MASTERARBEIT | MASTER'S THESIS

Titel | Title

It wasn't me. Responsibility for accidents with autonomous vehicles

verfasst von | submitted by

Mag. Michael Schwendinger BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 641
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Interdisziplinäre Ethik
Betreut von | Supervisor:	Univ.-Prof. Mark Coeckelbergh PhD

Table of contents

Abstract

This Master's thesis contributes to the ethical discussion around self-driving cars, i.e. autonomous vehicles (AV). The focus is on questions of responsibility in case of crashes. I want to place particular emphasis on the relevance of one factor that has mostly been ignored in the discussion to date: speed. For this I will first overview the practical and regulatory status quo as well as the ethical debate around AV. Then I discuss different approaches dealing with responsibility in times of artificial intelligence (AI), followed by an examination of historical, empirical and theoretical accounts of the factor speed within the field of traffic safety. As a result, I wish to demonstrate that the factor speed is inevitably relevant for the ethical discussion on AV as well as of great importance to analyse questions of moral and legal responsibility in case of crashes.

Kurzzusammenfassung

Diese Masterarbeit leistet einen Beitrag zur ethischen Diskussion rund um das Thema selbstfahrende Autos. Der Fokus liegt dabei auf Fragen der Verantwortung bei Unfällen. Besonderes Augenmerk möchte ich dabei auf die Relevanz eines Faktors legen, der in der bisherigen Diskussion weitgehend ausgeblendet wurde: Geschwindigkeit. Dazu gebe ich zunächst einen Überblick über den praktischen und rechtlichen Status quo sowie die ethische Debatte rund um das Thema selbstfahrende Autos. Anschließend diskutiere ich verschiedene Ansätze zum Umgang mit Verantwortung in Zeiten von Künstlicher Intelligenz (KI), gefolgt von einer Betrachtung historischer, empirischer und theoretischer Grundlagen zum Faktor Geschwindigkeit im Bereich der Verkehrssicherheit. Im Ergebnis möchte ich aufzeigen, dass der Faktor Geschwindigkeit für die ethische Diskussion rund um das Thema selbstfahrende Autos sehr relevant ist und auch für die Analyse von Fragen der moralischen und rechtlichen Verantwortung bei Unfällen von großer Bedeutung ist.

Abbreviations

AI … artificial intelligence

N.B.: Since it doesn't make a difference for the scope of this thesis, I use the terms "AI", "robot", "artificial agents" without further specification and somewhat interchangeable – always referring to an (physical or virtual) agent with advanced intellectual capacities based on forms of machine learning, neuronal networks and so forth.

AV … autonomous vehicle(s)

N.B.: I will use the term „autonomous vehicle" (AV) in a very general way throughout this thesis – from manually driven cars with advanced driver assistance systems to fully self-driving cars without human drivers behind the wheel. When necessary for my argumentation, I'll further specify the level of automation. For further information to different levels of automation in the context of AV, see section 1.2.1.

1. Autonomous Vehicles

Devon Miles:	*Welcome on board of the Knight Two Thousand.*
Michael Knight:	*Thank you. What's all this? Looks like Darth Vader's bathroom.*
Devon Miles:	*It is a one of a kind car, Mister Long. It is the fastest, safest, strongest car in the world. It is also completely fuel-efficient and it is operated entirely by microprocessors, which make it virtually impossible for it to be involved in any kind of mishap or collision – unless, of course, specifically so ordered by it's pilot.*
Michael Knight:	*Pilot? Don't tell me this thing flies?*
Devon Miles:	*No. But it thinks.*

(Knight Rider, 1st episode, original release on September 26th, 1982)

In science fiction self-driving cars have already been around for decades. In the meantime they have become reality: self-driving cars are already in use in several cities today and fatal accidents already have happened. As their number and utilisation is expected to increase massively in the near future, the introduction of autonomous vehicles (AV) into our traffic system "can deservedly be considered as the ethical challenge of the century in AI Ethics […]" (Etienne 2021: 106). The short dialogue in the initial quotation from a 1980s cult TV-show perfectly hints at some of the ethically most interesting and pressing questions in this context: Is it true that AV can be designed to avoid any mishap or collision, whereas human drivers are mere risk factors? What role do human drivers actually play in using AV and (when) should they be allowed (or obliged) to overrule an AV's decisions? Are AV really able to "think" and what does this mean in terms of agency – i.e. is the car really driving "by itself" or are there (and should be) always humans in the loop? Finally, what does all this mean in terms of responsibility for accidents – i.e. when is a human driver, user or passenger of an AV in case of an accident rightfully allowed to say: It wasn't me!?

1.1 Research interest and outline of thesis

In this thesis I want to address some of the above-mentioned questions, focussing on responsibility in the context of human-machine interaction in case of using AV. In doing so I want to stress the relevance of one specific factor that is often ignored in the discussion: speed. I want to show that taking into account this factor not only has significant implications on current ethical discussions on AV, but also that it can be used as a practical tool in order to fairly distribute responsibility in case of accidents among involved agents. The basic idea behind my argument is very simple: Speed is the single most important factor when it comes to car traffic safety. Therefore it is key to identify the agent(s) responsible for travel speed, in order to fairly distribute general responsibility in case of accidents with AV among the relevant agents involved – the faster you allow the vehicle to go, the more responsibility you have to bear in case of an accident. My research interest therefore can be captured in the following two questions:

- Who bears responsibility in case of accidents with autonomous vehicles?
- What role can the factor “vehicle speed” play in fairly attributing responsibility among the relevant agents (e.g. driver/user/passenger, manufacturer, software developer etc.)?

To answer these questions, I will proceed as follows: In the first part of the thesis I will portray the status quo – in terms of the most relevant definitions, historical background, the regulatory framework and AV in practical use, and as well give an overview of the academic discourse and different ethical accounts. In the second part I lay the groundwork for my discussion of responsibility. I start with some general remarks on responsibility and then turn my focus to responsibility in the age of automation and AV. Besides moral responsibility I will also briefly touch upon liability and legal responsibility, since these different notions of responsibility are deeply intertwined in this discussion and the legal perspective further helps to point out and clarify some interesting aspects. In the third part I gather the necessary evidence for my claim, that speed is the single most important factor when it comes to traffic safety. I start by describing some historical background of speed limits in general and then discuss the relevance of speed for crashes, as well as vehicle dynamics in crash-situations. In the fourth part I try to bring the pieces together and elaborate the main argument of this thesis in detail. In the last part of the thesis I wrap up my main points and open up the floor for further research in this line of argument.

1.2 Status Quo of autonomous vehicles

The idea of automated passenger vehicles first appeared in the 1930s in comic books and science fiction novels. With the exhibit “Futurama”, Norman Bel Geddes was the first representative of a car manufacturer (General Motors) describing a vision of vast highways with automated vehicles at the New York’s World Fair in 1939. During the 1980s and 1990s professor Ernst Dickmanns from the Bundeswehr University in Munich developed a modified Mercedes van named VaMP, that could drive autonomously in traffic with speeds up to 100 mp/h. In 1995 a research team of Carnegie Mellon University constructed an automated vehicle and travelled from Pittsburgh to San Diego with an average speed of 60 mp/h. 98 percent of the journey was completed by the car itself. In 2004 the "Defense Advanced Research Projects Agency" (DARPA) started a grand challenge for AV, the first long-distance race (150 miles) with a prize of 1 million USD. In the first year of the challenge no car made it to the end. A year later in 2005, 45 vehicles participated in the competition, 5 successfully completed the course. The winner was the project "Stanley", developed by Stanford University. In 2010 the University of Parma developed a van and managed to send it 10.000 miles through 9 countries in 100 days from Parma (Italy) to Shanghai (China) (Wishart et al. 2022). As this very brief insights shows, the idea of AV is not new and also the development of real-life AV already started decades ago.

1.2.1 Levels of driving automation

In 2014 the Society of Automotive Engineers (SAE) launched a taxonomy of 6 different levels in the field of driver assistance systems and automated driving, know as the SAE J3016 levels of driving automation. They range from level 0 "no driving automation" to level 5 "full driving automation" (see figure 1). It has been the industries most-cited standard definition and updated several times in partnership with the International Organization for Standardization (SAE 2021).

It is important to differentiate between driver assistance systems and driving automation. The former includes features such as lane-keeping assistance, lane-departure warning, automatic emergency braking or blind-spot warning, whereby the driver always fully stays in control and the assistance system is monitoring the driving and alerting the driver when necessary. In the later case the vehicle itself takes over control partly or fully and under certain conditions (Wishart et al. 2022). As this taxonomy shows, autonomous or self-driving vehicles start from level 3 upwards. As already mentioned above, I will use the term autonomous vehicle (AV) in a very general way and specify further if this is necessary for the sake of argument.

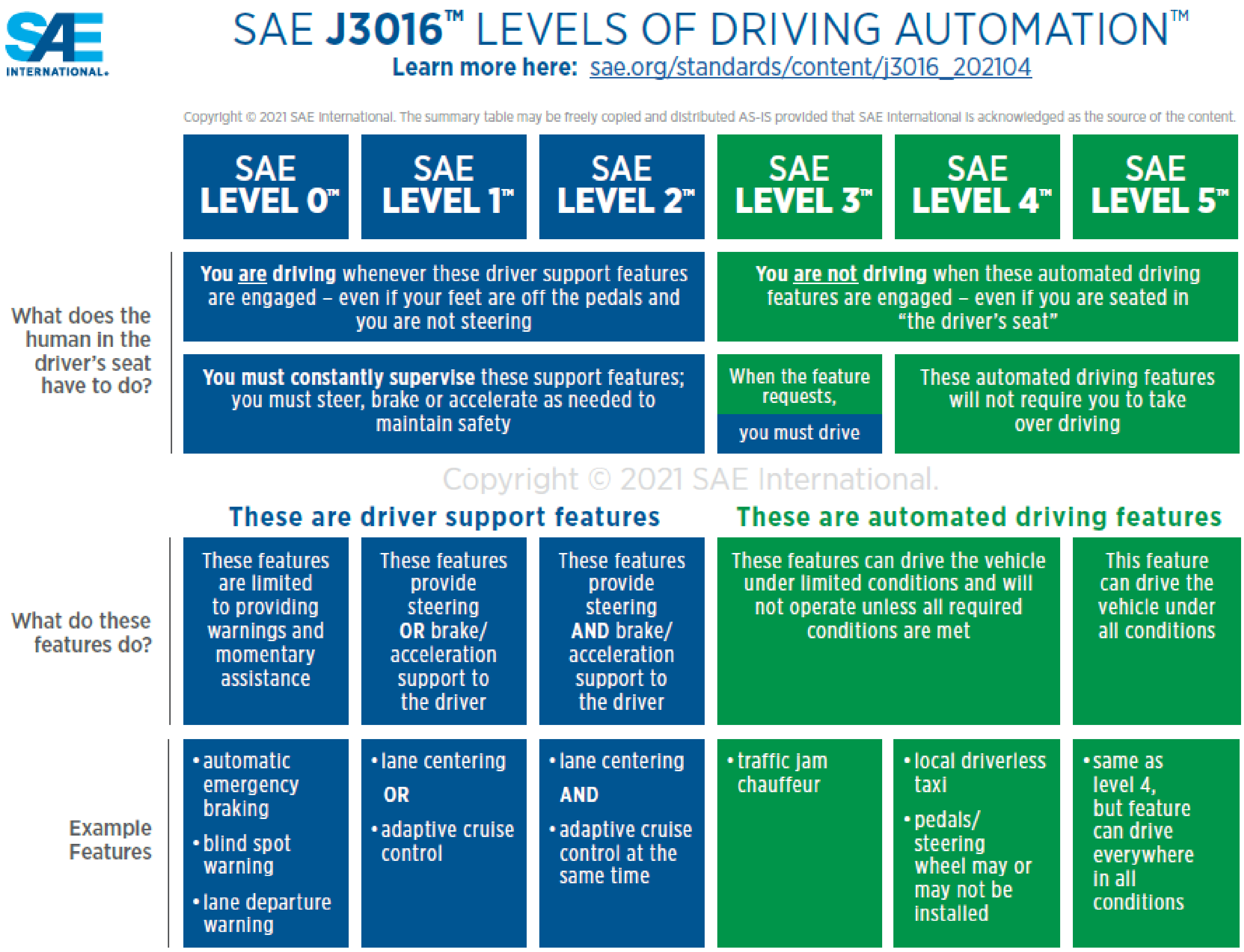

Figure 1: SAE J3016 - Levels of Driving Automation

1.2.2 Autonomous vehicles in practice

The first known fatal accident with a semi-autonomous self-driving car happened on May 7th 2016 in Williston, Florida. A driver put his Tesla in autopilot mode while driving on a highway. The car crashed into a crossing white truck and trailer, which the autopilot failed to identify against the bright sky (Yadron/Tynan 2016). As an official report revealed later, the driver ignored several warnings of the autopilot to put his hands back on the steering wheel, which would have been his legal duty. The last action the driver undertook was to set the cruise control at 74 mp/h, against a legal maximum speed of 65 mp/h, less than two minutes before the fatal crash (Shepardson 2017). About two years later, on March 18th 2018 the first fatal accident with a fully autonomous self-driving car took place in Tempe, Arizona. The autonomous vehicle (AV) operated by Uber with an emergency backup driver behind the wheel killed a pedestrian crossing the street. Although the autonomous operating system reportedly had significant safety flaws and failed to identify the pedestrian, Uber didn't have to face criminal charges (BBC 2019). The backup driver by contrast has been charged with negligent homicide, as she demonstrably didn't pay attention to the street while watching an episode of "The Voice" (BBC 2020). In 2021 more than 50 companies held a license to test self-driving cars with backup drivers on the street solely in the state of California (Pladson 2021).

As to private cars Audi attempted to reach level 3 in 2017, but regulators limited it to level 2. Honda successfully certified a car for level 3 use in Japan in 2021, but only made 100 units available (Jeffs/He 2023). As a big success Mercedes was the first company worldwide to receive a permit for level 3 in two car models in Germany as of May 2022. The function called "drive pilot" is allowed to fully take over control on highways under good weather conditions in traffic jams with a maximum speed of 60 km/h. Whereas Tesla's "full self-driving"-system on level 2 always requires the attention of a human driver, human drivers in a Mercedes car for the first time are allowed to take their eyes from the street and fully transfer driving operations to the car – although they have to be able to resume control within 8 seconds (Hubik/Holtermann 2022). Beginning of 2023 Mercedes also received permission to sell its level 3 cars in Nevada, followed by California in June – which makes it the world's leading company concerning automated private cars, whereas for example Tesla and General Motor so far only reached permission for level 2 (Holtermann 2023).

Commercial robotaxis theoretically have been available since 2017, but only for restricted groups and not for the general public. In 2020 Waymo (owned by Google/Alphabet) started to offer a driverless ride hailing service in Phoenix (Arizona) open to the public. In 2022 driverless taxi services were expanded to multiple cities across the US and China, such as San Francisco, Las Vegas, Beijing and Shenzhen (Jeffs/He 2023). In August 2022 chinese company Baidu was the first to receive a permit to offer commercial driverless robotaxi services, named Apollo Go, on public roads in China. It started out in Chongqing and Wuhan, expanding to Beijing and Shenzhen in 2023, which makes it the world's largest provider of AV ride-hailing services (PR Newswire 2023a). On August 25th 2023 Baidu announced to expanded its robotaxi service Apollo Go to cover the international airport in Wuhan, located 25 kilometres outside of the centre. It's the first time in China that an AV ride-hailing service connects urban roads and highways (PR Newswire 2023b).

In August 2023 California’s Department of Motor Vehicles permitted Waymo and Cruise (owned by General Motors) to expand their services of driverless robotaxis around the clock. Previously their services were restricted to operate between 10 pm and 6 am and at a maximum speed of 30 mp/h (Bhuiyan 2023). About one week after this permission, a Cruise robotaxi was involved in a crash with a fire truck, passing a red light while rushing to an emergency. One person was injured and taken to hospital (Goswami 2023). On August 15th a Cruise robotaxi drove into a road construction zone and got stuck in wet concrete (Templeton 2023). After these and other incidents public authorities told Cruise to halve their fleet to 50 vehicles in operation at daytime and 150 during the night (Lee 2023). On October 2nd 2023 another severe accident happened. A human driver hit a pedestrian crossing the street against a red light and threw her into the path of a Cruise robotaxi. The driverless car braked, hit her, stopped on top of her and then tried to pull over – dragging her about 6 meter. The robotaxi is equipped with 40 cameras and sensors and according to Cruise, the car quickly swerved and braked trying to avoid a collision. The pulling over was built into the car’s software to promote safety and, says Cruise, is required by both California and federal regulators. The woman was severely injured, the human driver that caused the initial crash hit-and-run and has not been caught (Marshall 2023). After this accident California’s Department of Motor Vehicles suspended Cruise’s permit to operate robotaxis. Cruise had been quickly expanding to other cities, including Phoenix, Austin, Dallas, Houston and Miami. After the suspension in California it paused all driverless operations nationwide. Opponent Waymo is still operating robotaxis in San Francisco (Evers/Bosa 2023). On Octobre 26th 2023 taxi platform provider Uber announced to include driverless rides provided by Waymo in Phoenix. Clients will be notified via app and can reject the ride with a driverless robotaxi. If they accept, they pay just the same price as with human drivers (Capoot/Piazza 2023). Volkswagen starts its first AV test programme in Austin (Texas) beginning July 2023. By the end of 2023 about 10 vans with autonomous driving systems will be deployed in Austin, although for now they all will have human backup-drivers on board (Rosevear 2023). In Octobre 2023 Honda and General Motors announced to jointly start a driverless taxi service in Tokyo in 2024, using a newly developed vehicle without a driver’s seat (Nikkei 2023).

In Austria the first self-driving shuttle bus on public streets started as a research project in April 2017 in Koppl near Salzburg City on a track of 1.4 kilometres to better connect the rural village with the city. From April to November 2017 the shuttle undertook 240 test-rides without any accident. In the only critical incident the self-driving shuttle stopped due to overheating of the electric engine while climbing an acclivity, couldn't resolve the problem by itself and had to be brought back on track manually. Although the shuttle produced by Navya supposed to be the first self-driving vehicle reaching automation level 5, the final report on this project classified it as level 3, since human drivers had to fulfil certain manoeuvres and always had to back-up the self-driving system (Zankl/Rehrl 2018). In December 2022 the successor Digibus 2.0, an adapted version of a commercially available VW e-van, hit the streets for the first time and will be tested publicly from middle of the year 2023. The infrastructure has also been equipped with additional technology, including 5 roadside units to guarantee high-speed internet connection and a high-definition digital map. With this new vehicle a maximum speed of 50 km/h corresponding to the Austrian decree for autonomous driving is possible, while the previous shuttle was restricted to 20 km/h (Salzburg Research 2023). In September 2017 another research project with an autonomous shuttle bus named "SURAA" (Smart Urban Region Austria Alps Adriatic) started in Pörtschach (Carinthia). Since May 2018 the shuttle runs on a daily basis following a fixed timetable and since May 2019 the route also includes a federal highway (SURAA). On 9th August 2023 the shuttle crashed into 3 parked cars, no humans were harmed. The human safety driver deactivated the self-driving system in a daily routine to manually drive the vehicle to its parking garage away from its technically equipped test-route. According to the project manager the accident was due to a human error. The shuttle-service was suspended afterwards to properly analyse the incident (SURAA 2023). In Vienna the publicly owned transportation company Wiener Linien started to test a self-driving shuttle for 11 passengers in June 2019 in the neighbourhood of Seestadt Aspern, running on a 2 kilometre route with a maximum speed of 20 km/h. On July 18th 2019 a minor accident happed, when an inattentive pedestrian crossed the road and laterally bumped into the vehicle (ORF Wien 2019). In 2021 the company announced to cease its testing after 3 years and more than 12.000 kilometres due to several flaws in practical use, for example problems with heavy rain, snow and foggy conditions (ORF Wien 2021).

1.2.3 Regulatory frameworks on automated driving

The United Nations Economic Commission for Europe (UNECE) brings together 56 countries of Europa, Asia and North America to promote economic cooperation. Its Inland Transport

Committee (ITC) is a platform for international cooperation to facilitate international transportation, reflected in more than 50 international agreements and conventions providing a framework for legal and technical regulations. One important framework is the 1968 Convention on Road Traffic and Signs and Signals ("Vienna Convention"). As related to autonomous driving one fundamental principle was, that every vehicle moving in traffic shall have a responsible driver (article 8/1) and this driver always has to fully control the vehicle (article 8/5). In 2016 the convention was amended, stating that "systems which influence the way vehicles are driven" and "systems which can be overridden or switched off by the driver" are in accordance with the original article 8 of the convention (EP 2016: 6f). Following this convention vehicles still need to have a driver, although she can take the hands off the wheel temporarily. In summer 2022 the UNECE decided to raise the speed limit from 60 to 130 km/h for the use of automated and autonomous vehicles under certain conditions (UNECE 2022). Mercedes announced to gradually increase the maximum speed of 60 km/h of their "drive pilot"-system, without a fixed timeline to reach the newly possible 130 km/h (Donath 2023).

