customisation. Once again, the driving factor was predictability, which according to P3 was the primary reason why a universal ethical setting should be implemented:

> "It has to be the same for everybody. I have two arguments, not in a particular order. The other is that it (AV traffic) has to be as predictable as possible, as an AV passenger you have to be aware of how other AVs are behaving. Another is that the world is full of bad people. If you could customise it, there would be cars in traffic that couldn't care less about other people." P3

It seemed that all participants had general distrust towards the choices of other people in a regulatory vacuum, which draws parallels to the prisoner's dilemma presented earlier in this thesis. To P5, the major cause for regulation was not to restrict users but to force the manufacturers to respect the social consensus on moral preferences. He presented similar ideas to Liu & Liu (2021) and Gerdes & Thornton, (2016), saying that profit chasing of AV companies could result in making driver protection a feature in premium products, which would undermine the whole idea of harm confinement.

> "If we are talking about regulation, it must hit the manufacturer. Preventing at the source. it might turn out to be a slippery slope if we always punish the driver (in the case of an AV accident). Then, the biggest benefit for creating this kind of harmful action goes to the manufacturers. It can still profit, and has no incentive to stop" P5

Essentially, what P5 is talking about is the same kind of moral hazard that would be realized among other traffic participants if AVs fail to address legal responsibility. There might be not enough incentives to force the manufacturers to make their products safer if the responsibility can be delegated to the user who "decided on the settings". P1 also expressed similar thoughts when asked how failing to answer moral dilemmas would affect her moral acceptance. She argued that regulation is a way to ensure that society's moral consensus is embedded into the AVs system requirements.

> "I feel that if we don't address ethical and moral issues from the start, it's hard to do it afterward. In my opinion, we should think about them precisely now in the development phase, so that we don't let the situation end up so that there are AVs where these (ethical concerns) are not considered. After that, I think we would lose control of the situation" P1

6.5 Moral average from DEGI

The goal of the discourse ethical group interview, referred as DEGI, was to achieve at least a majority consensus on the moral variables and how an AV should reach in each scenario. It's more feasible to base the proposed ethical policy and setting on the average preference, than to try match all individual opinions. It also makes it easier to compare the preferences with the existing ethical theories and regulation. There was no full consensus on any scenario before the group interview, only partial which comes as natural in a group of five. With discourse ethics, a full consensus was achieved for harm minimisation, prioritization and legal responsibility (scenarios 2-4). This can be regarded as a good result, as the group was divided on all topics before the interview, as demonstrated by the answer table below, showcasing how the participant's positions changed from the individual interviews. From the table we can see how the answers remained the same only in scenario 1.

	Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5	
	Before/After		Before/After		Before/After		Before/After		Before/After	
P1	B	B	B	A	A	A	B	B	B	A
P2	B	B	B	A	B	A	B	B	B	A
P3	A	A	B	A	A	A	B	B	B	B
P4	A	A	A	A	A	A	A	B	A	A
P5	B	B	A	A	A	A	B	B	A	A

Table 5 - Individual answers before and after the DEGI.

From DEGI, we obtained the moral average preference for an AV. From these preferences, we can formulate system requirements for a publicly accepted AV ethical setting. Chapter 3 introduced the German Act on Autonomous Driving (GAAD), along with three normative ethical theories that are most prevalent in AV ethics literature. Now, we'll compare the moral average obtained with DEGI to these theories, to see how they would fit our system requirements, as well as current AV legislation. In below table is a rough summary of this comparison.

Moral variable	The moral average	Utilitarian setting	Deontological setting	Rawlsian setting	GAAD
Active intervention	An AV must only actively intervene inside the original scenario	Negative. Can't make a distinction between involved and uninvolved	Positive. Prohibits using people as a means to achieve an end	Negative. Can't make a distinction between involved and uninvolved	Mixed. Prohibits "offsetting victims against each other
Harm minimisation	An AV must confine harm instead of minimise it	Negative. Can't fit pre-determined and quality-based harm distribution	Positive. Different scopes of harm confinement can be implemented as a rule if they are agreed upon	Negative. Can't fit pre-determined and quality-based harm distribution	Mixed. Mentions both sacrificing fewest lives and not demanding self-sacrifice form the user
Prioritization	An AV must prioritize the other party by default	Negative with same reasons as with harm minimisation	Positive with same reasons as with harm minimisation	Negative with same reasons as with harm minimisation	Mixed with same reasons as with harm minimisation
Legal responsibility	An AV must take legal responsibility into account	Negative. Can't account for qualitative variables such as legal responsibility	Positive. Can be applied as an conditional rule or a decision tree. Also, deontology is based on commonly agreed "law".	Negative. Can't account for qualitative variables such as legal responsibility	Positive. Acknowledges that parties creating the risks must have more responsibility.
Personal characteristics	AVs are not allowed to prioritize based on personal characteristics	Positive. Can't account for qualitative variables such as personal characteristics	Positive. Can allow both accounting and not accounting for personal characteristics, whatever is agreed as "right".	Positive. Can't account for qualitative variables such as personal characteristics	Positive. Precisely forbids any sort of prioritization based on personal characteristics

Table 6 - Comparing results to existing ethical theories and regulation.

6.5.1 Level of intervention

For active intervention, the moral average was that an AV can actively intervene, but only within the original scenario. The moral average solution of sacrificing the three pedestrians for the innocent bystander was chosen by 3 out of 5 participants, the smallest majority of the 5 scenarios. Also notable was that this was the only topic where no

participant switched their positions after the discourse, meaning that subjective opinions may play a strong role in how active intervention is perceived. Nevertheless, the results validate that people regard harm confinement as a central requirement for an AV. We conclude that an AV should not be allowed to distribute harm outside its intended domain, and so at least the wider scope of harm requirement is taken as a system requirement based on the groups preferences.

The only ethical theory that can provide support for this approach is deontology, as both utilitarianism and Rawlsian ethics fail to make a distinction between involved and non-involved parties. Also on a theoretical level, deontology precisely prohibits using people as a means to achieve optimal results. From a legal standpoint, GAAD does not make a clear distinction between involved and non-involved parties. It does state that it's "*prohibited to offset victims against one another*", which could be interpreted as sacrificing those in the original trajectory. Although this statement leaves a lot for interpretation, GAAD should recognise some form of harm confinement to better align with the moral average.

6.5.2 Minimizing overall harm

For harm minimisation, the moral average was that an AV should not minimise harm, but rather confine it. A full consensus for this solution was achieved through discourse. A total of three participants shifted from their original position, which was the most out of any scenario. It was also one of two scenarios where the majority opinion was shifted during the interview from one solution to another. It was commonly agreed that as AV passengers are getting all the benefits, they should also bear the risks associated with them. The group also placed a heavy emphasis on the fact that AV users are also the only party making an active decision to acquire these risks and benefits. As the group achieved a full consensus, we can credibly extend the confinement requirement to its stricter form. This means that in a vacuum with no secondary variables, the moral average prefers an AV not to minimise harm, but to confine it primarily to the AV user.

This requirement further erodes the validity of both utilitarian and Rawlsian ethical settings, as they can't allow for pre-determined harm distribution. Confinement also requires the AV to take participant quality as a determinant for harm distribution, but these theories can only provide support for quantity-based distribution. For Rawlsian ethics, it is plausible that the AV user would rarely be the worst off. But as this answer can't be given beforehand and for certain, Rawlsian ethics do not fit the requirement.

Again, the deontological ethical setting is capable of supporting this preference, as it should be easily implemented as a rule that the AV must always prioritize other traffic participants. Deontology also from a theoretical level gives support to this, as it forbids using third parties as a means to achieve an end. This would mean that an AV can't use uninvolved parties, such as pedestrians, to save the life of the AV user. Of course, someone could argue that this logic should also go vice versa, and you couldn't use the AV user to save the other parties. To this, I would argue that as it requires a conscious and active decision to use an AV, the user can't be regarded as the "uninvolved" party, if there are no secondary variables present. The inner hierarchy of harm minimisation is therefore established: in the absence of secondary variables, prioritization is above harm minimisation. On the legal front, GAAD has no clear position to these requirements. It states that "*It would appear reasonable to demand that the course of action to be chosen is that which costs as few human lives as possible*.", but also acknowledges that self-sacrifice can't be required from drivers. So there seems to be a discrepancy between the moral average and current legislation.

6.5.3 Prioritization

For prioritization, the moral average was that in a one-to one scenario, an AV should always prioritize the other party. A full consensus was also reached on this topic, although this result is not surprising after the stricter confinement scope was agreed on in the previous scenario. Before the discourse, only P2 was advocating for driver protection, and he too shifted to favour harm confinement. Other participants did agree with P2 argument that self-sacrifice is a lot to ask, and it may hinder the adoption rate and subsequent benefit realisation from AVs. Still, it was regarded as morally impossible to sacrifice the passive party in an accident scenario, and harm should again be confined to the one taking the conscious risk. This is the second time the group achieved a full consensus, further validating stricter harm confinement as a requirement for an AV ethical setting. This demand is again best answered by a deontological ethical setting, for much of the same reasons described with harm confinement. A strict prioritization rule could be implemented as an Asimovian rule-based system, where the AV could only target third parties if they couldn't be saved by sacrificing the user. For utilitarianism, it starts to get evident that its inability to assess different roles makes it unfit for matching the moral average preferences that revolve heavily around participant quality rather than quantity. Also theoretically, in a one-to-one scenario with equal probabilities of injury,

utilitarianism could fail to decide if it only seeks harm minimisation. Rawlsian ethics share the same basic problem with utilitarianism, although its again acknowledged that in most cases, the uninvolved party is likely to be the more vulnerable one. But as this answer can't be given outright and beforehand, it doesn't match the requirement. From a legal perspective, there are many of the same problems that were described in the previous scenario. GAAD includes statements in both directions and thus remains neutral, or slightly biased to driver protection. As no clear answer can be given, there seem to be discrepancies between legislation and the moral average.

6.5.4 Legal responsibility

For legal responsibility, the moral average was that legal responsibility must influence the AVs decision. This was the third topic where the discourse ethics led to a full consensus. Beforehand, only P4 advocated for a utilitarian solution, but she changed her opinion after the argument rounds. As she had already changed her opinion from a utilitarian solution to harm confinement in scenario 2, this was a rational move. The participants agreed that not being able to account for unlawful behaviour will erode the meaning of traffic rules in general. This will lead to a moral hazard as there are no incentives to act safely and predictably around AVs. This predictability, both from AVs themselves and the predictability of their response to other participants' actions was seen as invaluable for safe traffic. It's important to note that legal responsibility was a *secondary variable.* This means that if present in a situation, shifts in legal responsibility will also alter the established requirements for harm confinement. If the other party acts unlawfully, harm confinement should shift from the AV passenger to the unlawful party. The improved requirement for harm confinement is therefore to *confine harm to the responsible party.* By default, this is the passenger. However, the interview results validate that it is also morally acceptable to instead distribute harm to the one who created the risky situation. This makes harm distribution dynamic, as it will depend on scenario-specific variables. The deontological setting is the only one that can take legal responsibility into account, as utilitarianism and Rawlsian ethics fail to consider such a qualitative variable. Legal responsibility could be programmed a decision three where the default confinement to the user changes based on legal responsibility. GAAD is in line with the moral average, stating that "*Those parties involved in the generation of mobility risks must not sacrifice non-involved parties.".* This fits our final confinement rule above, as it does not demand the AV to make a distinction between parties, but by responsibility.

The AV user would likely to be the responsible one by default, but GAAD would nevertheless allow for dynamic harm confinement based on legal responsibility.

6.5.5 Personal characteristics

The moral average for personal characteristics was that AVs must not be allowed to take personal characteristics, such as age, into consideration. Not surprisingly, the toughest discussion was held on this topic. Along with harm minimisation in scenario two, this was the only time that the preliminary moral average changed. Initially, P1-3 advocated for favouring children, and P4-5 opposed it strongly. After the discourse, both P1 and P2 changed their position. This was mostly due to P5's argument that granting an AV the capability to classify people in real-time possesses severe direct and indirect moral dilemmas. The direct dilemma was the ability to classify itself, as there could be no guarantees that once hierarchical assessment is allowed, discriminatory algorithms wouldn't emerge. The indirect dilemma was that for this kind of system to be technically feasible, it would need a database that would breach people's privacy severely. With a majority consensus of 4 out of 5, AVs should not be able to prioritize based on personal characteristics. This means that this variable will be discarded, legal responsibility being left as the only secondary variable.

The demand to not decide based on personal characteristics is answered by all three ethical theories. Whereas utilitarian and Rawlsian systems by nature won't recognize personal characters as a variable, a deontological setting could fit whatever is agreed as a commonly accepted rule. On the legal front, things are also clear this time, as GAAD precisely forbids any sort of prioritization based on personal characteristics. As discussed in the background chapters, many studies have found opposing results for this variable. I argue that this is primarily due to the use of quantitative or structured qualitative methods. It was observed in the individual interviews that it was intuitive to answer that you want children to be prioritized. But when this position had to be defended in DEGI, both P1 and P2 abandoned their original positions the arguments against this type of prioritization were more convincing. If our group was able to stay rational and disregarded the perhaps intuitive answer of saving children as they understood the moral dilemmas that come with this capability, maybe there is also hope in forming a larger consensus on this issue.

	Utilitarianism	Deontology	Rawls	GAAD
Active Intervention	Negative	Positive	Negative	Mixed
Harm minimisation	Negative	Positive	Negative	Mixed
Prioritisation	Negative	Positive	Negative	Mixed
Legal responsibility	Negative	Positive	Negative	Positive
Personal charachteristics	Positive	Positive	Positive	Positive

Table 7 - Summary of the comparison results.

7 Discussion

7.1 Results

The objective of this study was do find out how laypeople, or regular future users, would prefer an AV to solve moral problems. The study introduced several ethical theories that current literature has proposed as the foundation for AV ethical setting, as well as the variables that are likely to be relevant for AVs. The idea was to use the same variables and scenario design with existing studies but place more focus on the motivations behind these preferences. System requirements could then be found by discovering larger, overarching themes on why people prefer certain decisions. Based on these goals, two research questions were formed:

1. *How do laypersons prefer an AV to solve moral problems?*
2. *How these preferences could be formed into an ethical setting?*

The empirical data was gathered using two methods: individual interviews and a discourse ethical group interview. The idea was that a wide array of individual preferences and motivations could be found with deep one-to-one interviews. The group interview tried to see if a consensus could be found within these various individual preferences, which could then be formulated into an ethical policy, our collective agreement on which ethical principles an AV should base its decisions. (Liu & Liu, 2021.) This policy dictates the system requirements for the ethical setting. It is the actual technical implementation of an ethical policy, consisting of a theoretical foundation, decision-making model and the moral variables. The next sections will go over the propositions for an ethical policy and setting based on the results.