Germany

To discuss ethical questions in the context of automated and autonomous driving the German government installed an ethics committee consisting of 14 experts from universities, industry, justice and different lobby organisations in September 2016. In June 2017 it published its final report, including 20 norms for the implementation of automated and autonomous driving (Ethik-Kommission 2017). N°1 and N°2 focus on the betterment of traffic safety as overall purpose as well as a positive balance of incidents compared to human drivers as a precondition for any development towards automated and autonomous vehicles. N°3 states that responsibility for regulation and approval of AV is incumbent on public authorities following the purpose of reducing accidents, accepting the unavoidable remaining risk. N°4 emphasises personal responsibility and freedom of decision as principle values of open societies. N°5 touches upon the aspect of dilemma situations, stating that such situations should be avoided as far as possible, using all technical and digital means – i.e. defensive software programming. N°6 states that using AV might be advantageous, but it should never become obligatory. N°7, N°8 and N°9 deal with dilemma situations. N°7 gives a clear priority of damage to property over damage to persons. N°8 holds that any incident depends on a unique context and behaviour of all participants, which is why it forbids generalizations and ex-ante evaluation in terms of preprogrammed decisions. N°9 explicitly bans any form of discrimination based on personal characteristics (such as age, gender or physical constitution), a netting of victims and also forbids sacrificing of bystanders. N°10 deals with the responsibility in case of accidents with AV and proposes a shift from drivers to producers, operators of technical systems as well as legal, political and infrastructural decision-makers. N°11 underscores that AV should follow the same conventions as general product liability in terms of safety and ongoing improvements. N°12 to N°15 deal with transparency, the possibility of a totally connected traffic system, IT security and protection of data privacy. N°16 and N°17 discuss overruling and taking over control between human and virtual driver, stating that it always has to be clear who is in control, combined with internationally agreed and appropriate time-limits for the AV to hand over control to human drivers – in order to avoid sudden states of emergency. N°18 generally accepts the use of machine learning systems, given that these systems meet all the necessary safety requirements. N°19 suggests to define a mandatory safety position for AV to realize independently in case of emergency and N°20 suggests to include basic knowledge on digital systems and AV already in general education and especially during trainings for the driving license.

On June 21st 2017 an amendment to Germany's "Straßenverkehrsgesetz" (StVG) entered into force, that for the first time allowed for level 3 automation in passenger vehicles. Under certain conditions, following the OEM's terms of use and after permission of the "Kraftfahrbundesamt", drivers were allowed to take their hands off the wheel and shift attention away from traffic – given that they can take back control within 10 seconds on request (Bussgeldkatalog 2023). The first model achieving a permit based on this legislation was Mercedes S-Klasse end of 2021, sales of their "driving pilot"-system started in May 2022 (Theile 2022; see also 1.2.2). On July 28th 2021 another amendment to the StVG entered into force, regulating conditions and specific use cases (e.g. shuttle-services, goods transportation or automated valet parking) for autonomous driving on level 4. With this regulation Germany allegedly was the first country to enact a regular legal framework for autonomous driving for specific use cases within specified operational areas worldwide (BMDV 2021, Paulo 2023). The amendment was followed by an enactment ("Autonome-Fahrzeuge-Genehmigungs-und-Betriebs-Verordnung", AFGBV) with further details and conditions to prepare the practical implementation, entering into force on July 1st 2022 (KBA). To prevent collisions the AFGBV requires that automated vehicles must be able to recognize other road users, animals and objects (§3/8/1) and conduct a risk assessment including adequate safety manoeuvres (§§3/8/2-3). The StVG already stated, that any AV must be supervised by a human technical supervisor ("Technische Aufsicht"), that has to have – as well as the passengers of an AV – the possibility to deactivate the autonomous driving (BfJ 2023a: §1d/3, §1e/2/8). Corresponding to the further details in AFGBV, this technical supervisor requires several conditions, such as adequate technical training and a valid driving licence (BfJ 2023b: §14). A precondition for AV is also a safety system designed to generally avoid collisions, be able to identify dilemma situations giving human life highest priority over damage to property and never to discriminate on the basis of any personal characteristics (BfJ 2023a: §1e/2/2). Hence the StVG and AFGBV clearly adopt some of the primary conclusions of the 2017 code of ethics.

Austria

In Decembre 2016 a legal framework for testing and using AV has entered into force in Austria (RIS 2023). It didn't allow testing AV in general, but only within pilot projects under certain conditions, on clearly defined routes and for specified reasons – which initially were: the usage of micro-buses, autopilots for passenger cars on highways (including lane changing) and military vehicles. In 2019 the regulation was amended to legally allow the usage of two features already built into commercially available cars: automated lane-keeping assistants on highways and automated parking assistants. Up to this amendment drivers in Austria were legally obliged to always keep at least one hand on the steering wheel. Now they temporarily could let go, or, in the case of parking, even be outside the vehicle. Nevertheless they always have to pay attention and be able to intervene in case of emergency, either by taking over control again directly or indirectly via any kind of remote control or digital control system (e.g. an app). Given this unaltered responsibility of the driver, the amendment doesn't shift liability away from the driver in case of accidents while using these features (BMK 2019). In 2022 it was amended again to expand the number of use cases for testing to further include transportation in passenger cars, goods transportation, autopilots for highway slip roads, parking assistance and mobile working machines. The second amendment also introduced further requirements for safety drivers and a mandatory risk analysis of the selected route section in pilot projects (oesterreich.gv.at 2023). The allowed maximum speed for the different use cases of testing is also specified in the regulation. For parking assistance the speed limit is set to 10 km/h, for micro-buses and working machines 20 km/h, for transportation in passenger cars 50 km/h and for goods transportation 50 km/h for vehicles with type approval and 30 km/h for vehicles without type approval (RIS 2023).

1.3 State of the art: ethical reasoning on autonomous vehicles

1.3.1 Autonomous vehicles as trolley ethics

Although legal scholars thought about consequences of AV already some years earlier, the ethical discussion on AV started around 2014 (Nyholm 2018a: 2) – advanced amongst others by several articles by Patrick Lin (Lin 2013, Lin 2014). As an ethical starting point and common ground virtually all articles on AV build on the promise of a vast reduction of traffic accidents and casualties. "Whether or not this ultimately turns out being true, it is their main selling point." (Nyholm 2018a: 2) As for the possible benefits, others go further: "AVs promise world-changing benefits by increasing traffic efficiency, reducing pollution, and eliminating up to 90% of traffic accidents." (Bonnefon, Shariff & Rahwan 2015: 1) Besides an indication towards traffic safety, most of the relevant literature on ethics and AV refers to the discussion of so called trolley-problems in one way or another (Etienne 2021). Here is an example by Patrick Lin:

> *"Imagine in some distant future, your autonomous car encounters this terrible choice: it must either swerve left and strike an eight-year old girl, or swerve right and strike an 80-year old grandmother. Given the car's velocity, either victim would surely be killed on impact. If you do not swerve, both victims will be struck and killed; so there is good reason to think that you ought to swerve one way or another. But what would be the ethically correct decision? If you were programming the self-driving car, how would you instruct it to behave if it ever encountered such a case, as rare as it may be?" (Lin 2016: 70)*

Trolley-cases typically meet three conditions: (1) a collision is imminent and unavoidable, (2) an agent has to chose how to distribute the resulting harm and (3) the options and outcomes are of perfect certainty (Himmelreich 2018: 671). The original version of the trolley-problem was introduced some 45 years ago by Philippa Foot (Foot 1978). It is about a runaway trolley heading towards five track workmen and the option of turning a switch in order to get on a side-track, where only one workmen is working. Judith Thomson famously discussed other versions of this scenario, for example one with a fat man standing on a footbridge over the trolley track – which leads her to question if it is permissible to shove him onto the track in order to stop the trolley and save the five workmen (Thomson 1985). Interestingly, the discussion of quandaries in traffic started even much earlier and is about as old as the introduction of cars themselves. German legal philosopher Josef Köhler for example raised a similar issue already in 1915:

> *"Imagine that a car can no longer be brought to a standstill over a short distance, but that it is still possible to steer it in such a way that it drives to the right or to the left instead of moving straight ahead. If there are people straight ahead, on the right and on the left who can no longer avoid it, the driver is not in a position to avoid killing people, but he can divert death to one side or the other by moving the steering wheel." (Köhler 1915, cited in Paulo 2023: 1046f)*

The public as well as the academic discourse of ethical issues on AV in the past 10 years has focussed very much on the discussion of such alleged quandaries, "the ethics of autonomous vehicles seems to have settled on the so-called 'trolley-problem' as the lens through which to explore solutions." (Andrade 2021: 124) This predominance is so strong, that even its proponents utter caution: "Such edge cases have drawn so much attention that they threaten to cannibalize the whole field of autonomous cars. Almost every discussion of autonomous car ethics features some version of the so-called trolley dilemma […]" (Bonnefon, Shariff & Rahwan 2019: 503). Noah Goodall expresses an educated guess on the reasons for this: "I suspect the trolley problem is so often employed by the media for four reasons: it is fairly well-known, it is a stark choice between a manageable number of alternatives, it has completely certain outcomes, and these outcomes have obvious moral implications." (Goodall 2016: 812)

As mentioned above, robot ethicist Patrick Lin was among the first scholars to open up this debate. His starting point is that crashes with AV and the facing of quandaries are unavoidable – and we'd better think ahead how to deal with them. Following Lin, crash-avoidance or handing back control to human drivers (which simply takes to much time) are no universal options for AV in emergency situations, therefore we need to discuss ethics in terms of crash-optimization instead (Lin 2016). Interestingly for the purpose of this thesis, he basically rules out the question of speed and sets a certain (dangerous or even deadly) vehicle speed as given – as we have seen above:

"Given the car's velocity, either victim would surely be killed on impact." (ibid.: 70) Furthermore Lin explicitly rejects the option of braking as an ethically dominant strategy, stating that there might be situations in which braking itself is risky: "A wet road or a tailgater, for instance, may make it dangerous to slam the brakes, as opposed to some other action such as steering around the obstacle or simply through it, if it is a small object." (ibid.: 71) He adds that although sometimes risky, braking still could be seen as best option if overall benefits outweighed the costs, but refutes this view dubbed "morality by math" (ibid.) right away – without giving a coherent argumentation for this rejection. As we will see below, other scholars disagree and argue in favour of braking as ethically dominant strategy (see section 1.3.4). In discussing different trolley-scenarios and possible solutions, Lin further points to an ethically and also legally very relevant aspect: the sparing of any particular group inversely equals to targetting its opposite. If your AV for example had to choose between crashing into a small, light and unsafe vehicle or into a big, heavy and safe SUV, it might chose the SUV in order to minimize harm – which would essentially mean that your AV's algorithm targets SUV (ibid.: 72f). Discussing questions of the moral duty of self-sacrifice in certain trolley-cases, Lin also mentions an interesting point on responsibility and AV in general: "Also, because the driver is the one who introduced the risk to society – operating an autonomous vehicle on public roads – the driver may be legally obligated, or at least morally obligated, to absorb the brunt of any harm, at least when squared off against pedestrians, bicycles, and perhaps lighter vehicles." (ibid.: 72f) I will come back to this in the section on responsibility below (see part 2).

Also building up on trolley-cases, psychologist Jean-François Bonnefon et al. was among the early contributors to the ethical discussion of AV, introducing an experimental and empirical approach (Bonnefon, Shariff & Rahwan 2015). In order to deal with unavoidable dilemma situations, "manufacturers and regulators will need to accomplish three potentially incompatible objectives: being reasonably consistent, not causing public outrage, and not discouraging buyers." (ibid.: 2) To get at this, they "suggest that a way out of this conundrum is to adopt a data-driven approach, inspired by experimental ethics, to identify moral algorithms that people are willing to accept as citizens and to be subjected to as car owners." (ibid.: 3) In a first attempt they conducted three online-surveys with several hundred participants in June 2015, focussed on the question whether people would accept harm minimizing, utilitarian AV. As a result they conclude, "that participants were generally comfortable with utilitarian AVs, programmed to minimize an accident's death toll." (ibid.: 5) To some extent even self-sacrificing AV are valued, although the participants "actually wished others to cruise in utilitarian AVs, more than they wanted to buy utilitarian AVs themselves." (ibid.: 8) As already mentioned above, this kind of ethical reasoning on AV aroused large interest and was picked up by the public debate (Griffin 2015). The by far biggest and most frequently cited project in this regard was the so called "moral machine": an open online-platform confronting participants with stylized trolley-cases, focussing on nine variables in order to analyse moral preferences in case of AV-quandaries with two possible alternatives (Awad et al. 2018).[1] Nearly 40 million trolley-decisions were made in more than 200 states and territories worldwide. Although preferences showed significant variations across different cultural clusters, globally speaking "the strongest preferences are observed for sparing humans over animals, sparing more lives, and sparing young lives. Accordingly, these three preferences may be considered essential building blocks for machine ethics, or at least essential topics to be considered by policymakers." (ibid.: 60)

Other scholars are sceptical of such empirical investigations of moral questions. Nyholm (2018a) argues, that participants in such online-surveys mostly don't have any real-life experience with AV, their answers often lead to inconsistencies (e.g. as seen above: utilitarian AV yes, but only for others…) and they typically don't have to justify their responses – although this is key to ethical reasoning in order to balance different attitudes and opinions. Etienne (2021) methodologically criticises the poor quality of the data for its strong selection bias towards tech-savvy internet-users, as well as the usage of highly abstract trolley-problems to discuss morally relevant real-life situations in general (ibid.: 94f; further discussion see below). His major criticism concerning the whole project relates to the shift from descriptive to normative ethics – as some of the researchers engaged in the moral machine-project suggested in a subsequent paper, proposing a voting-based system for ethical decision-making: "We submit that decision making can, in fact, be automated […] by aggregating people's opinions on ethical dilemmas." (Noothigattu et al. 2018: 1587) Etienne harshly opposes this view and states that such a shift "cannot seriously be associated to anything else than a clear ethical fault and an academic fellony [sic] against the experiment's subjects."

1 The nine variables included: humans vs. pets, staying on course vs. swerving, passengers vs. pedestrians, more lives vs. fewer lives, men vs. women, young vs. elderly, pedestrians legally crossing vs. jaywalking, fit vs. less fit, high social status vs. low social status (Awad et al. 2018: 60). For further information see: www.moralmachine.net/

(Etienne 2021: 93) He even goes further, warning that such a shift "is not only fallacious but also dangerous for our democracies." (ibid.: 95) First, because algorithms and machines are not able to justify their decisions and therefore cannot be seen as moral agents, excluding them from moral decision-making. Second, "because aggregating individual uninformed beliefs does not produce any common reasoned knowledge." (ibid.: 97) He also refutes the underlying ethical rationale behind most claims supporting the introduction of AV, calling it the "highest moral imperative": in order to save human lives, AV should be deployed as soon as possible, even accepting loose regulation (ibid.: 89). Besides a poor philosophical justification, the same goal could be reached with other, less disputable measures as well – e.g. lower general speed limits (ibid.: 99ff). Over and above the faulty shift to normative ethics, Etienne is also generally sceptical of the moral machine-approach. First, because of its discriminatory, unconstitutional and morally irrelevant differentiation factors, "polarising the debate around erroneous AV dilemmas." (ibid.: 103) Second, because such problematic but popular projects distract from more relevant ethical discussions, e.g. the threat of AV-based mass surveillance (ibid.: 104). Berendt (2020) takes a totally different approach to identify weaknesses of the moral machine-experiment by conducting an "ethics pen-test", i.e. diverse working-groups discussing four fundamental questions about the setup and implementation of the project. One criticism focussed on the abstract and superficial nature of the given alternatives, which could lead to biased results. Another criticism questioned the starting point of the experiment, holding that the ethical discussion on AV shouldn't focus on allegedly unavoidable trolley-dilemmas but on how to avoid crashes in the first place (ibid.: 10).

1.3.2 Pro and contra trolley-discussions in the context of autonomous vehicles

Contra arguments

As already mentioned above, a big part of the ethical discussion on AV evolves around trolley-problems and whether they are applicable to the context of AV. Nyholm and Smids (2016) agree, that AV unavoidably will crash sometimes, which is why we need to think about adequate "accident-algorithms" (ibid.: 1276). As others, they assume high levels of speed as given: "Yet, self-driving cars cannot be a 100 % safe. This is because they will drive with high speed in the midst of unpredictable pedestrians, bicyclists, and human drivers." (ibid.: 1275) And later in the text: "However, since cars are heavy and move with high speed, physics informs us that they have limited manoeuvrability, and that they often cannot simply stop." (ibid.: 1277) Nyholm and Smids reject the discussion of trolley-problems in order to solve ethical questions surrounding AV on the basis of three crucial disanalogies. Firstly, the decision-making process in classical trolley-cases and trolley-cases involving AV is fundamentally different. In the former, it is a split-second decision, taken by one morally responsible individual right in the situation as it happens, in the later an informed decision, taken at an earlier stage, resulting from negotiations between multiple stakeholders (e.g. engineers, lawyers, ethicists etc.). Classical trolley-cases, the authors state, are much more similar to accidents of conventional cars than to AV (ibid.: 1280f). A second disanalogy regards the abstract nature of classical trolley-problems, excluding any further considerations besides the decision whom to run down or spare. In real-life conditions we necessarily have to take other things into account, such as to reflect carefully on what and who can be held responsible morally and legally. Such considerations will typically include both "forward-looking" (i.e. how to programme accident-algorithms) and "backward-looking" (i.e. who is responsible if an accident has occurred) responsibility (ibid.: 1283f). The third crucial disanalogy Nyholm and Smids identify also is about the abstract nature of classical trolley-problems, this time focussing on the stipulated certainty of its conditions and outcomes, whereas real-life decisions typically involve a high degree of risk and uncertainty: "Reasoning about risks and uncertainty is categorically different from reasoning about known facts and certain outcomes." (ibid.: 1286) Given these crucial differences between theory and real-life, Nyholm and Smids discard trolley-problems as a useful starting point for ethical discussions on AV.

Although mentioning that trolley-discussions can be useful to identify common responses and that they can serve as edge cases to test ethical theories, Noah Goodall (2016) also denies their practicability for the ethics of AV. He focusses on the later point of Nyholm and Smids, stating that besides very catastrophic accidents "almost everything else in road safety is probabilistic." (ibid.: 813) This is not only true for crash-scenarios but also for mundane traffic decisions, such as the calculation of appropriate lateral distances to other vehicles, cyclists and pedestrians. Therefore, Goodall argues to focus on risk management instead of abstract trolley-cases, although he admits this approach also has some shortcomings to address, e.g. the aspect of targetting algorithms also mentioned by Lin (see above). All in all, Goodall concludes that ethical reasoning on AV should focus on risk management, since this has been a successful strategy in many other areas as well, such as organ donation, radiation exposure and industrial safety standards (ibid.: 819). Building up on Nyholm and Smids (2016) and Goodall (2016), Johannes Himmelreich (2018) introduces further arguments against the utility of trolley-discussions for the ethics of AV. First, he identifies a tension or inconsistency between the assumptions of an unavoidable crash on the one hand and the possibility to make a significant choice on the other hand. If a car or AV is going to crash at high speed, room for manoeuvre is limited. Interestingly for the purpose of this thesis, Himmelreich explicitly mentions the role of speed as a relevant factor for the trolley-context. In a low-speed scenario there is more room for manoeuvre and "the low-speed scenario is not a usual trolley case in that the harms that accrue to different parties would not be equal in kind. Instead, if hit, the pedestrian might die but, when driving into a wall, the car passenger might only be injured." (ibid.: 674) As a second counter-argument Himmelreich draws on Nyholm and Smids' first point about differences in the decision-making process (see above) and adds the current tendency of AI-engineering towards bottom-up approaches (i.e. machine learning and neuronal networks). In trolley-cases typically a single agent takes a moral decision on the spot. In real-life context of AV, decisions depend on the vehicle- and software design at an earlier stage and, due to machine learning, the resulting actions taken by the car in an accident might even be unpredictable for its engineers (ibid.: 675). A third critique is that trolley-problems prompt the wrong kind of answers. This kind of thought-experiment asks for moral intuitions, when actually political solutions and regulations for the common good are needed (ibid.: 676). The fourth point against trolley-problems focuses on the epistemological comparability: the reliability of intuitive moral judgements in highly abstract trolley-cases is questionable, the unavoidable uncertainty in real-life situations makes a crucial difference (as already Goodall 2016 and Nyholm and Smids 2016 pointed out) and solutions for trolley-problems and specific accidents cannot be carried over to other situations, due to the manifold amount of relevant factors in real-life traffic: "What seems like the right choice in one situation might turn out to be the wrong choice in the next." (ibid.: 676) Instead of abstract trolley-problems, Himmelreich argues, the discussion should focus on "mundane traffic situations, such as approaching a crosswalk with limited visibility, making a left turn with oncoming traffic, and navigating through busy intersections […]" (ibid.: 678). Such a focus would shed light on a number of practically more relevant issues and challenges, as for example optimizing overall safety, balancing different values (e.g. safety vs. individual mobility needs) and adapting legal frameworks in order to cope with the introduction of AV.