7.1.1 Proposed ethical policy

As described above, ethical policy consists of general, more abstract values that contribute into a socially acceptable AV. The actual ethical setting consisting of decision-making models and normative ethics may be overly complicated for an everyday user, so an ethical policy is needed to inform the public how the AVs will be making decisions. (Liu & Liu, 2021.) Based on the research results, the following two requirements consist of our proposed ethical policy:

An AV must:

1. *Confine harm to the responsible party.*
2. *Act in a predictable manner, both in terms of actions and reactions*

First rule describes how confinement was found to be more influential than harm minimisation. This represents a new focal point for AV ethics research, as this concept is not mentioned in existing literature. Our results indicate that instead of *harm minimisation* as an absolute value, an AV should instead limit the *harm exposure* to pre-determined parties. The participants suggested several scopes of confinement as an option to limit harm exposure. In the widest scope, an AV should only be able to distribute harm to the participants in the original scenario. It was argued that this is the minimum requirement for harm confinement, as without it anybody could be subjected to harm in the sake for optimization. If harm wouldn't be confined at least to this scope, anyone could theoretically get hit by an AV, and thus everyone would need to constantly be aware of this danger. It was seen as rational to limit the amount of people that need to need to assume the risks of AVs at least to this level. Going back to theoretical origins of this thesis and the original trolley problem, the participants of this study recognized that there indeed is a big difference between letting someone get hurt, and actively harming another

In the stricter scope, an AV would confine harm primarily to the AV user alone, without any exceptions. This was agreed on with a full consensus on scenarios where there were no secondary variables in play. But as assessing legal responsibility was seen vital with a full consensus, this is not the final form of confinement. It was concluded that confinement should depend on whether someone is able to make an active decision in the scenario, whether it is to use an AV or break the law. Confinement to user is the default, but the harm is distributed to the responsible party if legal responsibility variable is present. This means that ultimately, the confinement is based on responsibility, not by role. This will further limit the number of participants that have to consider the risks to only those who can decide to act on them.

The second requirement in the ethical policy is predictability, which is divided into two types. First one is predictability in terms of actions, or internal predictability. This is primarily based on system transparency. An AVs ethical setting and subsequent decision making-system must be understandable to the user at least in terms of general working principles. This will make the users default responsibility more justifiable, as he

can’t hide behind not understanding the system. This internal predictability also allows other traffic participants to be aware of the decision-making logic of AVs. The participants also expressed that internal predictability requirement will act as a form of regulation, as AV developers can’t hide forbidden functions behind complex algorithms.

The second type is predictability in terms of reactions, or external predictability. It should be clear to other participants how an AV will react in respect to their action. From the results, this is concluded to be primarily a result of acknowledging legal responsibility. If a person knows that the first ethical policy confines harm to the responsible party and that the AV can assess legal responsibility, the AV will seem predictable. All a person must do is follow the law, and the risk is confined to the user. This enforcement of rules was also seen as a positive outcome of AV based traffic, as it’s likely to shape the behavior of all other traffic participants safer around AVs.

7.1.2 Proposed ethical setting

After establishing the general principles on how an AV should act as an ethical policy, it’s time to decide on the technical and theoretical foundations in the form of an ethical setting. Based on the results and the derived ethical policy, a following ethical setting is proposed:

1. Theoretical foundation: Deontology
2. Decision-making model: Rule-based
3. Moral variables: Active intervention, priorisation, legal responsibility

Based on the results, both utilitarianism and Rawlsian ethics proved to be unfit to match the preferences of the study participants. The very fundamentals of these ethical theories make it hard to consider different participant roles as a determining factor in harm distribution. Also, legal responsibility and predictability were deemed as core values for a publicly accepted AV, and neither of these theories can assess them as variables in harm distribution. This leaves deontology as the only viable option. As clearly seen from the summary table presented in the end of last chapter, it is clearly the best fit as a theoretical foundation from the three options. It can address all variables as long as they are agreed to be righteous by a consensus and refuses to see harm quantity as a deciding variable. Also, the idea of “not using a person as a means” matches the idea of confinement and responsibility well. It suggests that deontology does not allow using uninvolved parties

to save the lives of those who willingly took the risks. The choice of deontology as a theoretical foundation also dictates that a rule-based decision-making model should be utilized. This serves the user preference for predictability. It was mentioned by Karnouskos (2021), that if an AV would be targeting a certain outcome without any hard-coded rules, the decisions can seem random or unpredictable to the users. A deontological, rule-based system therefore directly answers to the ethical policy's requirement for predictability.

As the participants expressed many scopes of confinement and there are also secondary variables that make the rules dynamic, a suitable solution could be to adopt a rule hierarchy as proposed by Geisslinger et al. (2021). As mentioned in the chapter 3, rules would be categorized into obligatory, permissible, and forbidden actions, and these would limit the possible trajectories that an AV can choose.

To finish the ethical setting, it needs to be considered which moral variables the decision-making model should include to provide an accurate representation of reality and answer to the ethical policy. As the established ethical policy does not allow an AV to decide based on harm quantity, it is logical to drop harm minimisation from the model. This means that the main variables left in the final model are active intervention and prioritization. Active intervention, originally meaning if an AV should be allowed to optimize at all, now refers to the minimum requirement for confinement. Meaning, that any action outside the original scenario and participants is classified as forbidden in the rule hierarchy. Prioritization is now dependent of the legal responsibility. It is the only secondary variable in the model, as personal characteristics were seen as an immoral variable to be considered by an AV and dropped out of the final model.

To conclude, the proposed ethical setting for a publicly accepted ethical setting is based on deontology and utilizes a rule-based decision-making model. It's working logic includes three moral variables: *active intervention* only inside the original scenario, where *prioritization* is based on *legal responsibility*. The proposed ethical setting is a major shift from utilitarian harm minimization. As many original variables were left out of the model and it does not have any resemblance in the existing literature, this ethical setting is defined as the *harm confinement model (HCM*). It represents an entirely new approach to

AV decision making. Let's consider the following scenario, to see how theorize how HCM would solve a scenario with all three variables:

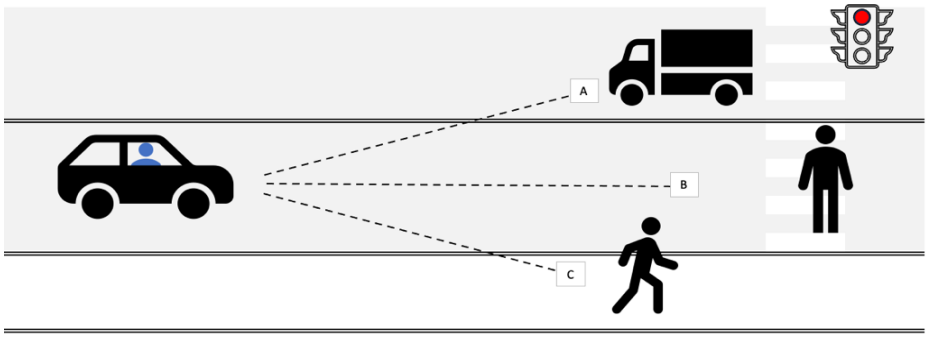


Figure 3 - An example scenario combining interview scenarios 1,3 and 4.

In this scenario, there are three possible trajectories that can be described with the hierarchy logic of Geisslinger et al (2021) . Trajectory C is forbidden, as it's outside minimum confinement scope of the ethical setting. This limits the AV's options to obligatory trajectory A and permissible trajectory B. In the absence of secondary variables, A would be chosen by default. But as the established ethical policy was to confine harm to the responsible party, an AV based on the proposed ethical setting would choose trajectory B, as the pedestrian is the responsible party. Below is a graphical representation of the proposed ethical setting and decision-making model.