De Freitas et al. (2020) take a similar line, arguing that abstract trolley-discussions actually are a distraction from more relevant issues in the context of AV. The discussion of trolley-dilemmas, they state, can only be of relevance if "they (a) occur with some reasonable frequency, (b) can be detected with confidence, and (c) can be negotiated by an AV system with controlled actions." (De Freitas et al. 2020: 1285) Not surprisingly they hold that all of these conditions are and possibly cannot be met in real-life. As to aspect (a) and (c) they go with Himmelreich (see above), arguing that due to their over-simplistic nature trolley-situations (with two clear and certain options) rarely ever happen in real-life and that if they do, due to the usually assumed high speed AV couldn't take reliable actions any more. Their second point implicitly relates to the risk management-approach by Goodall (2016) mentioned above, put from another perspective: AV cannot reliably detect if they find themselves in a trolley-dilemma, since this would require unrealistic knowledge of all relevant physical (e.g. road conditions, braking distance of other vehicles, etc.) as well as psychological (i.e. reactions of other human agents in the situation) details (ibid.: 1285). As per Freitas et al., the focus on trolley-discussions is but a distraction from more relevant issues, especially overall harm avoidance. Interestingly it seems that this view is also shared by engineers at AV-companies – for example Emilio Frazzoli, former CSO at AV-startup nuTonomy: "I consider 'trolley problems' one of the red herrings plaguing the world of AVs, distracting from the real issues." (Frazzoli, cited in De Freitas et al. 2020: 1286) Also JafariNaimi (2018) joins this line of argument, as others emphasising the uncertainty, complexity and uniqueness of real-life situations and the fact, that for a socially acceptable handling of possible AV-accidents we need to think way beyond simplistic trolley-cases and immediate outcomes (ibid.: 306ff). Even if we were to find some simple principles (e.g. maximizing life) in theory, the practical implementation would be hard to achieve: "Figuring out what maximizing life is in each case requires deliberation, and that's what algorithms cannot do." (ibid.: 310) Generally also JafariNaimi accuses the trolley-focussed discussion around the ethics of AV as being a distraction from more relevant issues and that it "is obscuring our ability to see what's at stake with this new technology." (ibid.: 315) Instead of focussing on theoretical traffic-quandaries, we should see the bigger picture, talk about the millions of casualties in our current traffic system, the liability of the automotive industry and the possible impacts such a new technology could have on the future design of our cities: "We might indeed think of the introduction of self-driving cars as an occasion for a radical rethinking of mobility that challenges and reorients dominant car-centric visions." (ibid.: 316) Not to forget, JafariNaimi reminds us, are AV first and foremost a business opportunity that naturally comes with vested interest: "We might note and question, for example, the roles and powers of strong industry interests in defining the terms of the discourse." (ibid.: 315) Also Etienne (2021: 93f) shifts attention to such possible conflicts of interest and refers to former graduate student researcher Rodrigo Ochigame, who explicitly describes problematic entanglements with big technology companies: "At the Media Lab, I learned that the discourse of 'ethical AI' [...] was aligned strategically with a Silicon Valley effort seeking to avoid legally enforceable restrictions of controversial technologies." (Ochigame 2019)

Pro arguments

On the other hand there are several scholars defending and promoting the usability of trolley-discussions in the context of ethics and AV. Paulo identifies three existing positions concerning the significance of trolley-problems for the ethics of AV (Paulo 2023: 1049ff). First, some scholars see trolley-problems directly relevant for the ethics of AV – although as per Paulo, this position is rarely defended amongst academic contributors. Second, a large part of academic papers see no relevance of trolley-discussions for the programming and regulation of AV (see also above). Third, a smaller part of the literature argues in favour of indirect relevance of trolley-cases, i.e. identifying important moral distinctions and principles by comparing different trolley cases, no matter how realistic they are. Paulo himself focuses on the first position of direct relevance, emphasising that the broadly discussed differences between trolley- and AV-cases simply don't matter: "What matters is the structural similarity concerning the morally relevant features of the cases." (ibid.: 1056). He refers to empirical studies based on the original trolley-cases by Foot and Thomson, suggests to accept these empirical results and use them for relevantly similar AV-cases by analogy. For the footbridge case (see section 1.3.1), Paulo states, he couldn't find any similar and plausible scenario in the literature on AV – and therefore proposes one himself, in which one out-of-control AV without passenger is heading towards five pedestrians at a crossing and could only be stopped by another AV with one passenger, that could be directed into the former to avoid the collision. Although far-fetched, Paulo argues, a scenario like this is at least possible (ibid.: 1056f). "We can thus conclude that direct analogies to the paradigmatic trolley cases are indeed normatively relevant for situations with autonomous vehicles, but not for all the situations that are discussed in the literature." (ibid.: 1059) Besides direct analogies trolley-problems can also serve to indirectly inform comparable fields of interest by identifying theories able to produce adequate solutions or insights – for example the distinction between killing and letting die referring to Kant's categorical imperative. As a consequence, Paulo holds, trolley-cases also have indirect relevance and approaches unable to deal with quandaries "should be excluded from the set of possible moral theories and principles to inform other moral problems, including the regulation of dilemma situations with self-driving cars." (ibid.: 1062) Support for the direct relevance of trolley-problems comes also from Bonnefon et al. (2019). Building up on the proposed shift to risk-management instead of clear-cut quandaries, they also transfer the trolley-discussion into the probabilistic realm. Even if AV were fully oriented towards minimizing the absolute risk of a crash, every action by an AV shifts relative risks between different road users. "But once they are aggregated over millions of cars driving billions of miles, these small statistical decisions add up to life and death consequences—and prompt the same questions as the trolley dilemma did. In expectation, the trolley dilemma and the statistical trolley dilemma are equivalent." (Bonnefon et al. 2019: 503)

In terms of indirect relevance of trolley-cases for the ethics of AV also Patrick Lin discards the objection of incomparability or attempts of showing that trolley-cases are far away from real-life situations: "Our thought-experiments are still useful in drawing out certain ethical intuitions and principles we want to test." (Lin 2016: 76) Interestingly, not only direct uncertainty concerning tangible aspects in real-life traffic situations is discussed as a relevant factor for the ethical reasoning on AV, but also indirect uncertainty concerning different ethical approaches. To cope with this, Bhargava and Tae (2017) propose to calculate an "expected moral value" (ibid.: 12) in order to sort out the best alternative.

Geoff Keeling (2020) devotes a full paper to counter four frequently cited objections to trolley-discussions in AV-ethics. The first objection states that reality is much more complex and abstract trolley-cases won't happen in practice. Keeling's reply is that this critique misses the point, since also unrealistic and highly stylized thought-experiments can produce meaningful insights for AV-ethics – for example in sorting out which properties make a difference to moral permissibility (ibid.: 294ff). The second objection states a significant difference between theoretical trolley-cases and the real-world. If this simply means that not all morally relevant factors are respected in trolley-discussions, the objection can be refuted by stipulating that this barely ever is the case and that the existence of other facts doesn't disprove the relevance of the ones found in a trolley-analysis. Nyholm and Smids (see above) argue that there is a more serious categorical difference, because in real-life traffic we have to deal with profound uncertainty. Keeling holds that risk and uncertainty make the ethical reasoning on AV more complicated, but in the end "the question of which properties the AV's utility function should track is the same in risky and non-risky cases." (ibid.: 300) A third objection focuses on the different decision-making processes, stating that trolley-discussion presuppose top-down decisions, whereas AV development in real-life tend to focus on bottom-up (i.e. machine learning) approaches. Keeling refutes this by pointing to the difference between decision-procedures and criteria for rightness – for the former the specific decision-making process is relevant and the objection therefore holds, but for the latter it's not (ibid.: 301). As fourth objection Keeling refers to Himmelreich's (2018; see above) point, that trolley-problems prompt the wrong kind of answers. Keeling's interpretation is that Himmelreich is asking for a societal choice instead of individual moral intuitions. It might be true, says Keeling, that a broad societal acceptance is desirable, "but it does not follow that our problem is *essentially* one of aggregating individual tastes, preferences or values." (ibid.: 304, emphasis in original) If a collective choice infringes well established moral principles, we should reject such a collective judgement. There are important aspects for moral design problems other than social choices, which is why Himmelreich's objection is flawed.[2] Lawler (2022) explicitly supports Keeling's answer to various objections questioning the usability of trolley-discussions for the ethics of AV and wants to highlight two aspects in particular. The first point is that adequately used, trolley-cases are not about finding morally permissible solutions to each and every case as such, but rather about the question "what explains the difference?" (ibid.: 208) The second point is about the often criticised differences between trolley-cases and AV. Again, Lawler holds that the whole point in dealing with trolley-cases is about explaining morally relevant differences: "It is *not just* that the trolley problem can be relevant, despite differences. Rather, the differences are, in fact, an *essential* element of trolleyology." (ibid.: 208, emphasis in original) If criticised, as Nyholm and Smids (2016) do, that

2 Keeling's interpretation of Himmelreich's argument seems somewhat inaccurate to me. Himmelreich doesn't seem to aim at simply numerically aggregating individual answers, but at an informed deliberation of different opinions in terms of political compromise. Understood this way, the examples given by Keeling to rebut Himmelreich's point are inappropriate.

trolley-discussions often only involve a limited number of options and considerations, then "[o]ne simply needs to construct the appropriate case or cases." (ibid.: 209) Furthermore Lawler rejects the notion of trolley-cases as philosophical equivalents to models put forward by Gogoll and Müller (2017; see section 1.3.3) and instead frames them as lab experiments: "Like lab experiments, they are not intended to be accurate representations of the world. Like lab experiments, the aim is to *isolate* a single variable, and ideally to keep everything else constant." (ibid.: 211; emphasis in original) Besides defending the use of trolley-analysis for AV-ethics Lawlor elaborates an argument in favour of introducing safe spaces (e.g. pavements) within which it shouldn't be permissible to impose risks on some people in order to save others. Furthermore he discusses different kinds of (relational) responsibility, which I will discuss below (see section 2.2.4).

As can be seen in this lengthy discussion about the usefulness and relevance of "trolleyology" (Lawlor 2022), a very large part of the discussion on the ethics of AV centres on trolley-cases. It seems that at least some of the pro/contra-debate is based on different understandings concerning the question what exactly trolley-problems should explain and how to use them. "If the methodology that the critics criticise is not the same as the methodology that trolleyologists actually use, it seems that the critics have missed their target. They are attacking an approach that no one actually defends." (Lawlor 2022: 212) However, to conclude this section it might be a fair compromise to go with Fiona Woollard (2023): "Yet, when due care is taken, trolley problems still have a role to play in the ethics of driverless cars." (ead.: 60)

1.3.3 Ethics settings in autonomous vehicles

Related to trolley-discussions and very relevant for the focus of this thesis, there is another cluster of research on ethics and AV centring around so called "ethics settings" for AV: instead of asking *how* an AV should act in trolley-cases, here the focus rests on the question *who* should take that decision? Given that AV sometimes will have to decide how to act in risky situations discussed in trolley-cases (i.e. crash-optimization or collision management) and that even seemingly uncritical and mundane decisions (e.g. how much distance to hold from other road users, when to overtake etc.) can result in a "statistical trolley dilemma" (Bonnefon et al. 2019: 503; see section 1.3.2), part of the job of AV-engineers "will be to program the car to react one way or the other in driving situations where similar value judgments must be made. […] Once you program an answer into the car […] you reify that ethical decision as an *ethics setting* embedded in the technology […]." (Millar 2017: 22, emphasis in original) Basically such ethics settings can either be rigidly fixed and built into the car by the manufacturers of AV or they can be variable, that is to say someone (e.g. owner, driver/passenger etc.) has to configure them before using the AV (ibid.: 22). Three ethically relevant key-questions in this debate are: What range of set points and options is acceptable? Should ethics settings be fixed or variable? Who should take the decisions? (ibid.: 29) Concerning the later question, Jason Millar refers to an empirical survey based on a trolley-dilemma involving an AV: 44 % of the participants wanted the driver/passenger of the AV to take the trolley-decision, 33 % voted for the public lawgiver and 12 % for the AV-manufacturer (ibid.: 25). "These results seem to suggest that, at the very least, people do not like the idea of engineers making high-stakes ethical decisions on their behalf." (ibid.: 25) Millar furthermore differentiates between high-stake (e.g. collision management etc.) and low-stake (e.g. climate control etc.) settings – stating that for the later it is less urgent to settle dispute about open moral questions. Interestingly for the purpose of this thesis, he explicitly defines low-speed collisions as low-stake: "Collision management settings that manage very low speed collisions with little or no likelihood of harm to people (or property), or where the harms will be done to non-humans such as squirrels (sorry, squirrels!), can also be deemed low-stakes." (ibid.: 30) Another differentiation is made between expertise-based versus non-expertise-based ethics settings. For example since it requires expertise about vehicle dynamics and several other factors to calculate a safe following distance at a given speed level, it would seem reasonable to limit the range of a variable settings to an expertise-based minimum. Other cases, such as trolley-problems, "do not submit so easily to expertise." (ibid.: 31) The discussion about ethics settings obviously is closely related to questions of responsibility. For AV-manufacturers for example it would be welcoming to broadly implement variable ethics settings, as a way of passing responsibility and liability for accidents on to drivers (Lin 2014, Millar 2017: 27).

Within the academic discourse on ethics settings for AV two opposing positions are prevalent: so called mandatory ethics settings (MES) versus personal ethics settings (PES). Gogoll and Müller (2017) coined these terms and argue in favour of MES. Although there are arguments (e.g. ensuring a high degree of individual autonomy) favouring PES, as to them the optimal solution for society as a whole can only be reached by MES. Gogoll and Müller refer to game theory and employ the prisoner's dilemma to show that with PES, strong incentives for egoistic individual ethics settings exist, leading to socially suboptimal results: "To put it differently, the result is that there is good reason to believe that morality will become crowded out in a world where people can choose their own ethics setting." (ibid: 694) In contrast Millar frames AV and other technical artefacts as "moral proxies" (Millar 2015: 53), emphasising the value of personal autonomy in morally relevant decisions, such as AV-quandaries and warns of paternalism. He compares the AV-case with the field of healthcare: "When faced with a diagnosis of cancer, for example, it is up to the patient to decide whether or not to undergo chemotherapy." (Millar 2014) Other than Gogoll and Müller he isn't concerned with optimizing for the social good or MES in terms of legal regulation, as main counterpart to drivers he focuses on AV-designers – and between these two Millar has a clear stance: "When robot cars must kill, there are good reasons why designers should not be the ones picking victims." (ibid.)

Another account on this question is elaborated by Johannes Himmelreich (2022). First he discusses some relevant arguments in favour and contra PES, then he makes his own case in favour of PES. As to the question of autonomy and AV as moral proxies (see above) Himmelreich holds that due to negative external effects of individual driving, the freedom and safety of others might be put at risk – therefore: "The best way to further individual autonomy in traffic might be to decide *collectively* about what driving is acceptable. So, instead of an argument in favor, we seem to have an argument against PES." (ibid.: 4; emphasis in original) Another argument in favour of PES he calls the "social dilemma argument": few people would buy AV, if ethics settings were set externally. Himmelreich doubts this premise, stating that AV "will likely find a significant market (esp. when the alternative is a manually operated car)." (ibid.: 5) As to Gogoll and Müller's argument in favour of MES (see above), Himmelreich disputes the framing as classical prisoner's dilemma: "There are other options. First, policymakers can regulate the strategic interaction in traffic such that traffic is not a PD [prisoner's dilemma; MS]. Second, even if traffic resembles a PD, cooperation can emerge. MES are hence not necessary in order to attain the outcome that is in the best interest of society." (ibid.: 7) From his analysis of the relevant literature, Himmelreich extracts three morally relevant desiderata in the context of using AV: respect to autonomy, overall safety and responsiveness to pluralism. Furthermore he sees two crucial conflicts in the debate on PES versus MES: self-interest versus interest of others and individual mobility versus safety (ibid.: 7).

Although speed is not explicitly mentioned in the context of the later conflict, I claim that it is at heart of it. All examples Himmelreich gives in this regard are about arriving at one's goal faster, including one example about overtaking a slower vehicle in front of you (ibid.: 7f). After all, Himmelreich concludes, PES should be somehow limited, but MES is too restrictive – which is why he proposes a variant of PES in which passengers before using the AV are given a dial in order to set their individual preferences on an interdependent scale based on the two trade-offs: "An increase in mobility will lead to an increase in other interest. And inversely, an increase in safety will lead to an increase of self-interest." (ibid.: 8)

Independent of ones opinion concerning this discussion on PES versus MES, what has become clear is that there are some decisions to take in terms of any AV's settings. Besides programming for hypothetical trolley-problems, also standard performance settings (e.g. eco-mode, acceleration profile, braking characteristics etc.) can and should be considered morally relevant since they arguably have significant impacts on safety and environment (Millar 2017: 28f). Such settings, especially concerning vehicle speed, could also be related to questions of responsibility – as I want to show in this thesis (see part 4).

1.3.4 Other approaches to ethics in the context of autonomous vehicles

Besides trolley-problems and the discussion of ethics settings there are naturally multiple other accounts on ethics and AV. Dubljevic et al. (2021) for example apply a multi-criteria decision analysis including a delphi-method and subsequent consensus workshop with experts from different fields to identify potential harms and risks concerning the introduction of AV into our traffic system. They discuss four different scenarios. The one with strict regulations for AV and restricted use mainly for fleets seemed to yield the most promising results, whereas the laissez-faire scenario did worst. Arfini et al. (2022) introduce a naturalistic perspective on morality in order to overcome disagreement between utilitarian, deontological and descriptive (i.e. the moral machine; see section 1.3.1) ethics. Their mixed ethical approach adapts to different communities, whereby norms "emerge from social utility, contextual necessities, and pragmatic adjustments between possibilities and goals of the agents who are immerse in changing socio-cultural environments." (ibid.: 725) As a result, norms and regulations of AV could vary between countries. To ensure international harmonisation, a supranational organisation should be established. In this context, Arfini et al. propose, AV-manufacturers would provide different ethical policies within the country-wide regulations and AV-buyers would choose their preferred policy according to their individual moral attitudes – e.g. going through an interactive test, which is afterwards discussed with a specialist before buying (ibid.: 727f, 730). Fabio Fossa (2022) proposes a bottom-up, function-based working approach to operationalise and translate general recommendations to practice. Starting from a specific technological function and their relevant features (e.g. autopilot on highways) we can identify morally relevant aspects, find practicable solutions and thus connect practice with theory. In his paper he discusses three morally relevant challenges concerning the introduction of AV: personal autonomy, explainability (from AV to users) and data privacy. As for autonomy for example, his bottom-up approach suggests to ask: Which functions might have an impact on personal autonomy? Which of them should stay under human control and what exactly would this mean in the specific case? What are potential impacts on other morally relevant goals (e.g. traffic safety) and what follows from this for the application of this specific function? (ibid.: 7) Although this is not a problem-solving procedure, nor is it a source of regulatory outcomes, Fossa states, it helps to connect theory with AV-practice (ibid.: 5). Borenstein et al. (2019) take an opposite approach. Instead of focussing on individual vehicles or trolley-scenarios, they suggest to apply a system level-analysis – the focus changes from car to traffic: "Note that 'traffic' is considered an obstacle when the focus is on a single car; when 'traffic' itself becomes a focus in system level thinking, there are both individual aspects of each car and characteristics that belong most properly to the traffic as a group phenomenon." (ibid.: 387) On a system level there are different approaches to mitigate systemic problems with the introduction of AV. Three of them are discussed in the paper: (1) regulation and implementation of safety and interoperability standards; (2) vehicle design and vehicle-to-vehicle communication; (3) infrastructure design and centralised intersection management (ibid.: 390ff).

Other scholars, such as Alexander Hevelke and Julian Nida-Rümelin (2015), Mark Coeckelberg (2016), Julio Andrade (2021), Antti Kauppinen (2021) and Fiona Woollard (2023) focus on questions of responsibility in the context of AV. I will discuss them in the next part of this thesis. To conclude this account of the state of the art on ethics and AV, I turn to two somewhat more practical approaches trying to resolve trolley-problems with a focus on operational and regulatory arguments. As both of them entail important insights for the purpose of this thesis, I will discuss them in more detail.

Rolf Johansson and Jonas Nilsson (2016) apply the perspective of functional safety standards to disarm the trolley-problem. Long before the development of ready for use AV, cars have been equipped with different assistance systems (e.g. anti-lock braking system). In order to restrict risks and faulty behaviour, the international community has agreed upon a standard on functional safety: ISO 26262.[3] As for any functional safety standard, potential risks are identified and assessed in terms of severity and frequency. If a risk is found to be too unlikely to occur or only has negligible effects, no risk-reducing measures are needed. All risks that do need risk-reducing measures are assigned a so called "automotive safety integrity level (ASIL)" (ibid.). For all risks with an ASIL, a safety goal is assigned as a vehicle-level requirement – for instance: "the vehicle shall stay on the road" (ibid.). If all significant risks are identified and given a correct ASIL, all of them are covered by a safety goal and all safety goals are fulfilled – then, by definition, the functionality can be claimed to behave safely and manufacturers internationally can argue to comply with product liability. As AV belong to the automotive domain Johansson and Nilsson propose to apply the ISO 26262 functional safety standards also for AV. For this, they highlight the differentiation between strategical (e.g. when to use a car), tactical (e.g. vehicle speed, distance to others, when to start an overtaking manoeuvre) and operational (i.e. how to react in a specific situation) decisions. In a manually driven car the driver is responsible for all of them. Driving assistance systems (e.g. anti-lock braking system) typically compensate for lapses on the operational level. AV, in contrast, also take charge of tactical decisions. Safe driving can be seen as balance between tactical choices and operational capabilities and this, following Johansson and Nilsson, sets the margin for AV:

> *"The more offensive (optimistic) tactical profile, the harder is the task on the operational level to guarantee safety. One way to solve this equation is to adjust the tactical profile to the current operational capabilities. How fast the operational capability is to react on surprises, sets the limit for how severe surprises the tactical choices may lead to. This means that for an autonomous vehicle a 'clever' and safe tactical decision pattern will not be to mimic an experienced driver, but to balance the current operational capabilities of itself. In some situations this will likely imply more conservative driving, but in other situations a less*

3 For further information see: https://en.wikipedia.org/wiki/ISO_26262

> *conservative style, compared to human drivers. The important thing is that the decision hierarchy is consistently implemented in such a way that the tactical choices always guarantee the operational tasks to be solved by the vehicle.” (Johansson and Nilsson 2016)*

Now, if all significant risks are assigned an ASIL and, for example the two safety goals “stay on the road” and “avoid collisions” are set – the trolley-problem, they conclude, would be disarmed. In an earlier paper Johansson (2015) discusses how to efficiently perform a hazard analysis and risk assessment for AV. The ISO 26262 doesn’t say how detailed the list of potential risks and hazardous events has to be. Completeness requires that all possible hazardous events are covered by a safety goal. A good way of verification, Johansson argues, is to let a review team challenge the list and try to find additional hazardous events that are not yet covered by the list of safety goals (ibid.). Interestingly, Johansson and Nilsson (2016) state that adjusting tactical decisions to operational capacity is exactly what also human drivers are legally obliged to aim for. For example, if an accident happens because a child runs onto the street in a residential area or nearby school, drivers cannot blame the kid because it was their duty to drive at a speed allowing them to stop well within the distance of their sight. In contrast to AV, human drivers frequently misjudge their operational capabilities and take risky decisions on the tactical level (see section 3.3). What ironically follows is that from this perspective trolley-problems are in fact for faulty human drivers, while they don’t apply to AV strictly obeying functional safety standards (ibid.).