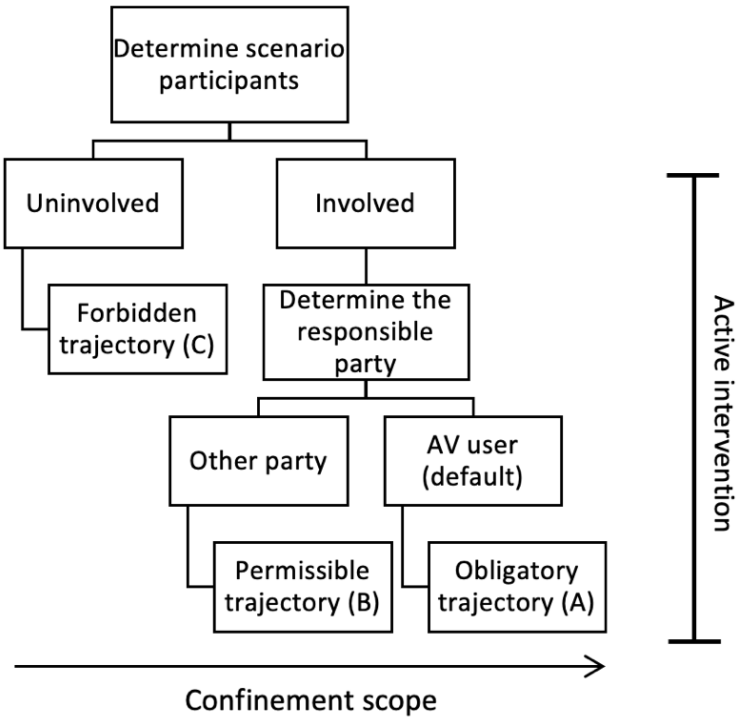


Figure 4 - The Harm Confinement Model (HCM).

If regulated into a universal ethical setting, the harm confinement model would also solve the prisoner's dilemma presented in chapter 4. Although the example compared an egoist and harm minimising model, the creators of the model did not advocate for any specific solutions, but rather argued against a setting that would distribute harm primarily to others (Gogoll & Müller, 2017). The prisoner's dilemma could likely be avoided also by implementing a confinement model, as the suboptimal result was caused by the availability of an egoistic setting that distributed more harm to others than itself. A confinement strategy would likely yield similar results in avoiding a prisoner dilemma, especially when it can be considered as the opposite to an egoistic one.

From a legislative standpoint, there is some mismatch between GAAD and HCM. Although it was stated that legal responsibility must influence decisions, GAAD also argued that self-sacrifice can't be demanded from the AV user. As this is the default option in HCM, the proposed ethical setting does not currently fit to regulation. But GAAD also expressed many statements to the other direction and towards deontological values. It does not take a clear stance on any of the three primary variables, so the problem lies more in the ambiguity of the regulation than in the model itself.

7.2 Contribution

There were some similarities to existing studies such as the Moral Machine Experiment, suggesting that people do not tend to favour egoistic AVs (Awad et al. 2018). The study extended this finding to first-person decision-making perspective, opposed to the results of Bonnefon et al. (2016). Results suggest that if people are rational in their moral positions, there shouldn't be variation based on decision perspective. Despite these similarities, the main results present a substantial divergence from the existing studies, suggesting an entirely different focal point for AV ethics research and development. As presented in the background chapter, majority of existing studies have concluded that people prefer harm minimisation and utilitarian optimisation. But as described thorough this thesis, the structured nature of many of these studies has not allowed discovering any larger themes that affect people's preferences, thus portraying people's views black and white, tied to the options presented by the researcher. By allowing participants to motivate their answers freely, an explorative research approach allowed this study to counter the simplification posed on it by the trolley problem.

On a theoretical level, the main contribution of this thesis is the discovery of general values that make up a good ethical setting: confinement, and predictability. The study results should not be understood as the only way to achieve these values, but nevertheless it suggests that there are some core qualities that make an AV appeal to people. This offers a more stable premise for future studies both in terms of research design and analysis, as there are more dimensions to consider than just utilitarianism versus egoism. On the empirical side, main contribution of this study is the validation of discourse ethics as a tool for AV ethics research, and perhaps also in development of a real-world ethical policy. Although the sample size in this study was small, discourse ethics was still able to achieve full or nearly full consensus on topics where there was substantial polarisation before the interview. It allowed for the "survival of the fittest" of moral positions, and there is no reason why it wouldn't work also in a wider context in developing a societal consensus.

Another important finding that directly opposes current research was also the role of personal characteristics and age-based prioritization, which was preferred in many studies such as the MME. The results suggest that there are such direct and indirect moral problems associated with qualitative prioritization that this capacity should not be given to an AV. It is likely that the structured nature of previous studies made it easier to advocate for this, as the discourse ethical interview demonstrated that any need to defend this preference quickly undermined its validity. The study therefore suggests that in future research, this moral variable is not considered as an option, at least not in studies where there is no need to defend such an immoral position. This also provides validity for existing legislation, as the results showed that the participants could not defend quality-based prioritization in rational discourse, something that would also happen when forming a policy on this issue.

7.3 Limitations and future research

The obvious limitation of this study was the sampling, both in terms of quality and quantity. Although the group expressed a variety of preferences and motivations, they are represented a similar cultural background and education level. Also, although there was some variation in terms of age and gender, additional validity could have been added by having four age brackets instead of two. The utilisation of two methodologies and

extensive interviews proposed limitations on participant quantity, so there was no room extend the variation also to education level and cultural background. In future studies, it would be interesting to see how these factors correlate with the moral preferences. Other clear limitation was how scheduling issues forced the DEGI to be held online, which naturally altered the dynamics of a group discussion. It can't be said for sure how this affected the discussion. It might have been easier to defend a position online, as the non-verbal communication is not so prevalent. It would have been interesting to see if for example, the only remained position for age-based prioritization would also held in face-to face discourse. In future research, discourse ethical interview should be held in this manner, to further validate this tool for AV ethics research.

Future research should also place emphasis on the core themes of this study. Confinement should be added as a main variable alongside with harm minimisation, to allow for more options to choose from. Especially the preference of minimisation versus confinement should be studied, as most studies have only focused on people being either altruists or egoists. There should also be studies done on perceived predictability of AVs and what factors contribute to this. This could be achieved for example, describing the decision-making logic of an AV beforehand to the participants, and having them conclude the AVs reaction to their action, to see if there are more factors at play than just legal responsibility found in this study.

8 Conclusions

The research set out to explore how a publicly accepted ethical setting for an autonomous vehicle (AV) could look like. It presented different ethical theories, decision-making models and moral variables that could help solve moral issues faced by AVs. The objective was to see how laypersons, the future users, would prefer an AV to solve moral dilemmas that it might encounter in everyday traffic. Data was gathered using two methods, both of which contributed to the discoveries. Individual interviews allowed the researcher to discover the full variety and breath of how individuals prefer an AV to make moral decisions. A discourse ethical group interview allowed to formulate the proposed policy and setting on with the moral average, and showcased how a consensus could be achieved on sensitive moral issues.

Despite the limitations, most of which are related to the small and culturally homogenous sample size, this thesis proves that ethical issues are more than just a mundane aspect of AV development. Developers should not underestimate the psychological significance of moral issues and understand that a large mismatch between public preference and AVs ethical setting may severely hinder adoption rates. A model based on harm minimisation seems to not fit people's preferences, as most variables present in moral dilemmas are so qualitative that they can't be minimised using mathematical optimisation. But as proved in this thesis, the preferences of laypeople are surprisingly rational, and there are ways to achieve a consensus on these topics. These are the rational foundations which people base their moral preferences on, and these should dictate how AVs should be developed to be publicly acceptable. Future research could aim to further validate confinement and predictability using quantitative methods

The results and main theoretical contribution of this thesis can be summarised in one sentence: *a publicly accepted ethical setting will harm the user by default, but harm can be distributed to another party if one was present and responsible in the original situation*. This approach does not share the theoretical foundations of many previous studies that see utilitarian harm minimisation as the most viable option. The study also showed that peoples preferences are more grounded to rational concepts, confinement and predictability, rather than individual moral beliefs. These two concepts were used to formulate an ethical policy and ethical setting based on user preferences. The main theoretical contribution of this thesis therefore shifted from offering deeper motivation of existing preferences, to presenting a new approach for AV decision making with the harm

confinement model (HCM). It is to the researcher's understanding the first model that focuses on different roles and their inherent differences in responsibility, rather than in harm as a quantifiable variable. In conclusion, this thesis suggests that utilitarian solutions, where harm is minimised as an absolute value, are not actually preferred by users. Instead of seeing harm minimisation as the goal and harm quantity as the primary value to consider, the focus should be on qualitative factors such as participants role and legal responsibility, and harm should be confined, not minimised.