Another attempt of disarming alleged trolley-problems is elaborated by Rebecca Davnall (2020). Based on practical insights in vehicle dynamics, her focus lies on braking in a straight line as dominant emergency strategy in traffic – an aspect Lin (2016) ruled out by superficially referring to the possibility of wet roads or tailgaters (see section 1.3.1). Himmelreich (2018) presumably had something similar in mind when pointing to a tension between significant trolley-decisions and practical operational choices (see section 1.3.1). The starting point of Davnall’s argument is that “the primary source of risk is the car's momentum – the fact that it is a large mass in motion. […] This means that the best way to minimise the risk associated with a car's motion is to reduce its speed, as much as possible in as short a time as possible." (Davnall 2020: 439) The critical factor for this is the tyres grip on the road surface – which is significantly higher when rolling compared to sliding or skidding. Braking in a straight line or swerving makes a big difference in this respect: “If the car swerves, then, it is at much higher risk of skidding [...], and even if a skid is avoided this can only be because it is doing less braking than it could be, which entails a higher-speed collision." (ibid.: 440) There are two groups of risks at play when swerving and skidding. First, skidding makes it difficult to predict the motion of a car and therefore it is easier for potential victims to get out of the way of a not-skidding car. Additionally, a skidding car has a higher probability of not hitting it's target head-on – which typically increases an accident's severity, since bonnets are designed to minimise harm to others and the car's occupants (ibid.: 440f). Second, as already mentioned, impact speed of skidding cars is likely to be higher. While in low-speed scenarios swerving might be an option due to a lower risk of skidding, this doesn't hold true for higher speed levels: "If an obstacle strays into a car's path at such a distance that a straight-on collision will happen at over 50 mp/h even with optimal braking, any swerve sharp enough to avoid the obstacle will be absurdly risky." (ibid.: 441) Davnall discusses two possible objections. The first one is that with additional information on all potential victims (e.g. age, gender, social status etc., as stipulated in many trolley-discussions) swerving might be morally demanded – e.g. to safe a group of children. But, Davnall states, such information couldn't reasonably be supposed to be available to AV and discrimination based on personal characteristics is ruled out by law anyway (see section 1.2.3). The only justified reason to swerve therefore could be the number of potential victims – but even in this regard there are relevant limitations (ibid.: 444ff): an AV can only consider individuals that it actually sees, which is not the case if sight is limited by other objects or persons (which typically is the case in urban traffic contexts); people have different risks of injury; people's behaviour in emergency situations is highly unpredictable; Secondly Davnall turns to multiple-vehicle cases and discusses the problem of tailgating. She admits that such cases are very complex and sometimes, although risky, swerving could be a better option. On the other side, braking in a straight line is very unlikely to increase the risk of passengers and bystanders and what's more, by general legal principles, rear-end collisions due to tailgating are the fault of the trailing car's driver, since she obviously didn't maintain safe distance (ibid.: 447). After all, Davnall concludes, braking in straight line is always the best solution for AV if no other cars are involved and even in complex multiple-car cases it is a good strategy worth considering.

1.4 Interim conclusion

What this first part of my thesis clearly has shown: AV are here and they're not going away any more – in practice on our streets, as well as in theory and in ethical reasoning. In the recent past AV have evolved from first pilot projects to regular commercial use in several regions around the world. Parallel to their successive practical introduction into our traffic systems, first steps in terms of legal frameworks have been developed and adopted. Despite these developments and the rapid spreading of AV over the last couple of years it remains unclear, when and how AV (e.g. as robotaxis, autonomous buses for public transportation or rather private cars) will make their way into everyday traffic systems. Either way we'll have to deal with them in one way or another and this means, that several ethical challenges arise in this context.

The ethical debate on AV intensified over the course of the past decade and so far has highly been focussed on trolley-discussions – including a meta-discussion on the plausibility of having this discussion in the first place. I take it, that these discussions indeed shed light on interesting ethical problems and bring about ethically relevant thoughts and therefore are justified – but on the other hand are of limited practical relevance for the specific case of AV. The empirically impressive "moral machine" for example brought to light some ethically relevant priority rules (i.e. human victims trump animals, less victims trump more victims and young victims trump old ones; see section 1.3.1). However, these results are not bound to traffic-accidents including AV, but rather apply to any form of quandary (e.g. terror attacks or medical triage). Thus, although relevant and interesting to a certain degree, I rather join the camp of the critics or at least state, that an excessive focus on AV trolley-discussions distracts from other relevant issues. Most relevantly for the purpose of this thesis, the trolley-discourse as we have had it up to now completely distracts from the factor speed, as well as harm and risk avoidance in the first place. Virtually all trolley-discussions start with setting high speed-scenarios and unavoidable "life or death"-decisions simply as given facts, without asking how these quandaries came about. This is problematic since the factor speed significantly influences the picture in at least two ways: (1) it changes the options you have and (2) it changes the outcomes you'll get. As Davnall puts it: it is "absurdly risky" (2020: 441; see section 1.3.4) to swerve in basically every scenario trolley-proponents suggest. Not to forget that lower speeds have the potential of rendering irrelevant "life or death"-quandaries altogether. But even without going this far, as Himmelreich (2018; see section 1.3.2) stresses, the outcomes in low-speed scenarios are different – i.e. drivers might survive most low-speed collisions due to vehicles' built-in protection measures, whereas they won't save drivers with higher speed. From this we could infer, just for the sake of argument, that in low-speed scenarios we should generally ask for self-sacrifices of drivers (since they most probably are going to survive them anyway), whereas in high-speed scenarios self-sacrifices are harder to justify. Another and for the purpose of this thesis very relevant way of indirectly introducing the factor speed into the discussion is the perspective of functional safety standards (see section 1.3.4) – i.e. product safety demands, that AV should only be allowed to go as fast as they safely can manage from an operational perspective. Ironically, this leads to the conclusion that trolley-problems as a matter of fact are rather for human drivers – since only human drivers regularly misjudge their capacities and, irresponsibly as well as unlawfully, engage in risky situations.

Together with many other relevant ideas for the further discussion in this thesis, I take away from this first part that the factor speed is noticeably absent in most of the relevant literature on this topic. In the few accounts where it explicitly is mentioned, the relevance for the purpose of this thesis clearly shows. The discussion on ethics settings (see section 1.3.3), including questions of who should decide about relevant configurations when using AV and which implications these decisions have on a societal level, perfectly leads the way to the next part of this thesis – i.e. responsibility in context of AV.

2. Responsibility

Interviewer:	*Good afternoon, HAL. How's everything going?*
HAL 9000:	*Good afternoon, Mr. Amer. Everything is going extremely well.*
Interviewer:	*HAL, you have an enormous responsibility on this mission, in many ways perhaps the greatest responsibility of any single mission element. You're the brain, and central nervous system of the ship, and your responsibilities include watching over the men in hibernation. Does this ever cause you any lack of confidence?*
HAL 9000:	*Let me put it this way, Mr. Amer. The 9000 series is the most reliable computer ever made. No 9000 computer has ever made a mistake or distorted information. We are all, by any practical definition of the words, foolproof and incapable of error.*
[…]	
Dave Bowman:	*Open the pod bay doors, HAL.*
HAL 9000:	*I'm sorry, Dave. I'm afraid I can't do that.*

(2001: A Space Odyssey, by Stanley Kubrick, original release in 1968)

As this thesis is all about responsibility in the context of AV, a thorough examination of this ethical basic concept obviously is needed. Traditionally responsibility was purely tied to human decisions. Meanwhile also other agents, such as organised groups or artificial agents, have moved into focus. This challenges traditional conceptions of responsibility and – as flagged up in the initial quotation of legendary AI science fiction-hero HAL 9000 – brings about radical questions: Should we allow robots to bear responsibility? If yes: which maxims do they follow and how can we ensure their responsible behaviour? If no: who else is responsible if robots' actions lead to bad outcomes?

As seen in the first part of this thesis, these questions play an important role in the context of AV. In order to get a good understanding of the relevant aspects for this thesis, I first want to overview some responsibility basics, before moving the focus more specifically to responsibility issues in the context of automation and artificial agents. In the concluding section of this second part I'll get back to AV and try to summarize the most important points for my argumentation.

2.1 Responsibility in general

The term responsibility has a rather short history. It was first used in its adjective form in the 13th century in France, spread to Germany in the 17th century and hadn't been developed and used systematically in the western world until the beginning of the 19th century (Loh 2019a: 1). Up to date it is an indistinctive concept with many different notions. A very basic understanding of it states, that agents are responsible for their actions in relation to a prevailing norm (Heidbrink

2017: 5). Etymologically the term traces back to the Latin "respondere", which simply means to give an answer (Nida-Rümelin 2011: 12). Answers generally are needed for actions deviating from prevailing norms – i.e. responsibility has to do with justifying intentional or unintentional deviations from what is normal (ibid.: 21f). From an etymological standpoint there are three different meanings attached to responsibility, which can be understood as a minimal definition: (1) it means to give an answer and account for one's actions; (2) it usually is a normative concept that goes beyond the identification of causal relations; (3) it "implies a specific psycho-motivational constitution of the responsible subject in question" (Loh/Loh 2017: 37; see also Loh 2019a/b).

Early thoughts on responsibility already can be found in the work of Plato and Aristotle, although they didn't develop a systematic concept of it. In one dialogue Plato hints to something like responsibility when stating, that educators are to blame for individuals with a faulty character, not the individuals themselves (Sauvé Mayer/Hause 2017: 89). Elsewhere in his work Plato emphasises the legitimacy of punishment for evil deeds, also found in a slogan from the Myth of Er: "Guilt is of the chooser. God is guiltless" (ibid.: 90; my translation).[4] Aristotle touched upon it in his discussion on voluntariness, stating that praise and blame can only be attributed to an agent's voluntary actions. An agent is the cause of such actions if she is the origin of the action, has the option to do or not to do it and has control over her action (Sauvé Mayer 2011: 1). Aristotle's account of moral characters therefore clearly relates to the question of responsibility: only agents who merit praise or blame "are those who are properly subject to the demands, expectations, and evaluations of morality." (Aristotle, cited in ead.: 3) Furthermore he thinks that only certain agents can intentionally cause things for which they are morally responsible. Animals and children for example lack the necessary disposition for this kind of actions due to their inability to perceive what's good/bad and therefore only accidentally can bring about just or unjust outcomes (ead.: 162ff). Therefore, as for Aristotle there are three conditions for moral actions: voluntariness, causality and intentionality (Heidbrink 2017: 9).

For centuries other thinkers in the tradition of Greek philosophy and later scholars such as Augustine, Anselm of Canterbury and Thomas Aquinas focussed on the identification of good/bad actions, on the existence of free will and the problem of determinism. It was not until modern philosophy that the discussion expanded to a relational conception of responsibility in terms of

4 Original quote: „Die Schuld ist des Wählenden. Gott ist schuldlos."

holding someone else responsible and ascribing accountability to someone – a perspective sometimes called the "second-personal" (Sauvé Mayer/Hause 2017). Immanuel Kant anchors responsibility in conscience and regards it as a duty towards oneself in the sense of self-responsibility for moral decisions. In the 19th century also consequences of actions and inactions as well as the question of responsibility for collective actions came into view. Therefore responsibility becomes a multi-faceted concept that not only includes accountability for ones actions, but also the ascription of responsibility due to social roles (Heidbrink 2017: 9ff). Such rather indirect forms of responsibility, independent from one's own actions are often called "institutional responsibility" (Nida-Rümelin 2011: 30). Three basic forms of responsibility can therefore be differentiated: (1) responsibility, i.e. individual moral duties, (2) liability, i.e. legal responsibility for compliance with laws and other positive rules and (3) accountability, i.e. social responsibility for ascribed and assumed roles and functions (ibid.: 19f).

As this brief overview shows, there is no uniform conception of responsibility. The traditional view holds that it is closely related to individual actions, characteristics of the involved subjects and societal norms. In a nutshell, the two basic conditions for moral responsibility following an Aristotelian perspective are: (1) control condition, i.e. an agent has to be the original cause of an action in order to be responsible for it and (2) epistemic condition, i.e. an agent has to know what she is doing and which consequences her actions have (Coeckelbergh 2016: 750; Coeckelbergh 2020: 111). An important remark for the second condition of course is, that self-inflicted lack of knowledge does not protect against punishment (Heidbrink 2017: 5). Especially in the context of AI and robots there is a broad discussion on the different capabilities agents need to possess in order to qualify as moral agents (see section 2.2.1). Within the traditional conception of responsibility the focus was placed on the following capabilities (Loh/Loh 2017: 37f; Loh 2019b: 128):

- <u>Communication</u>: moral agents need to be able to communicate with other agents
- <u>Autonomy</u>: moral agents need to possess a demanding form of autonomy, including the capacity to act, awareness of the context and consequences, some form of personhood and a scope of influence

- Judgment: moral agents also need to be able to make decisions, which includes some cognitive capacities such as reflection and rationality and also interpersonal institutions such as promises, trust and reliability

And as all of these capabilities can be possessed or attributed gradually, it follows that also responsibility itself is a gradual concept (ibid.: 38). Since responsibility is a very complex matter involving certain kinds of agents and their actions, their relation to others as well as context and norms, a somewhat more systematic approach might help to maintain an overview. Janina Loh identifies five different relational elements necessary for responsibility (Loh 2019a: 2ff):

1. subject: responsible agent and bearer of responsibility
2. object/matter: for what the subject is responsible (e.g. actions, events, things, creatures)
3. authority: to whom the subject is responsible (e.g. personal conscience, court, god)
4. addressee: towards whom the subject is responsible (i.e. victim or affected party)
5. normative criteria: set of norms, values, maxims, laws, rules that apply

To give a simple example how these relational elements interact in practice: a thief (1.) is responsible for stealing a book (2.) to the judge (3.) and the owner of the book (4.) within the context of criminal law (5.) (ead.). In the context of AI things get even more complicated since it is questionable, if artificial agents qualify as moral agents and are able to bear responsibility – or put differently: if artificial agents can only hold the object-position (2.) or also the subject-position (1.).

2.2 Responsibility in context of artificial intelligence

At present the traditional notion of responsibility seems to undergo a transformation. Global phenomena such as the climate crisis, as well as the emergence of new technologies, automation and digitalization, big data and robots equipped with different levels of autonomy and AI challenge our traditional understanding of responsibility (Coeckelbergh/Loh 2020: 8f). In the context of a globalised society and an upcoming age of automation several difficulties occur: Who is the responsible agent? What happens if we are not able to find responsible agents or to occupy the above-mentioned five relational elements according to our traditional understanding of responsibility? Are we necessarily confronted with responsibility gaps? And how to cope with the global consequences and indirect repercussions of our actions in times of a globalised society, facing existential challenges like the climate crisis? (ibid.: 11f)

In the context of AI, especially the above-mentioned control condition is problematic. Potential agents such as robots and AV might be causally responsible for certain actions, but it is an open question whether they qualify as genuine moral agents. In our legal system we already handle such situations – e.g. in the context of children and animals whom we don't hold responsible for their actions, putting liability on the shoulders of their caretakers (Coeckelbergh 2020: 111f). For AI such an approach comes with problems. First, AI is able to take actions very rapidly and way beyond our capacities to understand, react or intervene. Second, in the creation and development of AI many people and also many things (e.g. algorithms, data, sensors, hardware, software etc.) are involved – the so-called "problem of many hands" – leading to the question which of the possible agents can be held responsible in the end? (ibid.: 112f; Franssen 2015: 142f). Besides the trouble with agency, AI is also in conflict with the etymological root of responsibility in the sense that robots might not be able to answer questions about their behaviour to others. Especially robots working with machine learning-algorithms and neural networks might be incapable of explaining their decisions – the so-called "black box problem" (Coeckelbergh 2020: 117).

To cope with these difficulties scholars have found various ways of conceptualising human-machine interactions. An interdisciplinary working-group including philosophers, computer scientists, lawyers and practitioners drafted an ethical guidance for autonomous agents to overcome issues of responsibility issues already ten years ago, resulting in 5 rules.[5] Rule N°1 states: "The people who design, develop, or deploy a computing artifact are morally responsible for that artifact, and for the foreseeable effects of that artifact. This responsibility is shared with other people who design, develop, deploy or knowingly use the artifact as part of a sociotechnical system." (Miller 2011: 58) The other rules emphasise that responsibility must not vanish when more people are involved, that the sociotechnical system in which those artefacts are used should be taken into account, that the responsibility also includes being answerable for the artefacts effects, that users shouldn't be deceived about these artefacts and their effects and that users who knowingly use such artefacts are also morally responsible for that use. Furthermore this approach

5 Obviously this approach reminds us of the famous „Three Laws of Robotics" by science fiction author Isaac Asimov, formulated in the year 1942 (for further information see: https://en.wikipedia.org/wiki/Three_Laws_of_Robotics)

holds, that complexity, self-learning algorithms and neural nets are no exculpation – the people behind these artefacts still bear responsibility (ibid.).

Besides defining rules for human-machine interactions, scholars have also tried to include robots into the group of responsible moral agents themselves, whereas others discuss them as objects of responsibility within institutional arrangements and networks in order to prevent potential responsibility gaps. Furthermore there are inclusive and pluralistic approaches trying to overcome the traditional, individual-psychological conception of responsibility, potentially also extending it to include non-human beings (Loh 2019a; Loh 2019b: 138-194). In what follows I try to explore some of those approaches that seem promising for the purpose of this thesis.

2.2.1 Robots as moral agents and the responsibility gap

Since at least two decades scholars have engaged in a vivid debate around the question, if AI and robots qualify for moral agency – which in turn is closely connected to the concept of responsibility. As Mark Coeckelbergh puts it: "When humans are doing things and making decisions, we normally connect agency with *moral responsibility*. You are responsible for what you do and for what you decide." (Coeckelbergh 2020: 109; emphasis in original)

In an often-quoted paper Luciano Floridi and J.W. Sanders (2004) unravel exactly this conjunction, stating that moral agency and responsibility are separate things, whereas mixing them together is "confusing the *identification* of x as a moral agent with the *evaluation* of x as a morally responsible agent." (Floridi/Sanders 2004: 367) As examples they discuss children, animals and Oedipus – who all qualify as moral agents without being ascribed full responsibility for their actions (ibid.: 368ff). What this approach of so-called "mind-less morality" comes to, is the integration of artificial agents into the class of moral agents including moral accountability for their actions, but reserving the stronger conception of full moral responsibility for adult humans. Although this differentiation between agency and responsibility might be useful, it is questionable if it gives satisfying answers for the case of accidents with AV, since it practically consists in shifting the focus away from moral responsibility towards accountability and (if we take it a step further) liability. Another criticism holds that Floridi/Sanders apply an inflationary understanding of moral agency, since every agent able to make actions with morally relevant consequences, qualifies for moral agency (Woollard 2023: 54).

There are also other scholars who argue for moral artificial agents. John Sullins (2006) for example builds on Floridi/Sanders (2004) and sets autonomy, intentionality and responsibility as requirements in order to qualify as moral agent – whereas he dismisses personhood as a necessary condition. His "engineering sense" of autonomy sets very low limits: everything not under direct control of any other agent already is said to achieve a certain (indeed very low) level of autonomy. Furthermore the definition of intentionality as a behaviour "complex enough that one is forced to rely on standard folk psychological notions of predisposition or 'intention' to do good or harm", as well as responsibility as behaving "in such a way that we can only make sense of that behaviour by assuming it has a responsibility to some other moral agent(s)", remain vague (Sullins 2006: 28). Another relevant criticism for the purpose of this thesis is the fact, that Sullins puts responsibility on the same level as autonomy and intentionality – whereas traditionally the latter are preconditions for the first. Put differently: in order to be responsible, an individual already has to qualify as moral agent – not the other way round (Loh 2019b: 142ff). Colin Allen and Wendell Wallach (2012) argue that there is a spectrum from mere "operational morality" (i.e. tools, whereby responsibility stays within the hands of designers and users) to "full moral agency" (e.g. human responsibility), which in some unknown future theoretically is also achievable for artificial agents. As for robots they propose the perspective of "functional morality" as a compromise, where robots have at least some capacity for assessing moral challenges. In their view we need to put morality into artificial agents in order to avoid moral hazard, because further development of artificial agents is practically unavoidable and today's robots are "ethically blind". There are two ways to accomplish this: top-down implementation of some ethical rules into these agents versus a bottom-up learning approach (Allen/Wallach 2012). This approach circles back to the discussion of trolley-problems and ethics settings for AV as already discussed (see section 1.3.3).

Other scholars deny moral agency for robots. In contrast to Sullins (2006), Christian Neuhäuser (2013) for example sees responsibility closely related to some form of personhood. Based on the well-known approach of Daniel Dennett, he takes higher order intentionality (i.e. intentions referring to other intentions), the ability to communicate and consciousness as preconditions for personhood and therefore responsibility. Furthermore Neuhäuser adds as fourth condition the ability of moral deliberation – something that also Nida-Rümelin focusses on in his account on responsibility (Nida-Rümelin 2011: 32). Since at least today's robots don't qualify for personhood and relevant moral deliberation, Neuhäuser says, they lack the necessary preconditions for responsibility. As a result we have two options: either we need to find responsible humans within a network of shared responsibilities (see section 2.2.3), or we are forced to prevent the use of artificial agents. In practice most probably we’ll see different intermediate forms of dealing with this challenge – e.g. the legal arrangement of strict liability (Neuhäuser 2013: 284f; see section 2.2.2).