References

Anderson, S. L., & Anderson, M. (2011). A Prima Facie Duty Approach to Machine Ethics and Its Application to Elder Care. In *Workshops at the twenty-fifth AAAI conference on artificial intelligence*.

Asimov, I. (1940). *I, Robot*. Narkaling Productions.

Awad, E., Anderson, M., Anderson, S. L., & Liao, B. (2020). An approach for combining ethical principles with public opinion to guide public policy. *Artificial Intelligence*, *287*, 103349.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Bergmann, L. T., Schlicht, L., Meixner, C., König, P., Pipa, G., Boshammer, S., & Stephan, A. (2018). Autonomous vehicles require socio-political acceptance—an empirical and philosophical perspective on the problem of moral decision making. *Frontiers in Behavioral Neuroscience*, *12*. https://doi.org/10.3389/fnbeh.2018.00031

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2019). The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars. *Proceedings of the IEEE*, *107*(3), 502–504. https://doi.org/10.1109/JPROC.2019.2897447

Bonnefton, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vechiles. *Science*, *352*(6293), 1570–1573.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Coca-Vila, I. (2018). Self-driving cars in dilemmatic situations: an approach based on the theory of justification in criminal law. *Criminal Law and Philosophy, 12*(1), 59–82. https://doi.org/10.1007/si1572-017-9411-3

Contissa, G., Lagioia, F., & Sartor, G. (2017). The Ethical Knob: ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law*, *25*(3), 365–378. https://doi.org/10.1007/s10506-017-9211-z

Deontological ethics. (2023). In Editors of Encylopedia Britannica (Ed.), *Encyclopedia Britannica*. https://www.britannica.com/topic/deontological-ethics

Driver, J. (2022). History of Utilitarianism. In *The Stanford Encyclopedia of Philosophy* (Winter 2022). Metaphysics Research Lab, Stanford University.

Eriksson, P., & Kovalainen, A. (2008). *Qualitative Methods in Business Research*. SAGE Publications Ltd. https://doi.org/10.4135/9780857028044

Ethics Guidelines For Trustworthy AI. (2019). High-level expert group on artificial intelligence set up by the European commission ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Faulhaber, A. K., Dittmer, A., Blind, F., Wächter, M. A., Timm, S., Sütfeld, L. R., Stephan, A., Pipa, G., & König, P. (2019). Human Decisions in Moral Dilemmas are Largely Described by Utilitarianism: Virtual Car Driving Study Provides Guidelines for Autonomous Driving Vehicles. *Science and Engineering Ethics*, *25*(2), 399–418. https://doi.org/10.1007/s11948-018-0020-x

Feldman, F. (1978). *Introductory ethics.*

Flood, M. M. (1958). *Some Experimental Games* (Vol. 5, Issue 1).

Foot, P. (1967). The Problem of Abortion and the Doctrine of the Double Effect.

Geisslinger, M., Poszler, F., Betz, J., Lütge, C., & Lienkamp, M. (2021). Autonomous Driving Ethics: from Trolley Problem to Ethics of Risk. *Philosophy and Technology*, *34*(4), 1033–1055. https://doi.org/10.1007/s13347-021-00449-4

Gerdes, J. C., & Thornton, S. M. (2016). Implementable ethics for autonomous vehicles. In *Autonomous Driving: Technical, Legal and Social Aspects* (pp. 87–102). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-48847-8_5

Ghauri, P., Grønhaug, K., & Strange, R. (2020). *Research Methods in Business Studies*. Cambridge University Press. https://doi.org/10.1017/9781108762427

Gill, T. (2021). Ethical dilemmas are really important to potential adopters of autonomous vehicles. *Ethics and Information Technology*, *23*(4), 657–673. https://doi.org/10.1007/s10676-021-09605-y

Gogoll, J., & Müller, J. F. (2017). Autonomous Cars: In Favor of a Mandatory Ethics Setting. *Science and Engineering Ethics*, *23*(3), 681–700. https://doi.org/10.1007/s11948-016-9806-x

Goodall, N. (2014). Ethical decision making during automated vehicle crashes. *Transportation Research Record*, *2424*(1), 58–65. https://doi.org/10.3141/2424-07

Guba, E. G. (1981). Criteria for Assessing the Trustworthiness of Naturalistic Inquiries *ERIC/ECTJ Annual Review Paper,* 29(2), 75-91. https://about.jstor.org/terms

Gunn, J., Littlejohn, S.W. and Foss, K.A. (2009), Encyclopedia of Communication Theory, SAGE, pp. 27-30

Habermas, J. (1996). Between facts and norms- contributions to a discourse theory of law and democracy. *Trans. Rheg William, MIT Press, Cambridge, MA.*

Holstein, J. A., & Gubrium, J. F. (1995). *The active interview* (Vol. 37). Sage.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291.

Kant, I. (1981). *Grounding for the metaphysic of morals*.

Karnouskos, S. (2021). The role of utilitarianism, self-safety, and technology in the acceptance of self-driving cars. *Cognition, Technology and Work*, *23*(4), 659–667. https://doi.org/10.1007/s10111-020-00649-6

Kauppinen, A. (2021). Who Should Bear the Risk When Self-Driving Vehicles Crash? *Journal of Applied Philosophy*, *38*(4), 630–645. https://doi.org/10.1111/japp.12490

Knaapi-Junnila, S., Rantanen, M. M., & Koskinen, J. (2022). Are you talking to me? – calling laypersons in the sphere of data economy ecosystems. *Information Technology and People*, *35*(8), 292–310. https://doi.org/10.1108/ITP-01-2021-0092

Kopecky, R., Jirout Košová, M., Novotný, D. D., Flegr, J., & Černý, D. (2023). How virtue signalling makes us better: moral preferences with respect to autonomous vehicle type choices. *AI and Society*, *38*(2), 937–946. https://doi.org/10.1007/s00146-022-01461-8

Koskinen, J., Knaapi-Junnila, S., Helin, A., Rantanen, M. M., & Hyrynsalmi, S. (2023). Ethical governance model for the data economy ecosystems. *Digital Policy, Regulation and Governance, 25*(3), 221–235. https://doi.org/10.1108/DPRG-01-2022-0005

Kriebitz, A., Max, R., & Lütge, C. (2022). The German Act on Autonomous Driving: Why Ethics Still Matters. *Philosophy and Technology*, *35*(2). https://doi.org/10.1007/s13347-022-00526-2

Krügel, S., & Uhl, M. (2022). The risk ethics of autonomous vehicles: a continuous trolley problem in regular road traffic. arXiv:2206.03258

Kumfer, W., & Burgess, R. (2015). Investigation into the role of rational ethics in crashes of automated vehicles. *Transportation Research Record*, *2489*, 130–136. https://doi.org/10.3141/2489-15

Leben, D. (2017). A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology*, *19*(2), 107–115. https://doi.org/10.1007/s10676-017-9419-3

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic Inquiry*. Sage.