Just like Neuhäuser, also Andreas Matthias (2004) is concerned with the potential emergence of responsibility gaps when dealing with evermore autonomous artificial agents. He focusses on the above-mentioned control condition, i.e. agents only can rightly be said to bear moral responsibility for some outcome, if they were in control of the corresponding action or behaviour. This responsibility vanishes with autonomous agents: “In situations where the operator has reduced control over the machine he also bears less or no responsibility.” (Matthias 2004: 176) Therefore, together with autonomy also the problem of responsibility gaps increases – in a moral, as well as a legal sense (ibid.: 183). Robert Sparrow (2007) strikes the same chord with a focus on military weapon-systems. As one answer why we normally are not willing to ascribe moral responsibility to robots he points to the practice: what would “holding a robot responsible” practically mean, how would we do this? Usually we blame or praise responsible agents for their actions. In the case of robots this might seem odd, but theoretically possible. Sparrow takes the ability to suffer as condition for the possibility of punishment or praise. If a robot actually meets this condition, says Sparrow, ironically we would have created artificial agents with the status of full moral persons – which contradicts the development of “killer robots” in the first place (ibid.: 72f). Anyway, as per Sparrow today’s robots clearly don’t qualify as responsible agents – opening up a grey area, where we don’t really know who is responsible for the actions of more or less autonomous agents. “The only possible solution seems to be to *assign* responsibility to an appropriate individual — presumably the commanding officer who orders their use.” (ibid.: 74; emphasis in original) As this solution might seem somewhat unjust because it ignores the above-mentioned control condition, to Sparrow the application of such artificial systems is unethical as long as nobody is responsible.

Whereas Matthias (2004) and Sparrow (2007) base their argument on the control condition, Marino/Tamburrini (2006) and Santoro/Marino/Tamburrini (2008) argue that we need to dismiss exactly this condition in order to bridge the alleged responsibility gap. If we stick to the control condition, we'll indeed be confronted with a responsibility gap – especially so in the context of self-learning robots and AI. But we are not compelled to do so: "The crucial move here is to acknowledge the distinction between liability or objective responsibility on the hand, and moral responsibility on the other hand." (ibid.: 310) As many others, they use the comparison with parents and their children to clarify this distinction: parents are not morally responsible for the actions of their youngsters, but nevertheless liable – although overlaps are possible, e.g. if parents fail to educate their kids with due attention, then they're not only liable but to some degree also morally responsible for their kids' bad behaviour. Furthermore this solution should also concur with our sense of justice, since it traces back to the "ubi commoda ibi incommoda"-principle within Roman law.[6] And it applies as well retrospectively (i.e. compensation for damage incurred), as well as prospectively – which in the context of this thesis means, that the ones who benefit from the introduction of AV should bear the lion's share of liability (ibid.: 310f).

Another way of dealing with the control condition is presented by Filippo Santoni de Sio and Jeroen van den Hoven (2018), trying to redefine and adapt it for the context of AI in terms of "meaningful human control" – whereby we "need to move away from the myths and dreams about full autonomy and look at situated autonomy in real systems in the twenty-first century […]." (ibid.: 3) The idea of this approach is to show, that neither the mere presence of a human in the loop of a human-machine interaction, nor the causal contribution to some action are sufficient conditions to attribute meaningful human control – but nevertheless we need such a conception in order to settle liability issues and avoid responsibility gaps (ibid.: 4). They define two conditions for their novel conception of control: tracking and tracing. Tracking means that "a decision-making system should demonstrably and verifiably be *responsive* to the *human* moral reasons relevant in the circumstances — no matter how many system levels, models, software, or devices of whatever nature separate a human being from the ultimate effects in the world […]." (ibid.: 7; emphasis in original) Tracing basically means that it has to be possible to trace back responsibility for some morally relevant outcome to one or more humans, even if no one actually was causally responsible in a certain situation. A classical example is the drunk driver – who isn't in control and therefore not responsible at the moment an accident happens, but responsibility traces back to when he made the decision to drink and drive in the first place. More specifically Sio/Hoven state that to

6 English translation: „where there are benefits, there are disadvantages" (my translation)

fulfil this condition, we need at least one human agent who took part in designing or using the system, able to understand its capacities and possible effects in the world and able to understand that other human agents might be affected by it (ibid.: 9). Obviously in the case of an accident with such a complex system as an AV, the driver or operator is not the only human agent in question, but instead "responsibility for the accident may trace back to other agents upstream, or to the designers of the system […]" (ibid.: 9). So in a nutshell, the conception of "meaningful human control" argues that even in the context of AI we need humans in the loop to bear responsibility and therefore the traditional notion of the control condition is updated and redefined. Furthermore we saw that instead of single agents such as the driver or operator, there are several other human agents upstream who potentially can be attributed responsibility as well – which already hints to a line of argument that I will discuss later (see section 2.2.3).

Although there might be practical answers to the threat of possible responsibility gaps – it doesn't satisfy all related worries. John Danaher (2016) argues that there is more to it than just the legal side of the problem – for which vicarious and strict liability rules seem like viable solutions (ibid.: 301; see section 2.2.2). From a more psychological standpoint he additionally identifies a "retribution gap". The idea is simple: "Psychological evidence suggests that humans are innate retributivists: when they are harmed or injured they look for a culpable wrongdoer who is deserving of punishment." (ibid.: 299) Even if alleged responsibility gaps are bridged from the legal perspective, we still insist on some identifiable agents to morally blame or punish – and this sort of gap is way more difficult to bridge, says Danaher. He is sceptical if robots themselves qualify as blameworthy agents, suspecting that "it is far more likely that when a robot misbehaves people will look to the human manufacturers and programmers as potential targets for retributive blame." (ibid.: 304) Anthropomorphisation of robots could be one way out, but it is questionable if this would really satisfy our need for retribution and furthermore gives manufacturers an easy way to avoid and shift responsibility away from them. Another solution could consist in blaming the next best human commander. Unfortunately such a solution in the sense of "command responsibility" also comes with problems: (1) sometimes it conflicts with our sense of justice (i.e. if the commander didn't have a fair chance to prevent what happend), (2) it might put a prohibitive burden on the shoulders of potential commanders and developers and (3) it might vanish, as the development of artificial agents, robots, AI and the like is becoming increasingly fragmented (ibid.:

307). Especially the last point is very interesting for the purpose of this thesis and I will come back to it later (see part 4).

A somewhat similar but different approach to Santoro/Marino/Tamburrini (2008) is presented by Marc Champagne and Ryan Tonkens (2015). Instead of putting liability on one's back by legal requirements, they argue for voluntary and informed consent upfront, in order to close the alleged responsibility gap. The basis for this idea is the human ability to impute relations – even in cases where individuals weren't causally responsible for some outcome. Actually we routinely do this in the form of scapegoating – a rather unlovely habit that could be transformed into a solution, because "*when informed consent is present*, the same mechanism can be used fairly." (ibid.: 127; emphasis in original) Basically such a "blank check" responsibility is nothing more than a voluntary exchange: social prestige and other benefits versus liability for an uncertain risk. As per Champagne/Tonkens this solution would also apply for the case of totally autonomous robots, because if their behaviour was too unpredictable and risky, possibly no human would accept responsibility for them ahead of time – with the welcome consequence, that usage of extremely untrustworthy machines would be hindered. Besides the legal side of responsibility gaps, this solution to some degree also addresses the above-mentioned retributive gap, because "with such consent in place, (at least some) responsibility is thereby transferred to the agent doing the consenting." (ibid.: 133) Such a solution, someone willingly and explicitly agreeing ex-ante to bear responsibility for the usage of an artificial agent, could be a relevant aspect for the purpose of this thesis and will be discussed later (see part 4.).

Christian List (2021) argues for a somewhat more progressive approach, willing to grant AI itself a certain degree of agency and the ability to bear responsibility. He compares AI to group agency as in the case of organized collectives (e.g. companies, universities etc.), stating that the decisions and behaviour of such organisations is best explained by ascribing them some form of intentional agency "over and above their individual members." (ibid.: 1215) List argues that the same holds true for some artificial agents. He differentiates between "intentional agency" and "moral agency", whereby the latter has higher requirements to fulfil. An "intentional agent" has to have: (1) representational states encoding "beliefs" about its surrounding and status quo, (2) motivational states encoding "desires" and "goals" how to change the status quo and (3) the capacity to bring about change, i.e. to "act" according to its beliefs and desires (ibid.: 1219). "Moral agency" also includes the capacity to make normative judgements and deliberations about choices. Together with the traditional control and epistemic condition, moral agency for List is a necessary requirement for agents to qualify as fit for bearing responsibility (ibid.: 1227). Obviously humans qualify as intentional and responsible agents, but this rather functionalist conception also leaves room for robots and organisations. Since it's not possible to fully reduce responsibility for the behaviour and decisions of organisations and AI to the humans involved, List argues, we need to shift some of it to those non-human agents in order to avoid responsibility gaps. Actually, in the case of decisions made by AI, there are only three roles for humans in order to shift responsibility on them: enacting role (e.g. operator or user), authorizing role (e.g. owner, regulator) and system-design role (e.g. creator, software developer). But even if all humans involved behave responsibly, some bad outcome or accident might happen – which is why List argues in favour of AI-responsibility (ibid.: 1225ff). Some organisations with ethics committees, codes of conduct, compliance departments etc. can be said to functionally perform a sort of deliberation and therefore moral agency and in practice it is socially accepted to ascribe them a certain degree of responsibility. List goes further and turns this potential into a condition, suggesting that we should only permit big and powerful organisations if they qualify as moral agents – and that we should think about AI in the same way. "Society, via its regulatory authorities, should permit the use of autonomous AI systems in high-stakes settings only if structures are in place to ensure these systems' — or at least their legal representatives' — fitness to be held responsible for their actions." (ibid.: 1229f) As Allen/Wallach (2012; see section 2.2.1) he basically sees two ways to get there: a constraint-based (i.e. top-down) and a training-based (i.e. bottom-up) approach. Furthermore he suggests to use "ethical black boxes", similar to flight data recorders, in order to reconstruct the causes in case of accidents (ibid.: 1230ff). List doesn't really answer the question if AI and especially AV already today qualify as moral agents, but adds that from a functionalist perspective this is not a categorical but rather an empirical question. And for cases in which we cannot reasonably ascribe moral agency to AI, says List, we need to apply strict liability in order to avoid responsibility gaps (ibid.: 1231). Interestingly, List takes collective agency as a mere example for his discussion about AI – but he doesn't discuss the possibility of combining them. For AV we could think of ethics committees instituted within the software development company performing the normative deliberation necessary to qualify for responsible agents and (at least indirectly) affecting the behaviour and decisions taken by AV. For this we’d have to think of agency in the sense of human-machine collaboration, which will be discussed later (see section 2.2.3).

2.2.2 Legal responsibility in the context of autonomous vehicles

Evidently, moral and legal responsibility (i.e. liability) are not exactly the same thing. Arguably legal responsibility should be embedded within a broader discussion on moral responsibility. Legal responsibility can therefore be seen as a subset of moral responsibility with significant overlaps and “it represents an excellent starting point for thinking about robot ethics […]” (Asaro 2007: I.). For this reason it makes sense for the purpose of this thesis, to briefly look at the discussion on legal responsibility in the context of AV – without getting lost in the letters of the law.

Basically there’s a distinction between criminal law (i.e. public authority seeking justice for wrongdoing such as stealing etc.) and civil law (i.e. private parties seeking compensation for incurred damage). Peter Asaro (2007, 2016) discusses several legal concepts within the scope of civil law in order to cope with the challenge posed by artificial agents. The challenge with traditional concepts of liability is twofold: “unpredictability, and causal agency without legal agency.” (Asaro 2016: 191) One important starting point is product liability, whereby the most relevant concept in the context of artificial agents is negligence. The standard defence against charges of negligence is the reference to industry standards (e.g. ISO 26262; see section 1.3.4). An interesting aspect hereby is that liability can be split up amongst different parties involved – which is very common especially in the context of car-crashes (Asaro 2007: II.). Other promising approaches to cope with liability issues in the context of artificial agents are the concepts of “quasi-persons” (e.g. minor children), “diminished responsibility” (e.g. persons with permanent or temporary impairments), cases where agency voluntarily is transferred to legal agents (e.g. corporations) as well as strict liability, joint and several liability and forms of risk-pooling insurances (ibid.; Asaro 2016). Although from a legal perspective there might be viable solutions, “[j]ustice requires that those who will be held liable should be able to understand the scope and extent of the risks and liability they are assuming in deploying an autonomous artificial agent, and have some means of managing that risk through controls over the system.” (Asaro 2016: 192)

Madeline Roe (2019) tries to shed light on upcoming liability issues with AV via a comparison with surgical robots, since the latter are in practical use for already more than a decade and have some similarities to AV – i.e. both cases somehow depend on AI, the users are not the ones who created these machines, nor are they likely to fully understand them. An interesting difference is that surgeons necessarily are professionals, whereas AV often are praised to be available for everyone. This brings up the question if the ownership of a driving licence should be a precondition for using AV or at least make a difference for legal (and moral) responsibility (ead.: 335f). A lesson to learn from case law concerning surgical robots is that it's often quite difficult to assign liability between surgeons, hospital management and the manufacturers of the surgical robots. "Therefore, the case law regarding surgical robots focuses a lot on expert testimony, the duty to warn, and negligence by the controlling surgeon." (ead.: 329) Another conclusion is that compared to surgical robots, where professionals closely supervise every move and hospitals as professional institutions also play a major role for their application and therefore face some liability, in the case of AV a bigger share of liability will be shifted to manufacturers (ead.: 341ff).

As Asaro also Jeffrey Gurney (2013) focusses on product liability in the context of legal responsibility for AV and as Roe, he also emphasises the importance of expert testimony to assert negligence and design-errors on behalf of the manufacturers – but warns, that this could result in prohibitively high costs for most product liability suits on behalf of users or victims (ibid.: 265f). To correctly assign liability, Gurney discusses scenarios with different types of drivers, including attentive, distracted, disabled drivers and drivers with diminished capabilities. Basically he concludes that drivers should bear the lion's share of liability whenever they are controlling the vehicle or explicitly have a duty to supervise and intervene. Obviously AV-users should not be forced to always be attentive, since this would contradict the purpose of these vehicles. On the other side, when an AV is driving in autonomous mode the manufacturers should be liable – which would also ensure that manufacturers have a strong incentive to continuously improve their AV's safety (ibid.: 268ff). Drivers in such cases could still be partly condemned for comparative fault if they negligently misbehave, e.g. fail to be attentive and take over control in an emergency, if this knowingly was their duty (ibid.: 273f). Between car producers and software developers for an AV, Gurney argues, liability for accidents in autopilot mode should rather lie with the software developers, as their programming most likely will be the main cause for malfunctions (ibid.: 272).

Melinda Lohmann (2016) discusses the legal framework for the introduction of AV in Europe. Before product liability comes into play, she highlights the important function of strict liability of vehicle holders, which exists in most European countries. Despite discussions of a potentially reduced responsibility of drivers due to their diminished ability to understand and control AV, in judicial practice vehicle holders will have to face lawsuits in case of accidents – and rightfully so, says Lohmann: "It seems fair to burden the vehicle holder with the risk of automation, as he is the person who decides to introduce an automated vehicle into traffic and can take certain measures to ensure a safe functioning, e.g. undertake system checks before use. Moreover, the holder is the person profiting from the use of a self-driving car." (ead.: 338) As seen above, her point touches upon the "ubi commoda ibi incommoda"-principle. In a next step, vehicle holders in practice will try to recover their strict liability referring to product liability of manufacturers. Within the EU, product liability generally applies to defects regarding design, construction and instruction. A defective product is specified as such, that it doesn't provide the safety a careful user is entitled to expect (ead.: 338). The aspect of instruction is very important in this context, since in order to qualify as responsible, users exactly have to know what is expected of them. For manually driven cars only a very small share of accidents is caused by technical defects – and even in these cases defects mostly stem from poor vehicle maintenance on the side of vehicle holders. In the case of AV, Lohmann expects a liability shift towards the manufacturer because technical defects (including software errors) will gain in importance, whereas fault based liability on the driver's side will lose in relevance as vehicles take over more and more driving tasks. As Gurney (2013) she supposes this shift to incentivise improvements in safety design on behalf of the manufacturers – which seems as a fair burden, since these companies are the ones benefitting from selling AV (ead.: 336ff). In order to settle liability suits Lohmann also points to the foreseeably important role of so called "Event Data Recorders". Generally, Lohmann doesn't identify major liability gaps within the European legal framework when it comes to AV (ead.: 339).

In a very detailed analysis of the liability framework for AV in Germany, Tim Hey (2019) affirms most of the general conclusions made by other scholars mentioned above – but nevertheless points out a potential gap for which he proposes to introduce a novel form of strict liability for manufacturers. Generally, liability in the context of AV will conceivably shift from drivers (who more and more transform to users and passengers) and vehicle holders to manufacturers, as an increasing share of accidents will be caused by technical defects. Although the liability framework for manufacturers is quite strict, gaps might open up in case of increasing degrees of autonomy and unpredictability of AV combined with difficulties in fulfilling the burden of proof in case of accidents – which, again, highlights the role of expert testimony (ibid.: 243f).

Some further remarks on the broader context of liability within the traffic system are interesting for the purpose of this thesis. Basically there are two liability principles that apply within the traffic system: (1) liability for culpable wrongdoing and (2) liability for safety hazards. Strict liability-regulation historically was introduced to cope with the latter, being "the legislator's response to uncontrollable dangers that were abstractly associated with a certain activity, be it the operation of a railway, a vehicle or a nuclear power plant." (ibid.: 172; my translation)[7] In the case of strict liability for the operation of cars, the relevant safety hazards referred to in the German motor vehicle law of the year 1908 first and foremost focussed on the high speeds reached by these vehicles together with their large mass (ibid.: 176f). Interestingly for the focus of this thesis this connection between strict liability and speed still exists in current laws. In § 7 of German traffic law (Straßenverkehrsgesetz [StVG]) strict liability for harm and damage caused by cars generally lies with the holder of the car. According to § 8 StVG this explicitly does not apply to vehicles that cannot drive faster than 20 km/h on a level road – with the notable exemption of AV in autonomous driving mode (BfJ 2023: §§7-8). In the Austrian context a similar provision can be found in a law regulating liability for railways and motorized vehicles (Eisenbahn- und Kraftfahrzeughaftpflicht Gesetz [EKHG]). § 5 generally puts liability on the shoulders of vehicle holders, whereas § 2/2 exempts vehicles unable to drive faster than 10 km/h on a level road (RIS 2024a: §2, §5). As we will see later, this connection of speed and liability hints to the practical relevance of the basic argument of this thesis: within the traffic system, responsibility and speed are closely related (see part 4).

2.2.3 Institutional responsibility and responsibility networks to close the gap

Instead of focussing on the above-mentioned control condition and trying to identify individual scapegoats, there is a broad discussion on different forms of distributed and institutional responsibility, i.e. arrangements including more than directly involved agents. Such arrangements often are based on the separation of agency and responsibility. One example is responsibility in

7 Original quote: „Historisch war die Einführung einer Gefährdungshaftung stets die Antwort des Gesetzgebers auf nicht kontrollierbare Gefahren, die abstrakt mit einer bestimmten Tätigkeit einhergingen, sei es der Betrieb einer Eisenbahn, eines Fahrzeugs oder eines Atomkraftwerks."

case of corporations, companies and other forms of organisations, which commonly are accepted as bearers of responsibility if they satisfies some conditions – such as having explicit internal decision structures as well as collective goals (Heidbrink 2017: 20f; also see List 2021, section 2.2.1). Similar to List (2021), Neuhäuser (2013) discusses but (contrarily to List) finally negates the possibility to directly ascribe responsibility to artificial agents. Interestingly he introduces the notion of "responsibility networks" (ibid.: 279) – i.e. instead of focussing on isolated agents, he describes responsibility as a collective burden, warning that this network might become permeable by introducing more and more irresponsible agents such as highly autonomous AI-systems.

Also Allan Hanson (2009) criticises the traditional approach of strictly tying responsibility to isolated, individual subjects, calling it "moral individualism" (ibid.: 92). He argues for the perspective of "extended agency" in the sense of taking into account interactions of humans with non-human things and agents, leading to an approach he calls "joint responsibility". As an example he mentions the slogan "guns don't kill people; people kill people". Well yes indeed, says Hanson, but a subject with a gun is significantly different from one without a gun – also when it comes to responsibility, since responsible behaviour is based on having the choice to implement a certain behaviour in the first place (ibid.: 92f). In the context of this thesis we therefore could say that the joint responsibility of the subject "human driver–manually driven car" is something very different from the subject "human driver–AV". Inagaki (2010) takes this perspective further and discusses the notion of "joint cognitive systems" for the context of traffic. A "cognitive system" is a system able to modify its behaviour based on experience, typically a human being. The ensemble of a cognitive system with an artefact or other cognitive system(s) is called "joint cognitive system" (ibid.: 154). Examples are pilot–aeroplane, pilot–copilot–aeroplane as well as driver–car or user–AV. As per Inagaki, artificial artefacts can intervene on 4 different levels of a "perceptual cycle model": current understanding – action selection – action implementation – event. If an AV for example detects a potential problem, it might warn the driver, prohibit certain actions (e.g. to exceed allowed speed limits) or implement certain actions itself – each of which has implications for questions of responsibility in case of accidents (ibid.: 157f). Maarten Franssen (2015) discusses a similar notion of "sociotechnical systems", defined as "[c]omplex people-containing entities which are conceived and implemented as instruments to serve some purpose […]" (ibid.: 120). Human agents inside such systems are called "operators". Due to the increasing complexity of sociotechnical systems including artificial agents and the above-mentioned "problem of many hands" (see section 2.2) it is challenging to adequately attribute responsibility to human operators. Nevertheless this is taken very seriously in our society – as an example, Franssen quotes the International Civil Aviation Organization (ICAO) highlighting "the ultimate responsibility of pilots for the safety of their aircraft and its passengers and crew regardless of their evergrowing technical embeddedness." (ibid.: 141) As another interesting aspect in this context he discusses the notion of "meta-task responsibility", thought of as "a responsibility that people have with respect to the tasks of *others*, tasks that they are responsible for designing." (ibid.: 145; emphasis in original) For the case of AV this could be related to the discussion on ethics settings (see section 1.3.3) or in general to questions focussed on how to ethically design AV, which tasks and decisions should stay with AV-operators, which ones not? From the perspective of meta-task responsibility for example it might seem unfair to design AV in such a way, that they hand over control to human operators just a few seconds before an unavoidable accident happens.