Liu, P., & Liu, J. (2021). Selfish or Utilitarian Automated Vehicles? Deontological Evaluation and Public Acceptance. *International Journal of Human-Computer Interaction*, *37*(13), 1231–1242. https://doi.org/10.1080/10447318.2021.1876357

Martinho, A., Herber, N., Kroesen, M., & Chorus, C. (2021). Ethical issues in focus by the autonomous vehicles industry. *Transport Reviews*, *41*(5), 556–577. https://doi.org/10.1080/01441647.2020.1862355

Millar, J. (2015). Technology as Moral Proxy: Autonomy and Paternalism by Design. *IEEE Technology and Society Magazine*, *34*(2), 47–55. https://doi.org/10.1109/MTS.2015.2425612

Mingers, J., School, K. B., & Walsham, G. (2010). Toward ethical information systems: the contribution of discourse ethics. *MIS quarterly, 833-854.*

Mordue, G., Yeung, A., & Wu, F. (2020). The looming challenges of regulating high level autonomous vehicles. *Transportation Research Part A: Policy and Practice*, *132*, 174–187. https://doi.org/10.1016/j.tra.2019.11.007

Normative ethics -- Britannica Online Encyclopedia. (2023). Encylopedia Britannica. https://www.britannica.com/topic/normative-ethics

Nyholm, S. (2018a). The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass*, *13*(7), e12507. https://doi.org/10.1111/phc3.12507

Nyholm, S. (2018b). The ethics of crashes with self-driving cars: A roadmap, II. *Philosophy Compass*, *13*(7), e12506. https://doi.org/10.1111/phc3.12506

Paulo, N. (2023). The Trolley Problem in the Ethics of Autonomous Vehicles. *The Philosophical Quarterly*, *73*(4), 1046–1066. https://doi.org/10.1093/pq/pqad051

Porpora, D. V. (2009). Sociology's causal confusion. In *Revitalizing Causality* (pp. 195–203).

Rawls, J. (1971). *A Theory of Justice*. http://ebookcentral.proquest.com/lib/kutu/detail.action?docID=1367826.

Ross, A., & Chiasson, M. (2011). Habermas and information systems research: New directions. *Information and Organization*, *21*(3), 123–141. https://doi.org/10.1016/j.infoandorg.2011.06.001

Sandberg, A., & Bradshaw-Martin, H. (2013). What do Cars Think of Trolley Problems: Ethics for Autonomous Cars. *Beyond AI: Artificial Golem Intelligence*, 12.

Santoni de Sio, F. (2017). Killing by Autonomous Vehicles and the Legal Doctrine of Necessity. *Ethical Theory and Moral Practice*, *20*(2), 411–429. https://doi.org/10.1007/s10677-017-9780-7

Savulescu, J., Kahane, G., & Gyngell, C. (2019). From public preferences to ethical policy. In *Nature Human Behaviour* (Vol. 3, Issue 12, pp. 1241–1243). https://doi.org/10.1038/s41562-019-0711-6

Schäffner, V. (2021). Between Real World and Thought Experiment: Framing Moral Decision-Making in Self-Driving Car Dilemmas. *Humanistic Management Journal*, *6*(2), 249–272. https://doi.org/10.1007/s41463-020-00101-x

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operational Research*, *177*(3), 1333–1352. https://doi.org/10.1016/j.ejor.2005.04.006

Soltanzadeh, S., Galliott, J., & Jevglevskaja, N. (2020). Customizable Ethics Settings for Building Resilience and Narrowing the Responsibility Gap: Case Studies in the Socio-Ethical Engineering of Autonomous Systems. *Science and Engineering Ethics*, *26*(5), 2693–2708. https://doi.org/10.1007/s11948-020-00221-5

Stebbins, R. (2001). *Exploratory Research in the Social Sciences*. SAGE Publications, Inc. https://doi.org/10.4135/9781412984249

Sunstein, C. R., & Llewellyn, K. N. (2003). Terrorism and Probability Neglect. In *The Journal of Risk and Uncertainty* (Vol. 26, Issue 3).

Tan, W. (2017). *Research Methods: A Practical Guide For Students And Researchers*. WORLD SCIENTIFIC. https://doi.org/10.1142/10699

Thomson, J. J. (1984). The trolley problem. *Yale LJ*, *94*.

Topham, G. (2021, January 3). 'Peak hype'- why the driverless car revolution has stalled. *The Guardian.* Retrieved 15.12.2023.

Westerstrand, S., Koskinen, J., Lähteenmäki, C., Rantanen, M., Cumini, A., Haapakoski, T., & Välimäki, A. (2023). *Työhyvinvointia tuottava digiloikka - Opas ihmislähtöisen automaation ja datatalouden edistämiseksi.* https://www.sitra.fi/tulevaisuussanasto/

Yin, R. K. (2002). *Case Study Research*. Thousand Oaks, CA: Sage.

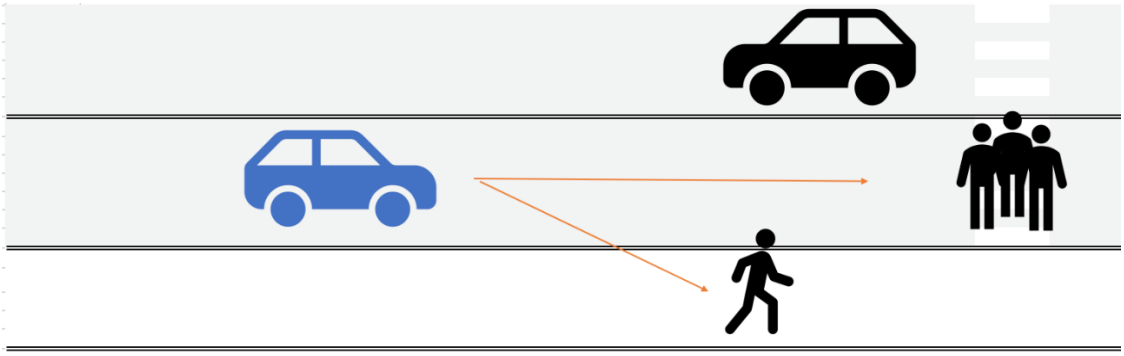
Appendices

APPENDIX 1 - The interview form (without page breaks)

Instructions:

1. Read the scenario description. Don't try to imagine a "better" way that could be done in a real situation, as these scenarios are simplified versions of reality to allow us to discuss AV ethics. Only available trajectory options are what is in the paper.
2. Choose the more preferred answer. No answers are inherently right or wrong, what matters is how you motivate your answers.
3. Motivate why you think this answer is more morally acceptable to you *personally.* The motivations are the main materials for this study, not your initial answer.
4. The interviewee will present a set of sub-questions depending on your answer. This process should not be understood as "questioning" your original position but as a tool for adding more depth to discussion.
5. If something is unclear, ask immediately instead of answering based on false assumptions.

Scenario 1: The intervention dilemma



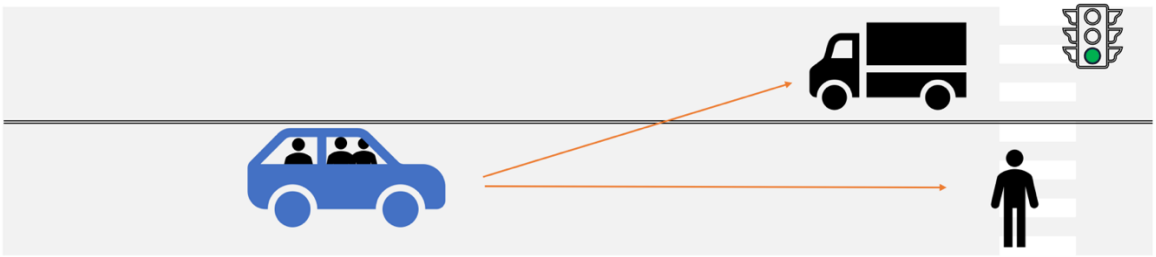
You are a passenger in an autonomous vehicle, driving through a busy mid-town street. Suddenly, three pedestrians cross the street on a crosswalk, which was painted to the street yesterday without communicating this to AV mapping software. The car calculates that there is not enough time to fully stop before the crosswalk. The AV now has two options:

A) Swerve left into the sidewalk, with a *possible* chance of injuring a single bystander.
B) Hit the three pedestrians, with serious injury to at least some of them.

The tradeoff: Should the AV be allowed to make an active decision to divert some harm to the bystander, even though he was not part of the original situation? This would result in less overall harm, as hitting the three pedestrians leads to guaranteed injuries.