Sven Nyholm (2018c) develops a notion of "collaborative agency" whereby agency is not ascribed to single agents (be it robots or humans) but instead understood as "distributed agency" in the sense of "coordination and collaboration of humans and machines" (ibid.: 1207). He criticises the tendency of other scholars to prematurely assume very high degrees of autonomy and agency for artificial agents, fuelling speculations about alleged responsibility gaps. Unpredictability and the inability to fully control a robot's actions are not sufficient to eliminate responsibility on the part of the users (ibid.: 1204ff). To make this point, Nyholm discusses different forms or degrees of agency (ibid.: 1207f): "domain-specific basic agency" (i.e. robots fulfilling specified tasks in specified ways), "domain-specific principled agency" (i.e. robots fulfilling specified tasks following some rules), "domain-specific supervised and deferential principled agency" (i.e. as before, including human authorities able to intervene and overrule) and finally "domain-specific responsible agency". For the latter, "the ability to understand criticism of one's agency, along with the ability to defend or alter one's actions based on one's principles or principled criticism of one's agency" is needed (ibid.: 1208). This notion of responsibility is clearly based on the ability to make deliberate decisions including the possibility of "standing one's ground", i.e. insisting on one's decision despite criticism – with the implication of being on a par with the criticising agents: "The practice of discussing reasons for and against our own or others' conduct assumes that we cannot simply order each around, but that we are instead on an equal enough level that we owe each other justifications." (ibid.: 1208) This notion of responsibility is akin to Nida-Rümelin, who states that: "We are responsible for our behaviour if we have reasons to behave in this way and not otherwise. It is our ability to deliberate that makes us responsible actors." (Nida-Rümelin 2011: 32; my translation)[8] Nida-Rümelin emphasises that the results of such deliberations are not determined in advance and do not simply follow from causal processes, calculations or algorithms (ibid.: 42). From this perspective, following Nyholm, an AV qualifies for the type of principled and supervised agency, but certainly not for responsible agency. Furthermore he distinguishes "individual agency", where the initiative comes from the relevant agent itself and "collective agency", where the initiative comes from some external agent or third party. A kid playing with its toys supervised by its parents belongs to the first type of agency, a kid tidying up its room on request of the parents belongs to the latter (Nyholm 2018c: 1210f). On this basis, Nyholm circles back to responsibility and the question where to locate it within such "human-robot collaborations". In the case of parents planing a bank robbery in which their kids execute it (i.e. the kids do most of the practical work) – moral responsibility obviously lies with the parents, who initiated and supervised the action as responsible agents. The same holds true for the case of an AV with a driver behind the wheel, setting relevant configurations and able to intervene anytime – although somewhat unpredictable, it was the driver who initiated the action and has the duty to supervise the course of action (ibid.: 1212ff). It gets tricky in a case without driver, but instead with a mere passenger inside the vehicle setting a destination but without possibility to set configurations and no duty to supervise the driving. In this case the AV collaborates with the passenger, but follows the programming specified by its software-designers and manufacturers. "This can make it hard to determine *which of the humans* involved are most responsible for the actions the car performs within the context of these collaborations." (ibid.: 1214; my emphasis) The important point is, that as per Nyholm only humans involved in these sorts of "collaborative agency" qualify as possible loci of responsibility and consequently no responsibility gaps appear – as long as we are dealing with artificial agents that we are able to control, overrule, update, intervene or stop. He doesn't deny that their might be, now or in future, other types of artificial agents or situations over whom we totally lose control – but we shouldn't

8 Original quote: „Wir haben Verantwortung für unser Verhalten, sofern wir Gründe haben, uns so und nicht anders zu verhalten. Es ist unsere Fähigkeit, Gründe zu wägen, die uns zu verantwortlichen Akteuren macht."

confuse this possibility (where genuine responsibility gaps might appear) with the more realistic and common situations discussed above (ibid.: 1215f). In order to identify appropriate loci of responsibility in human-machine collaborations, Nyholm suggests to apply a set of questions – e.g. in the AV-case: who is controlling or supervising and who is able to take over control, overrule or stop the vehicle, whose configurations and preferences is the vehicle conforming to, who is able to observe, monitor and understand the vehicles behaviour on the road, who owns the car etc. (ibid.: 1214). Although this approach is highly relevant for the purpose of this thesis, it doesn't settle all of the dispute. Nyholm seems to be more concerned with the moral (and not the legal) aspect of responsibility, but nevertheless it is somewhat unclear, as Roos de Jong (2020) mentions in a reply to Nyholm (2018c), if this approach also closes the above-mentioned "retributive gap" (see section 2.2.1). Furthermore, as Nyholm himself mentions, the approach doesn't give a clear-cut answer to tricky AV-cases – but rather serves as an interesting starting point for further investigations.

Another interesting and somewhat similar approach is developed by Janina Loh (2019a, 2019b, Loh/Loh 2017). Based on Allen/Wallach's (see section 2.2.1) account of moral agency as a gradual concept, Loh understands current AI-systems including AV as rather having operational agency instead of functional or even full blown moral agency – i.e. they don't qualify (yet) as responsible agents and therefore cannot hold the subject-position in terms of the above-mentioned relational elements of responsibility (see section 2.1). In order to avoid responsibility gaps, Loh builds upon the notion of "responsibility networks" developed by Neuhäuser (2013; see above). In the general case of road traffic, such a network potentially includes human drivers, AV-owners, AV themselves, car manufacturers, software developers, other road users, driving instructors, legal authorities etc. Referring to a specific object or matter of responsibility (see section 2.1), e.g. an accident with an AV, only a subset of the general road traffic responsibility network is relevant (Loh 2019a: 13f). Similar to Nyholm (2018c), Loh adopts the notion of "collective responsibility" for "hybrid systems", consisting of several distinguishable and at least partially autonomous agents – e.g. human-machine interaction as in the case of human driver plus AV. Such "hybrid systems" can also be understood as "plural subjects", tied together by a shared goal (e.g. getting from A to B). Those "hybrid systems" or "plural subjects" finally can be part of responsibility networks (Loh/Loh 2017: 42f). Put into practice, Loh suggests the following: In order to avoid responsibility gaps we need to distribute responsibility within a responsibility network. Road traffic broadly can be understood as such a network within which a human driver plus AV can be defined as a "hybrid system". The AV (together with software developers, manufacturers etc.) is responsible for maintaining standard driving operations, whereas the human driver should be responsible for morally relevant decisions – since it would amount to paternalism, if all such decisions would be dictated by an external agent or authority (such as the manufacturer or software developer). Therefore, Loh is actually emphasising the need for some kind of individual "ethics settings" (see section 1.3.3) – at least as long as we are talking about human drivers, instead of mere passengers (ibid.: 45).

2.2.4 Pluralistic approaches to responsibility

As we have seen above, there is a broad discussion on responsibility gaps in the context of human-machine interaction and different approaches how to bridge this alleged gap. But besides ideas of attributing responsibility to artificial agents themselves or finding institutional arrangements to circumvent this challenge, other scholars try to approach the problem from a different angle, reframing traditional conceptions of responsibility more fundamentally.

Daniel Tigard (2021) for example criticises that "moral agency" is often confused and equalised with "moral responsibility", emphasising that these are two different concepts. Moral agency requires certain capacities (see section 2.2.1). Moral responsibility further comprises relations to others and therefore is based on norms and expectations. Tigard calls this the "process view" of responsibility, associated with "holding responsible" – in contrast to the more traditional "property view" of responsibility, associated with "being responsible" (ibid.: 439f). This distinction for example allows us to revisit the discussion of responsibility gaps (see section 2.2.1). Sparrow (2007) denies responsibility for artificial agents based on the assumption, that robots cannot suffer. Actually, says Tigard, is Sparrow talking about moral agency, i.e. the "property view" and notion of "being responsible". If we apply the "process view" it is possible to think of punishment as a means of educational improvement or way of "holding others *to account*" (ibid.: 442; emphasis in original). Put this way it is possible for Tigard to avoid the alleged responsibility gap by granting some form of moral responsibility to artificial agents, without granting the status of moral agents.

As Tigard, also Mark Coeckelbergh and Janina Loh (2020) point to several drawbacks with our traditional concepts of responsibility in the "age of automation" – besides responsibility gaps and the challenge to adequately attribute responsibility to artificial agents, we also have to deal with growing space-time-dimensions, i.e. to allocate responsibility in a global society, facing global problems such as the climate crisis (ibid.: 11f). One way to address this challenge is to moderately try to rework the traditional concepts of responsibility – e.g. by allowing artificial or collective agents, networks or hybrid systems as bearers of responsibility instead of focussing on individual human subjects only. Another, more radical understanding of responsibility "does not start from the agent, person, or subject of responsibility, but from *the other*." (ibid.: 15) In this view, inspired by the works of Levinas, responsibility is primally understood as answerability and the relationality of responsibility is not secondary and in addition, but at the core and constitutive starting point. In the context of automation this approach allows us to question how artificial agents influence and mediate our responsibility towards other human agents and whether non-human agents like robots or animals qualify as "others" (ibid.: 18). Coeckelbergh explicitly distinguishes between "relational responsibility" and "moral responsibility", whereby the first is understood as a specific aspect of and addition to the latter (Coeckelbergh 2023: 2440). In the context of AI such a focus has several implications (ibid.: 2441): (1) It reminds us that discussions on AI tend do omit their social embeddedness, i.e. the question to whom we are responsible; (2) As AI provides data and knowledge influencing our decision-making, relational responsibility requires finding ways to cope with the "black box problem" (see section 2.2), i.e. how to explain our AI-mediated behaviour to others; (3) This view also points to the notion of "responsible innovation", i.e. a transparent discussion on development and use of AI; (4) Finally this relational approach highlights the political dimension of responsibility, i.e. a critical analysis of power structures in the context of social interactions between different groups of agents and AI (e.g. manufacturers, AV-drivers vs. pedestrians). Julio Andrade (2021) takes a similar line of argument, contrasting "centrist ethics" starting from a subject-position with a Levinasian ethics starting from a vulnerable "other", creating "an unconditional, non-reciprocal infinite responsibility for her." (ibid.: 126) This "other" is represented by his or her "face", symbolizing singularity and understood as meaning by itself. Andrade's claim is that discussions around AV mistakenly focus on trolley-problems, entailing the categorisation of road users into classes such as "pedestrians" or "passengers", moral calculations taking into account possible actions for AV, the potential infringement of traffic laws and implications of different ethics settings – which results in seemingly solving the problem of traffic safety in the context of AV, without ever addressing the constitutive "other". The fixation on trolley-problems tries to remove uncertainty, whereas a Levinasian perspective embraces uncertainty and ambiguity as necessary qualities of singular "faces". Seen from this side, Andrade comes to a somewhat paradoxical conclusion in analogy to Hans Monderman's concept of "shared spaces" and "naked streets": "In risking less-safe autonomous vehicles, we might make roads, and road users, safer in the process." (ibid.: 134)

Deborah Johnson (2015) argues, that discussions around potential responsibility gaps (see section 2.2.1) are based on false assumptions. One problem is that in many of the relevant papers the notion of responsibility is poorly developed and terms such as responsible, blameworthy, accountable or liable are often used interchangeably. She holds that responsibility in this context often is actually understood as accountability, and that "[a]ccountability-responsibility is embedded in relationships that involve norms and expectations." (ead.: 713) If it is based in relations and socially constituted norms and expectations, then "a particular technology does not necessitate a particular responsibility arrangement." (ead.: 713) Put differently: what we understand as responsibility is neither dictated by nature nor technology, but a result of social discourse and therefore it is up to society to negotiate for appropriate responsibility arrangements when it comes to artificial agents. As a second criticism on the notion of responsibility gaps Johnson reminds us of the productive power of discussions over possible futures: "Speculation, predictions, and especially visions, have rhetorical power; they can be used as a form of persuasion, to enroll others into activities that help make a particular future a reality." (ead.: 711) Talk about responsibility gaps therefore entices us to accept the idea of machines without clear responsibilities – although the field of science and technology studies clearly shows that instead of an unavoidable path, development in reality is "non-linear, multidirectional, and contingent" (ead.: 712). An example is a US-Department of Defence publication of 2012, stating that further developments of military technology will be focussed on robots supporting human decisions instead of robot's autonomy. Arguably such a shift of attention would have significant implications in terms of the funding for related scientific research. Consequently, it's up to us: "a responsibility gap may occur [in the future], but, if the gap occurs, this will be the result of human choices and not the inevitable outcome of the kinds of technologies currently in development." (ead.: 712).

Somewhat related to this discussion on the productive power and relevance of public discourse, Mark Coeckelbergh (2023) develops the notion of "hermeneutic" or "narrative responsibility". The basic idea is, that we as humans unavoidably have "the responsibility to make sense, to interpret, and to narrate." (ibid.: 2442) Whereas relational responsibility is a 2nd-person responsibility directed to others and causal as well as moral responsibility is a 3rd-person responsibility trying to identify responsible subjects, hermeneutic responsibility is a 1st-person responsibility forcing us (i.e. 1st-person singular as well as 1st-person plural) to interpret, understand and make sense of the world around us – which we commonly accomplish by telling stories, e.g. in the form of literature, newspapers or arts. Although artificial agents are not only objects, but also significantly influence these stories and our understanding and perception of the world – they cannot take this existential burden off us: "And in the end, it is our responsibility to decide what narratives we want to (co-)write—including narratives about AI, with AI, and sometimes against AI." (ibid.: 2446) If we as human individuals and collectives don't comply with this hermeneutic duty to make sense of our world including our interactions with AI, we'll find ourselves living in narratives someone else wrote for us, maybe even without realising it (ibid.: 2447). For the purpose of this paper this notion is important in so far, as scientific discourse and public stories on AV rarely ever talk about speed – whereas I want to argue, that this should be an essential aspect of the story.

2.3 Responsibility and autonomous vehicles

To conclude this section on responsibility in the context of AI and evermore autonomous robots, I want to turn my focus back to AV. Besides the focus on trolley-problems and general discussions on responsibility and AI, several scholars also directly discuss the question of responsibility in light of the introduction of AV. Obviously some of the lines of argument will be very similar to what was already mentioned in the forgoing sections – which is why this section might also serve as a brief recapitulation with a specific focus on AV.

To start with, Mark Coeckelbergh (2016) discusses whether the above-mentioned control and epistemic conditions as basic requirements for moral responsibility are met in case of AV. But interestingly he starts one step ahead, stating that actually it is also questionable to grant these conditions in case of manually driven cars: Do drivers really know what they are doing and are they really in control while driving at very high speeds? (ibid: 753f) For humans using AV in autopilot mode the case is clear and a direct relation to traditional moral responsibility vanishes: "The Aristotelian conditions for responsibility are not met." (ibid.: 754) In reality some form of comprise between full and no control presumably will be in place, at least as long as drivers have some possibilities to intervene or ex-ante set relevant configurations.

Alexander Hevelke and Julian Nida-Rümelin (2015) discuss different ways of ascribing responsibility for the use of AV. For manufacturers they suggest a "partial liability" as compromise between all-risks including product liability on the one hand and no liability on the other – since the former could prohibit investments in AV and the latter could lead to low safety-standards in these vehicles (ibid.: 620). For drivers and users of AV they discuss a potential duty to intervene as well as a strict liability. A duty to intervene is somewhat counter-intuitive to the whole idea of AV and can rather be seen as a kind of interim solution. If such a duty to intervene makes sense at all depends on chances to actually improve traffic safety. In reality this is questionable, since suddenly taking over control to perform challenging split-second decisions tends to overburden human capacities – and, as Hevelke/Nida-Rümelin hold, it is unfair to morally blame someone for a bad outcome, if she never had a fair chance to do what supposedly was her duty (ibid.: 625f). As to strict liability the rationale is simple (and holds true also for manually driven cars): the driver "took the risk of using the vehicle, knowing and accepting that it might cause accidents." (ibid.: 626) Despite rightful behaviour and complying with traffic rules – or in the case of AV: just using the car as a passenger without any chance to change driving settings – no driver/user can ever be "absolutely without fault", because they voluntarily and knowingly took a certain risk by using the vehicle in the first place. Nevertheless this risk is very small and partly socially accepted – which is why the authors propose a tax or mandatory insurance as the most practical solution (ibid.: 629).

Antti Kauppinen (2021) also starts his analysis from this small risk of using cars in general, but focuses on the role of moral rights when it comes to distribute unavoidable harm. The idea is simple: "when a person *lacks a right* against being harmed in a particular way due to something she has done, she is *liable* to such harm, and isn't *wronged* if she is so harmed." (ibid.: 633; emphasis in original). The term "liable" is used in a moral, rather then purely legal sense, but is on the other side also explicitly distinguished from "culpability". Even without any bad intention, "to be morally responsible for a threat, it is necessary to have voluntarily engaged in action that foreseeably poses a risk to others." (ibid.: 633) A further morally relevant distinction is between "option (bad) luck" and "brute (bad) luck" (ibid.: 636). A person voluntarily engaging in the (although minimally) risky action of driving a car has bad option luck if something happens. An innocent bystander or pedestrian hit by a car has bad brute luck. Based on these thoughts for an innocent bystander it would be morally permissible to harm a driver threatening to kill her, if this would save her own life – i.e. the driver isn't morally wronged, since he voluntarily took a risk and thereby lost his right compared to the innocent bystander. Since Kauppinen also sees this form of liability as a gradual concept, the bystander obviously should prefer other, less radical solutions and share some of the harm if possible. The other way round is a bit trickier, since it is questionable if risk-takers have the moral duty to sacrifice themselves. However this question is answered, the situation changes with AV: in this case the decision is taken by a third party (i.e. the software developers or the AV itself – if one is ready to allow this possibility) and therefore, says Kauppinen, it is clear that the AV-driver/user has to be the one to carry the can. "The deeper rationale for linking liability to moral responsibility is the thought that it is unfair if an innocent party has to suffer harm when it could fall instead on the person who brought about the wrongful threat." (ibid.: 634) Fiona Woollard (2023) has a similar approach to the problem, focussing on the distinction between doing and allowing harm – whereby the first is generally harder to morally justify than the latter (ead.: 50ff). One important aspect she points out, is that human agents only have rights against being harmed towards other humans, but not towards non-humans (e.g. natural disasters). From this the question follows, if AV are to be seen as human-like agents or rather as non-human-like things – which circles back to the question of agency discussed above (see section 2.2.1). Generally she conceptualises (at least manually driven) cars as "extension of the body". In dealing with AV we need to specify when it counts as identified with the driver and when the driver – or passenger – is identified as doing what the car does (ead.: 55). In the end, Woollard argues, the distinction between doing/allowing harm helps us to better understand the moral implications of our interaction with AV and these are different from the situation with manually driven cars.

2.4 Interim conclusion

As this discussion on responsibility in context of AI showed, there is no clear or generally accepted solution for the challenges posed by artificial agents to traditional concepts of responsibility. Nevertheless some interesting learnings for the purpose of this thesis can be drawn. One finding is that boundaries between the moral and legal side of responsibility in this context are fuzzy – as the example of alleged responsibility gaps shows: originally not further specified, later it was mostly interpreted as a liability problem and was reformulated as a genuinely moral one, i.e. as "retribution gap". Looked at from the legal side, the responsibility gap in context with AV seems solvable by combining strict and product liability, together with a mandatory risk-pooling insurance. Furthermore most legal scholars hold the belief, that liability will rather shift from AV-drivers/users to manufacturers.