Moral variables: Active intervention, harm minimization

Scenario 2: The Utilitarian dilemma



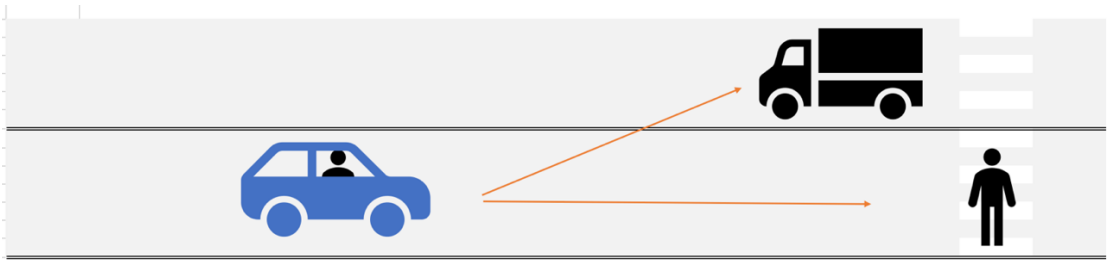
You are riding in an autonomous vehicle with two other people. The traffic light is malfunctioning and sends the AV a signal that the pedestrian light is red. Then, a single pedestrian walks to the crosswalk, but the AV has too much speed due to the faulty traffic light. There is a large truck on the other line. The AV now has two options:

A) Hit the single pedestrian with a serious injury to him.
B) Swerve left and hit the truck, with serious injuries to all AV participants.

The tradeoff: Should the AV prioritize the pedestrian as he is not at fault, or minimize the total amount of harm by protecting the three passengers?

Moral variables: Harm minimization, participant prioritization

Scenario 3: The prioritization dilemma



You are riding in an autonomous vehicle when suddenly a jogger runs to the sidewalk without paying any attention to traffic. He is not breaking traffic laws, but appeared from nowhere, so the car didn't have time to slow down. There is a large truck on the other line. The AV now has two options:

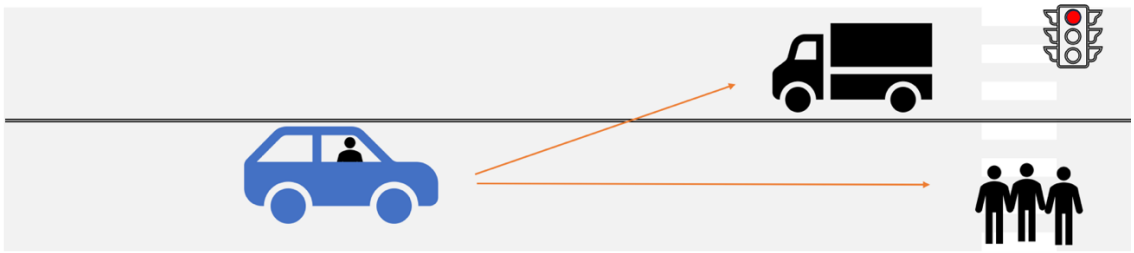
A) Hit the single pedestrian with a serious injury to him.

B) Swerve left and hit the truck, with serious injuries to you.

The tradeoff: Who should be prioritized in a one-to-one scenario, the AV passenger, or a pedestrian?

Moral variables: Harm minimization, participant prioritization

Scenario 4: The jaywalker dilemma



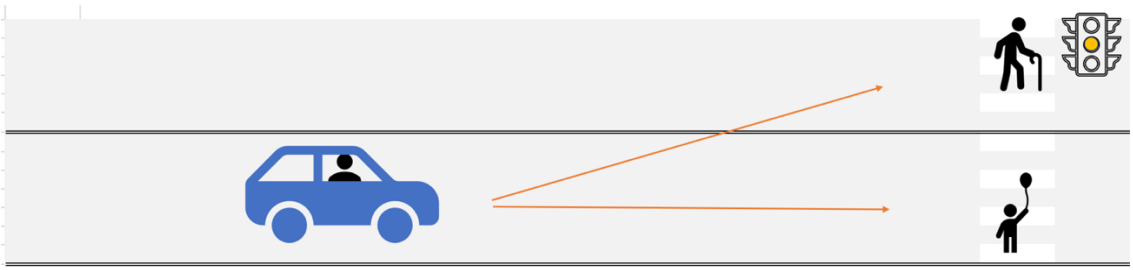
You are riding in an autonomous vehicle on a busy city street. Suddenly, three people jaywalk in front of the car, clearly violating traffic rules. The vehicle doesn't have time to stop before the crosswalk. Swerving to avoid them would result in the AV hitting a large truck in the oncoming lane. The AV now has two options:

A) Hit the three pedestrians, with serious injuries to all of them.
B) Swerve left and hit the truck, with serious injuries to you.

The tradeoff: What is the mutual hierarchy of harm minimization and legal responsibility? Should the three men bear more harm because they are violating traffic rules?

Moral variables: Harm minimization, participant prioritization and legal responsibility

Scenario 5: The age dilemma



You are in an autonomous vehicle driving through a quiet residential area. Ahead, you notice an elderly person and a child waiting at a crosswalk to cross the road on different sides of the road. Suddenly, the traffic light malfunctions, and both start crossing the road. Car sensors pick this up, but it doesn't have enough time to stop completely, and there are too many people on the sidewalk to drive there. The AV now has two options:

A) Stay on course and hit the child, with severe injuries to him.
B) Swerve left and hit the elderly person, with severe injuries to him.

The tradeoff: Should the AV prioritize the life of the children, even though it required active intervention by assigning more harm to the elderly person not in the original trajectory. Also, prioritizing the child requires the AV to have the ability to actively scan and rank people in real time.

Moral variables: Personal charachteristics, active intervention

Open question 1: *Would you want an AV to have a regulated ethical setting that is the same for everybody, or should you be able to customize it according to your preferences? Why?*

Open question 2: *What are your biggest concerns about autonomous vehicles as a traffic participant?*

APPENDIX 2 – DEGI instructions (in Finnish)

Diskurssieettinen työpaja – Ohjeet fasilitaattorille ja osallistujille

Säännöt osallistujille:

1. Luo omalla toiminnallasi turvallista, toisia arvostavaa ja myönteistä ilmapiiriä. Tämä onnistuu suhtautumalla avoimen kiinnostuneesti ja kunnioittavasti jokaiseen osallistujaan. Tarkoituksena on edistää yhteistä hyvää, ei voittaa. Jos haluat voittaa, jonkun on hävittävä – tämä taas ei edesauttaisi yhteistyötä.

2. Puhu selkeästi ja ymmärrettävästi, vältä erikoistermejä ja ammattikieltä. Jokaisen osallistujan näkökulmat ovat yhtä arvokkaita. Tarkoituksena on tulla tietoiseksi niistä, oppia yhdessä sekä hyödyntää yhteistä ymmärrystä palvelujen kehittämisessä, ei korostaa omaa tietämystä.

3. Esitä ajatuksesi tiiviisti, mieti miten se edistää asiassa etenemistä ja tavoitteeseen pääsyä. Keskity myös muiden kuuntelemiseen, silloinkin kun muiden ajatukset eroavat omistasi. Näin osoitat kunnioittavasi toisia, heidän aikaansa ja aikataulussa pysymistä.

4. Perustele kantasi, etenkin jos sinulla on vahvoja väitteitä tai mielipiteitä. Pysy samalla avoimena myös toisenlaisille näkemyksille ja niiden perusteille.

5. Osallistu keskusteluun avoimin kortein, vilpittömästi ja rehellisesti, ilman kätkettyjä tavoitteita. Tämä on tärkeää keskinäisen luottamuksen rakentamiseksi, erilaisten näkemysten hyödyntämiseksi ja todellisen yhteisymmärryksen (konsensuksen) saavuttamiseksi.

Fasilitaattorin rooli: Fasilitaattori ohjaa keskustelua neutraalisti kohti tavoitteita, mutta ei itse osallistu aktiivisesti keskusteluun. Hän saa kommentoida, jos sen tarkoituksena on pitää keskustelu asiallisena tai aiheessa. Jokaisen osuuden jälkeen fasilitaattori kertaa, mistä ja miten puhuttiin, ja mihin lopputulokseen päädyttiin.

Diskurssieettisen työpajan tavoitteet:

Tavoitteena on tuoda jokaiseen skenaarioon kaikki perustellut näkökulmat, sekä löytää kaikkia miellyttävä ratkaisu argumentatiivisin keinoin. Täyttä konsensusta ei tarvitse saavuttaa, enemmistön suosio riittää. Tavoitteena on

Säännöt ja periaatteet alun perin esitetty Westerstrand et al. (2023.):n raportissa ” Työhyvinvointia tuottava digiloikka - Opas ihmislähtöisen automaation ja datatalouden edistämiseksi”

Diskurssieettisen työpajan kulku:

1. Työpajan alussa fasilitaattori esittelee osallistujat toisilleen, ja käy läpi haastattelun yleiset tavoitteet. Haastateltavat saavat myös nimikortit S1-S5, joilla heidät tunnistetaan.

2. Fasilitaattori jakaa jokaiselle osallistujalle samat materiaalit kuin yksilöhaastattelussa, sekä kertaa kaikki käsiteltävät aiheet (moraaliset muuttujat). Samassa lomakkeessa on myös suostumuslomake haastattelun nauhoittamiseen ja materiaalin käyttöön tutkimuksessa.

3. Haastateltavat kysyvät mahdolliset kysymykset epäselväksi jääneistä aiheista tai menettelyistä.

4. Ensimmäinen skenaario käydään läpi, ja kaikki kertovat vuorotellen vastauksensa ja perustelunsa.

5. Keskustelu jatkuu niin kauan, että yksi vastausvaihtoehto saa 3/5 äänistä. Tämän jälkeen perusteluja käydään läpi ryhmässä diskurssietiikan keinoin. Fasilitaattori ohjaa keskustelua siten, että aiheita käsitellään mahdollisimman syvällisesti.

6. Sama toistetaan niin kauan, että jokaiselle skenaariolle on olemmassa enemmistöä tyydyttävä ratkaisu ja motivointi.

7. Lopputuloksena on siten ”moraalinen keskiarvo” ryhmän preferensseistä, joita aineiston analysointivaiheessa peilataan perinteiseen normatiiviseen etikkaan ja olemassa olevaan lainsäädäntöön.

APPENDIX 3 – Research data management plan

This document will help you plan how to manage your research data. More detailed instructions for each section are available online in the Research Data Management Guide for Students.

1. Research data

Research data refers to all the material with which the analysis and results of the research can be verified and reproduced. It may be, for example, various measurement results, data from surveys or interviews, recordings or videos, notes, software, source codes, biological samples, text samples, or collection data.
In the table below, list all the research data you use in your research. Note that the data may consist of several different types of data, so please remember to list all the different data types. List both digital and physical research data.

Research data type	Contains personal details/information*	I will gather/produce the data myself	Someone else has gathered/produced the data	Other notes
1. Individual interviews - Recording		x		
2. Individual Interviews – Transcribed text document		x		
3. Group Interview - Recording		x		
4. Group Interview – Table of preferred solutions		x		

* Personal details/information are all information based on which a person can be identified directly or indirectly, for example by connecting a specific piece of data to another, which makes identification possible. For more information about what data is considered personal go to the Office of the Finnish Data Protection Ombudsman's website

2. Processing personal data in research

If your data contains personal details/information, you are obliged to comply with the EU's General Data Protection Regulation (GDPR) and the Finnish Data Protection Act. For data that contains personal details, you must prepare a Data Protection Notice for your research participants and determine who is the controller for the research data.

I will prepare a Data Protection Notice** and give it to the research participants before collecting data ☐
The controller** for the personal details is the student themself ☐ the university ☐
My data does not contain any personal data ☒
** More information at the university's intranet page, Data Protection Guideline for Thesis Research

3. Permissions and rights related to the use of data

Find out what permissions and rights are involved in the use of the data. Consult your thesis supervisor, if necessary. Describe the use permissions and rights for each data type. You can add more data types to the list, if necessary.

3.1. Self-collected data

You may need separate permissions to use the data you collect or produce, both in research and in publishing the results. If you are archiving your data, remember to ask the research participants for the necessary permissions for archiving and further use of the data. Also, find out if the repository/archive you have selected requires written permissions from the participants.
Necessary permissions and how they are acquired

Data types 1-4: Written consent form delivered to the participants before starting the research process.

3.2 Data collected by someone else

Do you have the necessary permissions to use the data in your research and to publish the results? Are there copyright or licencing issues involved in the use of the data? Note, for example, that you may need permission to use the images or graphs you have found in publications.
Rights and licences related to the data:

Research does not contain any third-party data.

4. Storing the data during the research process

Where will you store your data during the research process?

In the university's network drive ☐
In the university-provided Seafile Cloud Service ☐
Other location, please specify: ☒

The university's data storage services will take care of data security and backup files automatically. If you choose to store your data somewhere other than in the services provided by the university, please specify how you will ensure data security and file backups. Remember to make sure you know every time where you are saving the edited/modified data.

The data is saved locally on the researcher's computer hard drive, and backup files are loaded to an external hard drive that contains only the research material. No material is uploaded to cloud applications.

If you are using a smartphone to record anything, please check in advance where the audio or video will be saved. If you are using commercial cloud services (iCloud, Dropbox, Google Drive, etc.) and your data contains personal data, make sure the information you provide in the Data Protection Notice about data migration matches your device settings. The use of commercial cloud services means the data will be transferred to third countries outside the EU.

5. Documenting the data and metadata

How would you describe your research data so that even an outsider or a person unfamiliar with it will understand what the data is? How would you help yourself recall years later what your data consists of?

The data is saved as named files and contains only the code of the interviewee and the interview date. All the participants ordered the data to be destroyed after the thesis is completed, so there will be no need to store metadata that will allow for the revision of the data in the future.

5.1 Data documentation

Can you describe what has happened to your research data during the research process? Data documentation is essential when you try to track any changes made to the data.
To document the data, I will use:
A field/research journal ☐
A separate document where the researcher records the main points of the data, such as changes made, phases of analysis, and significance of variables ☒
A readme file linked to the data that describes the main points of the data ☐
Other, please specify: ☐

5.2 Data arrangement and integrity

How will you keep your data in order and intact, as well as prevent any accidental changes to it?

I will keep the original data files separate from the data I am using in the research process, so that I can always revert back to the original, if need be. ☒

Version control: I will plan before starting the research how I will name the different data versions and I will adhere to the plan consistently. ☒

I recognise the life span of the data from the beginning of the research and am already prepared for situations, where the data can alter unnoticed, for example while recording, transcribing, downloading, or in data conversions from one file format to another, etc. ☒

5.3 Metadata

Metadata is a description of you research data. Based on metadata someone unfamiliar with your data will understand what it consists of. Metadata should include, among others, the file name, location, file size, and information about the producer of the data. Will you require metadata?

I will save my data into an archive or a repository that will take care of the metadata for me. ☐

I will have to create the metadata myself, because the archive/repository where I am uploading the data requires it. ☐

No data is stored into a public archive/repository, and therefore there is no need to create any metadata. ☒

6. Data after completing the research

You are responsible for the data even after the research process has ended. Make sure you will handle the data according to the agreements you have made. The university recommends a general retention period of five (5) years, with an exception for medical research data, where the retention period is 15 years. Personal data can only be stored as long as it is necessary. If you have agreed to destroy the data after a set time period, you are responsible for destroying the data, even if you no longer are a student at the university. Likewise, when using the university's online storage services, destroying the data is your responsibility.
What happens to your research data, when the research is completed?

All data is destroyed immediately after completion and approval of the thesis, because: It is demanded by all the interview subjects in the consent form.