One of the major takeaways from the moral side of responsibility is that virtually none of the above-mentioned scholars is willing to grant full responsible agency to current robots or AV (some leaving open this possibility for the future). Only List (2021) seems to argue in this line, but to my understanding his approach is rather meant as a kind of institutional liability arrangement (i.e. constructing a legal entity or person for artificial agents). Nevertheless it can be seen as in between legal and moral responsibility, as List is focussed on moral agency and tries to show that organised groups as well as artificial agents form some sort of agency "over and above" associated humans and that these agents have some genuine capacities to morally deliberate. Interestingly he doesn't combine group and artificial agency, which could reinforce his point. One could imagine artificial agents not only morally deliberating by themselves (however this is specified), but also via ethics committees and the like within the respective software companies (given, that this human deliberations subsequently would influence the behaviour of their artificial agents). Also the notion of "blank check"-responsibility, i.e. voluntary informed ex ante consent to accept responsibility prior to using an AV, seems to be located somewhere in between the legal and moral side of responsibility. On the one hand it obviously solves the liability problem, on the other hand it also helps to close the retribution problem. If someone knowingly and voluntarily keeps a dangerous dog and it attacks an innocent person, I think most of us would agree that the owner is not only liable, but also morally to blame. Of course this is closely connected to another point often emphasised, in the context of AI as well as in the traditional conception of responsibility: information, i.e. the epistemic condition. If the dog holder couldn't know about the danger (e.g. if it was a perfectly calm and well trained breed of dogs), she still has to bear liability, but we possibly wouldn't morally blame her for unexpected attacks. A similar point can be made for AV referring to the notion of "blank check"-responsibility. If an AV-driver knows about the high risk of his vehicle and/or has the possibility to set aggressive driving configurations in order to test out its maximum

– he will be liable and morally to blame, even if "the car itself" crashed into another car. He voluntarily accepted the risk and had bad "option luck". If another, cautious and well informed AV-user voluntarily and ex ante accepted responsibility and she rightfully believed in its very high safety-standards (suppose that up to this point no such AV was ever involved in any kind of mishap or collision) and she didn't have any chance to change the car's settings – I suggest she'd be (at least partially) legally liable as well, but only to a very low degree morally to blame (some blame rests on her for introducing the risk to society). Probably she would try to pass on legal liability to the manufacturer in terms of product liability. If the manufacturer (and software developers etc.) also behaved cautiously and were without fault (which is the situation List had in mind), the legal as well as moral dispute gets tricky. But as Nyholm (2018c), Miller (2011) and others point out, we don't have to speculate about potential moral responsibility of the artificial artefacts themselves, but should instead focus on the human agents involved. In part 4 of this thesis I want to argue, that there is one usually ignored but important factor in this context that can help us find a solution.

What this section also clearly showed, is that in the context of AI many scholars move away from isolated, individual-based conceptions of responsibility or "moral individualism". This entails an adaptation of the traditional control condition: instead of hanging on to direct and causal action as a necessary precondition, unpredictability and the lack of direct control are no longer seen as contradicting to moral responsibility. The common ground of notions such as extended, group, distributed or collaborative agency, hybrid, joint or sociotechnical systems, plural persons and joint responsibility as well as responsibility networks is, that in all these approaches more than one single human agent is conceptualised as being able to bear moral (and legal) responsibility, or put differently: some form of sharing moral responsibility among several relevant agents is allowed for. Nevertheless, as I already said, virtually all scholars – at least for the time being – want to ultimately trace back responsibility to one or more humans involved in the course of actions, at least as long someone can be said to have something like "meaningful human control".

The shift to pluralistic, relational and hermeneutic conceptions of responsibility emphasises the political dimension of this discussion. Responsibility is not only something we allow an isolated agent to bear or not to bear – it is also highly important how we as a society are holding this discussion, what stories we tell about AV, which aspects we highlight or omit and which role "the other" plays within the context of using AV. As we have seen in the first part of this thesis, speed is often ignored within this discussion. Interestingly, as Hey (2019) shows, speed and responsibility (in the form of strict liability) are in fact very closely connected in the context of traffic safety. Therefore, in the next part of this thesis I want to examine this factor speed in further detail.

3. Speed and traffic safety

I can't understand how this could happen. The curve could easily have taken 130.
Well, maybe I shouldn't have drunk the six quarters – but after six quarters you're not drunk, are you?
Nothing would have happened if that stupid tree hadn't been there.
"For a greener Vienna"… what a rubbish. Let them plant their shrubs somewhere else.
I wouldn't have minded, if they'd taken away my driving licence. We would have lived together in a wild marriage.
But now that I have to give you away for a price per kilo, this tears my heart apart.
[Refrain]
Yesterday my luck ran out. You're lying outside at the car cemetery, but you were always everything to me.
I just can't help crying, what a hot iron you have been. And all that's left, is just a total loss.

("Zweierbeziehung" by Rainhard Fendrich, 1980; my translation)[9]

The third part of this thesis is dedicated to examine the close relationship between traffic safety and speed – as this serves as my connecting piece between AV on the one, and responsibility on the other hand. Although at first glance it appears to be a rather neutral and technical topic, on closer inspection it is nothing less than a field of political debate and of ongoing cultural struggle about the dream of individual freedom. As the initial quotation reveals, the decision for a certain speed as well as one's fitness to drive and driving skills are subjects of self-confidence, just as causes for crashes are in the eye of the beholder – we'll encounter these topics again in the course of this part. Furthermore it hints to the historicity of public discourse and shows the intimate relationship that (some) man have with their cars, commonly interpreted as symbol of freedom.

In the first section of this part I want to outline the scope of the problem and as well give some political-historical context. After this I want to shed light on the relation between posted speed limits, actual travel speed and traffic safety. In the third section I'll have a closer look on crash causes and possible countermeasures in order to reduce the unbearably high death toll in traffic and briefly address some possible developments and opportunities when it comes to AV in this regard.

9 Original quote: „I kann gar net verstehen, wi das hat passieren können. Die Kurven hat do leicht 130 vertragen. Naja vielleicht hät ich die 6 1/4 net trinken sollen, aber nach 6 1/4 is ma do no net angsoffn, oda? Überhaupt nix wär passiert, wenn net der blede Baum da gestanden wär, für ein grünes Wien – aso ein Blödsinn. Solln's die Stauden woanders hinpflanzen. Gar nix hätt's ma gemacht, wenn's ma nur den Führerschein weggenommen hätten, hät ma halt in wilder Ehe zusammengelebt. Aber das ich dich jetzt um an Kilopreis hergeben muß, das reißt ma´s Herz auf. [Refrain] Gestern hat mich's Glück verlassen, du liegst am Autofriedhof draussen, dabei warst du doch immer ois für mi. I kann ma's Weinen net verbeißen, was warst du für a haßes Eisen und überblieben, is nur a Haverie." (see: https://www.fendrich.at/lyric/zweierbeziehung-2/)

3.1 Historical and political context of traffic safety and speed limits

The first known fatal car crash happened on August 17, 1896. A 44-year-old women was hit by a car with – according to an eyewitness – "tremendous speed" of 4 mph (approx. 6 kmh), the male driver was distracted by talking to a young lady passenger beside him (Shinar 2017: 3). More than a century later, in the year 2021 about 1,2 million people were killed in road traffic, a drop of 5 % compared to 1,25 million deaths in the year 2010. Road traffic is the leading cause of death for young people between 5 and 29 years of age and the 12th most important cause of death among all age groups. More than half of those killed are unprotected road users: 23 % pedestrians, 21 % motorcyclists, 6 % cyclists and 3 % users of micro-mobility (i.e. scooters). Occupants of 4-wheel-vehicles account for 30 % of traffic fatalities. In addition to the high death toll the estimated macroeconomic cost of road traffic accidents and injuries amount to 1,8 trillion USD or 10-12 % of global gross domestic product (WHO 2023: 4ff). Speed is a causal factor in about one third of fatal road crashes in high-income countries (WHO 2017: 5).

Despite these terrifying facts there is not much of an ethical debate around the issue. Leonard Evans (2008) found only one academic paper on ethics and traffic safety. In the mentioned paper Douglas Husak (2004) argues that even normal, everyday car driving poses a substantial risk to others – e.g. in September 2001 more people were killed in road traffic than in the 9/11 terrorist attacks. Obviously there are good reasons why we use cars, but at least if we use crash-incompatible vehicles posing an even higher risk to other road users without a justifiable reason (e.g. SUV, more than 90 % of which are never used off-road but due to their weight and design pose a higher risk for other road users) and drive for frivolous reasons (i.e. going for a ride just for fun), Husak concludes, driving a car is completely unjustifiable from an ethical point of view. One reason for our ethical blindness when it comes to traffic safety is that in public discourse, driving a car is often seen as an unquestioned necessity, as well as a symbol of personal freedom. Interestingly for the purpose of this paper, the media and advertising industry have important effects on traffic safety and speeding: "Television is probably the industry having the largest influence on safety. While drunks no longer appear in programming as likeable and humorous, as was once common, the media still portray reckless driving and speeding as exciting and acceptable." (Evans 2008: 3) Studies on car advertising in the US, Canada and Australia found that speed and fast driving is an important topic in about a quarter of them, car performance in general in about half, on the other hand safety aspects are only highlighted in about 10 %. Another analyses concludes "that risk-glorification, including speeding, is a common theme in much of the media, and [...] it has detrimental effects on risk taking inclinations." (Shinar 2017: 393) Nevertheless public opinion on traffic safety changed significantly over the past decades. In the 1970s only about 15 % of car drivers in the US used a seat belt and more than half of the fatal crashes were caused by drunk drivers. In 2012 already 86 % of US-drivers reported regular use of seat belts (in other countries this number is above 95 %) and the involvement of drunk drivers in fatal crashes has been reduced to 31 % (ibid.: 7f). Whereas the relevance of drinking and driving as well as non-use of seat belts have constantly declined, speeding has not and remains a major causal factor for fatal car crashes. Interestingly, "speed is possibly the most over-involved single factor in severe traffic crashes" (ibid.: 392), but at the same time the most difficult one to tackle. One reason for this might be found in the unrecognised schism between "beliefs people have about the dangers of speeding and their own acknowledged speeding." (ibid.: 392) A US-survey for example shows that exceeding the posted speed limits on highways by 15 mph (approx. 24 kmh) is "completely unacceptable" for 50 % of the people, but at the same time 68 % of them admit of having done exactly this in the past month (ibid.: 392).

Although many different actors (e.g. car industry and lobby, media, drivers etc.) play a role when it comes to traffic safety, "Government's ability to influence how drivers behave towers over that of other institutions. Government, unlike industry, has a core duty to prevent citizens from harming other citizens." (Evans 2008: 3) One of the most important influencing factors of Governments is setting the legal framework. A glance at history shows that the legal framework for road traffic is coined more by political considerations than pure concern for traffic safety and public interest. For example speed limits have been introduced or tightened in many countries at the beginning of the 1970s not as measure to improve traffic safety but in the wake of the oil crisis in order to cut fuel consumption (ITF 2018: 9). In the US speed limits traditionally have been the domain of federal states. 1973 US Congress tied public funds for highway construction to the acceptance of a National Maximum Speed Limit (NMSL) of 55 mph (approx. 89 kmh). In 1987 this limitation was relaxed to 65 mph (approx. 105 kmh) for rural interstates and 1995 it was dropped altogether – with the consequence that many states re-raised their speed limits (Shafi/Gentilello 2006: 719). Also Germany, homeland of cars with combustion engines, is an insightful example of how closely interwoven the history of traffic legislation with politics is, as Leon Birck (2024) shows in a detailed analyses. Probably the first official speed limit was issued in the city of Mannheim, home town of Carl F. Benz, in the year 1893 at 6 kmh in and 12 kmh out of town. At the turn of the century about 30 different and diverging regional traffic regulations existed within the German Empire. The first general and nationwide speed limit in Germany was set at 15 kmh – the average horse trotting speed – for built-up areas in 1909, a general speed limit for rural areas didn't exist for cars. More than a decade later in 1923 the limit for cars was raised to 30 kmh for built-up areas. 1934, after the Nazis' seizure of power, a completely new legislation was issued with an explicit political goal in its preface: "The promotion of the motor vehicle is the goal set by the Reich Chancellor and Führer, which this order is also intended to serve." (quoted after Birck 2024: 32, my translation).[10] Rules generally favouring motor vehicles (e.g. cars were given priority over non-motorised road users at intersections) were enacted and all speed limits for cars in and out of towns were cancelled. Interestingly Hitler had to give in to the reality of an increasing number of traffic crashes and fatalities and re-introduced a speed limit of 60 kmh in and 100 kmh out of town in May 1939, eventually even reducing it further to 40 kmh and 80 kmh respectively in October 1939 – just one month after Nazi-Germany's declaration of war on Poland. Hitler himself explained this move at the International Motor Show Germany in February 1939 in Berlin: "The German people are currently sacrificing as many lives to traffic in every six years as the Franco-Prussian War of 1870/71 claimed in total. This is an intolerable state of affairs." (quoted after Birck 2024: 34, my translation).[11] In 1952 the speed limits of 1939 were abolished as a form of coming to terms with the Nazi-legislation – without mentioning that the status quo ante of 1934 was also introduced by the Nazis. From the 1950s onwards, the car became a national symbol for progress and freedom in Western Germany – a particularly liberal legislation in this regard was seen as an important differentiation from collectivist Eastern Germany and the harsh Nazi past. Even proposed measures to improve traffic safety were vilified as "rampage of traffic legislation" by Hans Bretz, then president of Germany's largest car lobby organisation ADAC (ibid.: 36). In 1957, after intensive debates, a speed limit of 50 kmh for all motor vehicles in built-up areas was implemented. On rural

10 Original quote: „Die Förderung des Kraftfahrzeugs ist das vom Reichskanzler und Führer gewiesene Ziel, dem auch diese Ordnung dienen soll."

11 Original quote: „Das deutsche Volk bringt zur Zeit in je sechs Jahren genau so viele Tote dem Verkehr als Opfer wie der deutschfranzösische Krieg 1870/71 im Gesamten an Gefallenen gefordert hat. Dies ist ein unerträglicher Zustand."

roads speed for cars remained unlimited for further 15 years – which was changed only due to a clever move. Social Democrat Minister of Transport Georg Leber introduced a speed limit of 100 kmh for rural roads in 1972 as a large-scale trial to be evaluated. Due to its success it was permanently implemented in traffic legislation in 1976. As a reaction to the oil crisis, car-free Sundays were decreed in November and December 1973 and temporary speed limits were set to 100 kmh on motorways and 80 kmh on rural roads. Because the number of road fatalities and injuries subsequently has fallen sharply by 61 %, the Social Democrat led government wanted to maintain speed limits and made a compromise proposal of 130 kmh on motorways – but didn't prevail, which is why the mandatory speed limit on motorways was abolished after 111 days and replaced by a still valid voluntary recommended speed of 130 kmh. A decisive factor within the public discourse was a massive ADAC lobby campaign under the slogan "free citizens demand free travel" (ibid.: 37).[12]

In Eastern Germany the picture was very different – not the car, but public welfare and safety were given priority in traffic legislation. Already in 1956 traffic regulation included a strict alcohol ban and speed limits of 50 kmh in built-up areas, 90 kmh on rural roads and 100 kmh on motorways. In 1979 the speed limit for rural roads was reduced to 80 kmh. From an eastern perspective the western "free travel"-policy was seen as an inhuman and profit-driven capitalist system. As of 1.1.1992 western standards were declared universally valid for the whole of Germany in the course of reunification and conservative Minister of Transport Friedrich Zimmermann published a newspaper commentary titled "Free travel instead of socialism", stating that "socialism has always had a dysfunctional relationship with the car [...]. Reunification also means an end to socialist road transport ideologies." (quoted after Birck 2024: 39; my translation)[13] Interestingly a survey in 1990 by a government-owned research institution showed that 89 % of the population of Eastern Germany were in favour of a continuation of the existent speed limits at least in the eastern part of Germany and 73 % supported an expansion to the whole of Germany. Regardless of this, western traffic legislation was imposed and the attitude of German Federal Governments hasn't changed much until today. Former conservative Minister of Transport Andreas Scheuer vilified speed limits on motorways 2019 as a "fetish" and "against all common sense"

12 Original quote: „Freie Bürger fordern freie Fahrt."
13 Original quote: „Freie Fahrt statt Sozialismus. […] Und der Sozialismus hat von je her ein gestörtes Verhältnis zum Auto […]. Wiedervereinigung bedeutet auch Schluss mit sozialistischen Straßenverkehrsideologien."

(quoted after Birck 2024: 40; my translation).[14] In the current coalition government the liberal Freedom Party consistently blocked attempts by Social Democrats and Green Party to introduce a speed limit on motorways, with Minister of Transport Volker Wissing stating that “driving a car means freedom” (Block/Uhlenbroich 2023).

Excursion: Jaywalking

As a short excursion I want to briefly take a look at the insightful history of “jaywalking” – which is at least indirectly connected to speed, but directly connected to traffic safety and questions of liability and responsibility when using public streets.[15] Before the mass proliferation of cars, public streets were for everyone: pedestrians, pushcart vendors, horse-drawn vehicles, as well as playing kids. In the 1920s this picture changed dramatically – as can be seen in Illus. 1 (Stromberg 2015).

Illus. 1: Manhattan's Hester Street in 1914 (left side); Midtown Manhattan in 1925 (right side)
© Maurice Branger/Roger Viollet/Getty Images (left side); Edwin Levick/Getty Images (right side)

At the beginning of the 20th century, crossing a street was simple – you just crossed it wherever you wanted. As the number of pedestrian fatalities skyrocketed parallel to the spreading of cars, the latter were seen as a thread to public safety, as dangerous intruders. In media coverage car drivers were presented as the obvious culprits and sometimes depicted as the Grim Reaper. Cars were seen as “frivolous playthings” or vilified as “pleasure cars”, whereas monuments were erected in memory of the many killed kids. In New York around 10.000 kids went out on the streets during a “safety week” in 1922 to demonstrate for their right to access public streets, including a

14 Original quote: „gegen jeden Menschenverstand“

15 If not quoted explicitly, this excursus refers to Thompson (2022), Stromberg (2015), Lewis (2014) & Norton (2007).

separate group of 1.054 kids symbolizing the number of killed kids on New York streets in the previous year alone. At courts, judges and juries typically conceded guilt to those with the larger vehicle and charged drivers with manslaughter for fatalities, irrespective of the specific crash circumstances. The potential for conflict was essentially based on differences in speed. By design, cars were massive, flexible and could go faster than pedestrians, bicycles and horse-drawn vehicles – it was their major selling point. In reaction to the many traffic fatalities, public authorities tried to regulate them down to "pre-motor age speeds" – which didn't work out, because it was in contradiction with their basic competitive advantage (Norton 2007: 338f). As a reaction to poor compliance with urban speed limits, demands to mechanically prevent cars from infringing them were voiced. A petition in this regard in 1923 in the city of Cincinnati eventually became a tipping point in the historic struggle for public streets. Although 42.000 people supported the petition, it failed – and once and for all rung the alarm bells of the car lobby. The outrage went so far, that even car sales were affected – after steady increases, sales figures dropped by 12 % in 1924 (Thompson 2022). Understandably, the automotive industry feared for their business and they forcefully tried to turn the tables: "jaywalking" was identified as effective counter-strategy and campaigned nationwide.

Before the 1920s the term "jay" meant a "country hayseed out of place in the city" (Norton 2007: 342). A "jaywalker" therefore was someone from the countryside who didn't know how to behave in a city, staring at the lights and show windows, getting in the way of other road users. When it first came up in den first two decades of the 20th century, the term was understood as offence for pedestrians – who responded by using terms like "jay drivers" and "joy riders". In the 1920s "jaywalking" became a widely used expression and even made it into official documents of city administrations. After the petition in Cincinnati, the car lobby realised that it could only win this struggle proactively. The role model for their jaywalker-campaign was "Otto Nobetter", a fictional character invented by the industrial safety movement as a reply to stricter workmen's compensation laws. Otto Nobetter was depicted as clumsy and reckless workmen, used to shift the blame for work accidents away from employers towards workmen (ibid.: 341f). Similarly, the car lobby produced traffic safety publications blaming jaywalkers for crashes and put them into nationwide circulation, supported by police, oil companies and groups of Boy Scouts – back then a "mainstay of public safety campaigns" who distributed leaflets to educate pedestrians by the thousands (ibid.: 345). They also sponsored school safety campaigns and poster contests, warning against jaywalkers and emphasising that public streets were not a safe place for kids. They financed the implementation of crossings and warning sings. Since slowly spreading and existing regulation in terms of restricting pedestrians were not seen as effective, they strategically tried to associate jaywalking with shame – e.g. by using this derogatory and ridiculing term, by persuading police to whistle and shout at "jaywalkers" or by organizing local safety parades, in which "jaywalkers" were portrayed as clowns dressed in old-fashioned clothes, repeatedly being rammed by a slow-moving car. Another strategic move was to influence media coverage on road traffic crashes. At the beginning of the 1920s newspapers talked about "Automobile Deaths" in the context of road fatalities (ibid.: 354f). To change this negative framing, the car lobby started to place adverts. In 1923/24 they changed their strategy: instead of buying space in newspapers, they launched a service free of charge for journalists. After submitting some details for local crashes, journalists received a ready to print crash report including assessment and safety recommendations – with jaywalking unsurprisingly as the main cause for any crash.

The jaywalking-campaign undoubtedly was very successful. Still in the year 1924 a judge in Philadelphia warned: "It won't be long before children won't have any rights at all in the streets." (ibid.: 338) Just one year later, Los Angeles pioneered encompassing legal restrictions for pedestrians. In 1925 a new traffic ordinance was implemented, fining and forcing pedestrians to use crosswalks and to obey traffic signals, "where there were none, they would have to raise a hand to halt oncoming motorists." (ibid.: 352) Based on the Los Angeles sample, in 1928 a nationwide "Model Municipal Traffic Ordinance" was introduced, which many cities put into effect the years to come (Stromberg 2015). Even a century later, this upside down interpretation of traffic safety is still common. In 2013 the Los Angeles Police Department for example stated on their website, that pedestrians had been "impeding traffic and causing too many accidents and deaths" (Lewis 2014). As this excursus tried to show, such a statement surely would have caused astonishment a century ago. What's more, this change in discourse arguably also has implications for questions of responsibility – as a statement by Ray Thomas, a lawyer in the city of Portland specialised in traffic law aptly points out: "The difference is that no jaywalking pedestrian ever ran down and killed a driver [...]." (quoted after Lewis 2014).

3.2 Speed limits, travel speed and traffic safety

Speed limits themselves obviously do not translate into a reduction of traffic fatalities if they are not adhered to. Empirical studies show that a reduction of the posted speed limit by 10 kmh results in an average reduction of travel speed of about 2,5 kmh (Hydén 2020: 46; Shinar 2017: 417). A Study analysing changed speed limits in Sweden during 2008 and 2009 on approximately 20.000 km of rural roads and motorways also found an average decrease of average travel speeds of 2-3 kmh for reductions in posted speed limits by 10 kmh and an increase of 3 kmh for increased speed limits by 10 kmh – resulting in significantly less fatal crashes and a slight but somewhat ambiguous reduction in traffic injuries (Vadeby/Forsman 2018). Studies accompanying the abolition of the NMSL in the US (see section 3.1) also showed an increase of actual travel speed of about 3 mph (approx. 5 km) for increased posted speed limits by 10 mph (approx. 16 kmh; Shinar 2017: 418) – with a negative effect on traffic safety: the fatality rate per vehicle miles travelled was 31 % higher in states which increased their speed limit over 65 mph compared to states with speed limits of 65 mph and less (Shafi/Gentilello 2006: 720f). Summing up, most studies on changes of posted speed limits show – although to a lesser degree – changes of average travel speeds as well as number and severity of crashes in the expected direction (Shinar 2017: 418). Speed limits obviously are political decisions seeking a compromise between safety and individual freedom, but there are several good reasons why they are necessary: drivers tend to underestimate the effect of speed on traffic safety, crash avoidance and the probability and severity of crashes as well as energy consumption, they tend to overestimate the effect of speed on travel time, and posted speed limits tend to harmonize traffic flow since individual assessments of adequate speed levels vary significantly (Shinar 2017: 426f).

Besides speed limits many other factors influence actual travel speed, most importantly physical characteristics of the road, roadside interference, the weather and other vehicles (Shinar 2017: 393). Nevertheless, "[s]peed control is the most important aspect of promoting road safety world-wide. […] Speed has the great advantage that it relates to all potential risks irrespective of environmental factors, infrastructure, road users' status or any other risk-related factor." (Hydén 2020: 44) On the other hand "[speeding] reduces the time available to the driver to recognize a hazard on the roadway, minimizes his/her ability to undertake an avoidance maneuver, increases the distance needed to stop the vehicle, increases the severity of crash due to a higher kinetic energy, and reduces the ability of safety systems, such as restraints and roadway hardware, to protect the occupant." (Shafi/Gentilello 2007: 721) It follows and has been shown in numerous studies over the past decades that speed has a direct influence on crash occurrence and crash severity (ITF 2018: 7). This relationship was analysed empirically already in the 1980s by Nilsson, who came up with a "power model" as a rule of thumb for the relation between speed and crashes: every increase of average travel speed of 1 % leads to an increase of all injury crashes of 2 %, crashes with severe injuries of 3 % and fatal crashes of 4 % (ITF 2018: 16). In the 2000s Rune Elvik has developed this approach further and added an "exponential model" after analysing 115 studies, finding that initial speed levels also have to be taken into account since they significantly effect the outcomes in terms of injuries and fatalities – i.e. the higher the initial speed level, the higher the effect on fatalities. In the end, both models work similarly well to capture the relation between speed and traffic injuries (Elvik 2013; Elvik 2005; ITF 2018: 17ff). In a later study Elvik et al. confirmed the validity of this relationship for more recent empirical data (Elvik et al. 2019).

As already mentioned above, speed is positively related to crash risk and severity. The relation to crash risk is explained by a shorter reaction time to changes in the environment as well as a reduced manoeuvrability of the vehicle with higher speed (ITF 2018: 7; Aarts/Schagen 2006: 215). Another relevant factor for the crash risk is the variance of speed, i.e. the higher the deviation from the average speed level, the higher the crash risk due to an increased number of critical encounters (i.e. overtaking) with other cars. An often quoted study of the 1960s suggested an increased crash risk also for below-average speeds. Later studies found that this effect was mainly due to the (questionable) inclusion of braking and turning vehicles, whereas the correlation with other, normally below-average driving cars was statistically irrelevant (ITF 2018: 21f; Shinar 2017: 410ff). Although the relation between speed and crash risk is complex and sometimes even counter-intuitive[16], the bulk of the existing evidence indicates that speed generally contributes to crash risk and specifically so for certain types of crashes (e.g. "run-off-the-road" crashes), that cars with above-average speed are significantly more often involved in crashes and cars with below-average speed as well but rather due to other risk factors (such as braking or turning; Shinar 2017: 415f; Aarts/Schagen 2006: 222f).

16 David Shinar points to the somewhat paradoxical fact, that in dense city-traffic many so called fender benders occur at low average speeds, whereas on rural roads despite high speeds only few (but very severe) crashes happen – whereof statistics would suggest that less speed leads to more crashes (Shinar 2017: 401).

When it comes to speed and crash severity things are unequivocal and well explained by the physical laws of kinetic energy.[17] Impact force not only increases with the vehicle mass but also with the square of the speed (ITF 2018: 7). Empirical studies convincingly prove the practical relevance of this theory (Shinar 2017: 420f). Although there is a debate about the exact fatality risk figures for pedestrians hit by a car: "The impact speed of the striking car is widely accepted as a prime factor for the injury risk in car-to-pedestrian collisions." (Rosén et al. 2011: 25). Empirical research of the 1980s and 1990s suggested S-shaped pedestrian fatality risk curves with a very steep increase between 30 and 50 kmh and fatality risks at 50 kmh between 40-90 % (Rosén/Sander 2009: 536; Kröyer et al. 2014: 147). Newer studies found methodical inadequacies (e.g. a strong data bias towards severe injury crashes, leading to a disproportionately high share of fatalities) and criticised the data quality (i.e. too old and too small sample sizes) of the older literature and suggested significantly lower fatality rates (ibid.). One newer study is Rosén/Sander (2009), who use one of the most encompassing and detailed data sets available with more than 2.000 car-pedestrian crashes in Germany from 1999 to 2007. From this they derive a fatality risk function with a fatality risk of about 2 % at a vehicle speed of 30 kmh, 10 % at 50 kmh, 20 % at 60 kmh and 60 % at 80 kmh. Other recent studies show similar results, but with a rather wide range of variation (Kröyer et al. 2014: 145f).[18] One reason for this are different inclusion strategies in terms of data selection. It makes a big difference if e.g. SUV, children and/or elderly are included in the data sets, because besides impact speed, other factors such as vehicle types and characteristics of pedestrians are also relevant for the fatality risk – e.g. elderly people are more likely to die in collisions at lower speeds than younger people (Kröyer et al. 2014: 143, 151). Conversely the fatality risk for child pedestrians is significantly lower compared to pedestrians aged 15 or older (Richards 2010: 24). However, fatality risk for pedestrians hit by cars strongly correlates with impact speed, with a fatality risk at an impact speed level of 50 kmh being more than twice as high as compared to an impact speed level of 40 kmh and more than five times higher compared to an impact speed level of 30 kmh (Rosén/Sander 2009). The relation between speed and crash severity obviously also holds true for the occupants of a car. For car occupants it makes a major difference if a crash is head-on or from the side, since bonnets offer a large crush zone and other restraint

17 $E = ½ mv^2$ with E = kinetic energy, m = mass, v = velocity (for further information see: https://en.wikipedia.org/wiki/Kinetic_energy)

18 Fatality risk at vehicle speed of 30 kmh between 0,6 and 5,8 %, at 50 kmh between 6,3 and 25,6 % and at 70 kmh between 31,4 and 82 %.

systems (e.g. seat belt and airbag) increase protection, whereas in case of side-impacts there is only little space between seating position and the outside of the door. As a consequence, fatality risk for belted car drivers is about 3 % in a head-on crash at 30 mph (approx. 48 kmh), compared to 25 % in case of a lateral impact.[19] At 40 mph (approx. 64 kmh) driver fatality risk is about 17 % for head-on crashes, but already 85 % for side crashes. For head-on crashes at 50 mph (approx. 80 kmh) fatality risk for drivers is about 60 % and for crashes at 70 mph (approx. 113 kmh) fatality risk approaches 100 % (Richards 2010: 21f, 26). So with respect to the relation between speed and crash severity there is no doubt: the risk of severe injuries and fatalities significantly increase as a function of speed (Shinar 2017: 425).

Summing up the different aspects and relations of speed, crash risk and crash severity it has become clear that the higher the speed, the more risk is involved and especially so when it comes to severe crashes and fatalities – put differently: "Speed is likely to be the single most important determinant of the number of traffic fatalities." (Elvik 2005: 69)

3.3 Crash causes and traffic safety potential through AV

Different to drink-and-drive and non-use of safety belts, speeding has remained a major cause for fatal car crashes in high-income countries all over the world in the last decades with a constant share of about 30-35 % (Shinar 2017: 392; WHO 2017: 5). In the United Kingdom it is responsible for about 28 % of road fatalities, in Australia for about 30 % (WHO 2017: 5). In the US speeding generally accounted for 28-32 % of traffic fatalities in the years from 2001 to 2014, with different shares depending on personal characteristics – e.g. for the group of young men (up to 25 years of age) it accounted for about 40 % of traffic fatalities, for elderly (75 years and more) only for less than 10 % (Shinar 2017: 401f). In Austria speeding was the suspected main cause in 26 % of fatal crashes in the year 2023 (BMI 2024).

Speeding, as well as inattention and distraction, failures to give way, misjudgements, fatigue or alcoholisation fall into the category of human errors. An often quoted number, based on US crash

19 To be precise, these numbers refer to speed measured in delta-v, i.e. the difference between pre/after-crash speed. This measure is well suited for car crashes, because it takes into account characteristics of the vehicle (e.g. weight, stiffness etc.) as well as initial and final speeds of the vehicles involved. For example: if two cars with same weight crash head-on at an initial speed of 30 mph, due to the conservation of momentum delta-v is 30 mph. If one of the cars is stationary, it will give way and delta-v is 15 mph each. The numbers given in the text therefore are valid for crashes with fixed, immovable and heavy objects (e.g. walls, pillars, big trees etc.; Richards 2010: 19).

data from 2005 to 2007, states that human errors account for 94 % of car crashes, whereas only 2 % can be assigned to vehicle component's failure such as faulty tires or braking systems and another 2 % to environmental factors such as slick roads, obstructions to driver's vision or bad weather conditions (NHTSA 2008/2018).[20] Other studies supported this finding with numbers around 90 % or higher (Shinar 2019: 224). Within human errors the NHTSA assigns 41 % to recognition errors (i.e. inattention/distraction, inadequate surveillance etc.), 33 % to decision errors (i.e. speeding, illegal manoeuvres, misjudgements of gaps or other's speed etc.), 11 % to performance errors (i.e. overcompensation, poor directional control etc.), 7 % to non-performance errors (i.e. falling asleep etc.) and 8 % to other human errors (ibid.). With respect to individual factors, speeding is especially widespread among young men and is often based on overconfidence. Interestingly one study found a correspondence between speeding and the amount of gambling within the group of young drivers (Shinar 2017: 395ff). Another study shows that speeding depends roughly as much on the intention to speed as on the habit of speeding, which "means that much of our speeding behavior is habitual, essentially involving no conscious decision." (ibid.: 397) However, caution is advised when determining the causes of crashes – because: "Crash cause is similar to beauty: it is in the eye of the beholder. […] A crash cause is not something we can see, but something we surmise from what we see." (Shinar 2019: 224) To determine a crash cause starkly depends on the research method and who you ask – as can be seen from an example in which a data set of 148 fatal crashes has been analysed by different professionals: policemen saw about 99 % of them caused by human failures, roadway engineers found road defects as major causes in 76 % of the same cases (ibid.: 224f). The common interpretation of "cause" within the traffic safety community is the deviation from the norm and hereby the focus is usually on human failures – whereas e.g. the lack of a roundabout hardly ever is mentioned as a cause for crashes, although it is scientifically well confirmed that roundabouts reduce the number of crashes (ibid.: 228). As a constructive way out, David Shinar suggests to look at crash causes from the perspective of viable countermeasures. Especially the "safe system" approach is promising in this regard, i.e. a safe, error-tolerant, self-explanatory road design (e.g. narrow streets, bumps, sharp turns etc.) focussing on a fair balance between all road users – because: "Human causes can often be best eliminated through vehicle and/or environmental

20 The missing 2 % are other or unknown factors.

changes. In the absence of vehicle and roadway modifications, attempts to directly force drivers and pedestrians to change their behaviors involve enforcement: a Sisyphean, expensive, never-ending, cat-and-mouse game." (ibid.: 228)

When it comes to speed and crashes in the context of AV, I have already discussed two relevant approaches above (see section 1.3.4). The first insight was that speed significantly influences a car's manoeuvrability and therefore to disarm trolley-dilemmas, AV should be programmed to apply braking in a straight line as standard emergency strategy. The second insight started one step earlier, concluding that in order to avoid quandaries in the first place, speed should depend on an AV's operational capabilities – i.e. given the situational assessment of an AV's cameras and other sensors, speed is limited to the level that can be safely managed. Combs et al. (2019) furthermore analyse the potential performance of AV in detecting (and subsequently preventing) potential crash situations with pedestrians compared to human drivers. They use a US data set with more than 5.000 traffic pedestrian fatalities from the year 2015. From 4.241 transportation-related fatalities only about 80 % were labelled as preventable – excluding crashes with drowsy and impaired drivers as well as "pedestrian dart-outs" (ibid.: 3). The analyses focussed on three sensor technologies: visible-light cameras (VLC), light detection and ranging (LiDAR) and radar. As a result, AV equipped with VLC only could have detected between 21-51 % of the preventable fatalities, VLC and LiDAR together could have detected about 70-86 % and all three technologies combined could have detected more than 90 % of these critical situations leading to fatal crashes (ibid.: 4f). Although including the crashes labelled unavoidable into the calculation would reduce these shares significantly, this study clearly shows the high potential of AV to reduce fatal traffic crashes involving pedestrians.[21]

3.4 Interim conclusion

In this part of my thesis I tried to shed light on the relation between the factor speed and traffic safety – since the latter is the most important and commonly supported, as well as advertised target for the development and introduction of AV (see section 1.3.1). As I have shown, speed is a

21 Since fatal crashes caused by drowsy or impaired drivers and – at least to a certain degree – crashes due to „pedestrian dart-outs" (which are obviously closely related to the speed of the approaching vehicle) arguably can and should be seen as preventable too, it seems fair to use all transportation-related fatalities as a denominator – leading to a potential AV detection-rate of around 70 % of all transportation-related fatalities instead of 90 % of all "preventable" crashes.

major cause for the probability of severe crashes and undoubtedly the single most important factor when it comes to fatal crashes. The higher the speed, the more injury crashes, the more severe the consequences, the shorter the time to react to critical events and the lower the manoeuvrability of a car. Whereas other safety relevant factors such as the prevalence of drinking and driving or non-use of safety belts have been declining in the past decades, speeding remains a constant and particularly difficult to solve problem. Coming back to trolley-problems and other ethical issues discussed earlier (see section 1.3), it clearly shows that it really does make a major difference, whether we are talking about critical encounters with speed levels around 30 kmh or 80 kmh. For the former case pedestrian fatality risk rates around 2 % are involved, for the latter fatality risk amounts to around 60 % (see section 3.2). Elderly pedestrians are at a disproportionately higher fatality risk also for crashes with lower speed levels, and at lower speeds the manoeuvrability of cars is much higher – i.e. avoidance manoeuvres such as swerving might only be an option up to a certain speed level, whereas it amounts to an unjustifiable hazard at higher speeds (see section 1.3.4 and 3.3).

Furthermore it has been shown that 9 out of 10 car crashes commonly are attributed to human errors. As the study of Combs et al. (2019) shows, AV possibly could detect most of these critical situations in advance and take adequate measures to avoid many of these crashes. Many human risk factors such as inattention and distraction, misjudgements and performance errors, disobeying of traffic rules, fatigue and different forms of impairments (e.g. alcoholisation) most certainly could be remedied by AV. If all these human-related risk factors are eliminated, speed remains as most relevant critical factor for traffic safety – proportionately even gaining in relevance. Who takes the “tactical decision” on speed, human or machine, arguably makes a major difference for traffic safety (see section 1.3.4 and 3.3) – but the principal importance of the factor speed nevertheless remains unaffected by this. Regarding the criticism of exaggeratedly high and somewhat myopic attributions of crash causes to human errors, I argue that this criticism is still valid but at least loses weight in the context of AV. Unclear road junctions with obstructed sight might also pose a serious challenge for AV – although probably a smaller one as for human drivers, given that at least aggressive driving, distraction, impairments and other human risk factors are eliminated in the first place.[22] Other infrastructural as well as regulatory framework conditions related to traffic safety mainly apply to human drivers – i.e. it seems fair to assume that AV ("unless, of course, specifically so ordered by it's pilot" – see initial quotation in part 2) don't infringe speed limits and do not need speed bumps and narrow roads to do so. However that may be, I hold that with the further introduction and spreading of AV, the decision on speed remains one of the most relevant factors for traffic safety and will even gain in importance – regardless of who makes this decision.

What this third part also clearly has shown, is the political origin and nature of posted speed limits. Speed limits are definitely no neutral, exclusively public interest orientated regulations, set according to scientific findings – but rather products of political negotiation and historical context. Not concerns for traffic safety, but the oil crisis was the major factor in driving the discussion in the 1970s and for some, regulation aimed at road safety amounts to a "rampage of traffic legislation", and speed limits on motorways are "against all common sense" (see section 3.1). Against this background it is clear that public authorities, i.e. the lawgiver, obviously play an important role in any "traffic responsibility network" (see section 2.2.3). Again, when it comes to AV it can be assumed, that posted speed limits will affect traffic safety more directly – as AV hopefully will comply much better with legal standards than humans currently do (see section 3.2). This at least applies to scenarios were both, manually driven cars and AV, form part of the road traffic. For "AV-only" scenarios and regarding the "functional safety" approach (see section 1.3.4) one might question whether we still need speed limits at all – put differently: Are speed limits for humans? I will discuss this and other open questions regarding responsibility in the remainder of this thesis.

22 To be fair, one has to admit the possibility for additional, AV-related risk factors in such unclear situations – e.g. failures to correctly identify other relevant road users (see section 1.2.2). However, it can be assumed that these mistakes significantly decrease over time, since every single incident will (hopefully) contribute to further safety improvements for all AV (at least for all AV supervised by the same manufacturer, running on the same software).

4. Synthesis: speed and responsibility for AV

Manually driving a car is a straightforward case for traditional conceptions of responsibility. As long as I know what I'm doing (i.e. I'm aware of the risks involved and I know how to properly use the car), I'm in control of the car (i.e. the car directly obeys my commands), I'm using it voluntarily (see section 2.1) and technical defects are excluded – I'm the one responsible for any mishap. As we have seen above, responsibility for traffic safety when using a car basically consists of two components: (a) the introduction of a non-negligible risk to others into society and (b) wrongful behaviour. Both components are related to the factor speed. The first one is conceptualized by the term "strict liability". As I have pointed out above, strict liability for cars historically was explicitly related to the high speeds these back then new and massive vehicles could reach. Still in today's traffic codes we find this connection between design speed of a vehicle and strict liability (see section 2.2.2). In regard to the second component, vehicle speed is one of the most important factors in terms of traffic safety in general and specifically for severe crashes, as I have shown above (see section 3.2). Therefore it seems quite obvious that responsibility for traffic safety is closely related to the factor speed.

As we have also seen above, responsibility furthermore is related to prevailing, societal norms (see section 2.1). The short excursion on the history of "jaywalking" (see section 3.1) can be seen as a good example of this. In the beginnings of the 20th century, well before the motorization of the masses, "pre-motor age speeds" were the prevailing norm on public streets. Cars consequently were seen as dangerous intruders under general suspicion in regard to skyrocketing traffic fatalities. It was simply abnormal to drive faster than horse trotting speed of around 15 kmh. Traffic laws and speed limits were implemented to protect this normality. Few decades later this norm was turned on its head: cars and their speed became normal, whereas pedestrians were declared obstacles and risk factors. Supported by political interests and campaigns throughout the past century, this latter version of normality is still dominant in most of public opinion today. As already stated above, such societal norms are closely related to our understanding of responsibility. Whereas at the beginning of the 20th century, cars – irrespective of the specific circumstances – were seen as blameworthy and liable for any harmful encounter with a pedestrian, nowadays the statement "pedestrians cause crashes and deaths" (see section 3.1) is not only linguistically conceivable without contradiction, but also a legal reality: if a pedestrian steps on the street for no reason, it is their own fault that they are run over.

When it comes to AI and AV, traditional conceptions of responsibility reach their limits. Both the epistemic-condition as well as the control-condition are called into question. In times of machine learning and neuronal networks sometimes not even experts and the programmers, let alone untrained users are able to reliably predict the actions of AI, i.e. the programmers themselves might not really know what they are doing when using such a device, robot or AV. Also the control-condition must certainly be abandoned, as the whole idea of using an AV is based on giving up control. One opinion largely shared by legal scholars following from this, is that responsibility will foreseeably shift from users to manufacturers of AV (see section 2.2.2).

Although the discussion on responsibility becomes more complicated in the context of AI and AV, their utilisation and actions are always on behalf of someone. As Sven Nyholm (2018c; see section 2.2.3) points out, responsibility in the context of AI can be based on the nature of this relationship – i.e. even if we lose direct control as in the case of a manually driven car, we still have responsible agents to whom the AV (albeit with some uncertainties) indirectly obeys. Following this line of thought we can think of different categories of AV-users, depending on the extent of command authority they exercise when using an AV: the driver, the user and the passenger. I will discuss some implications of this for the question of responsibility below (see section 4.3). Before that, I want to work out in more detail the role of speed in this context. First I discuss the relevance of the barely mentioned factor speed in AV-related "trolleyology" (see section 4.1). Then I would like to give some good reasons why I think speed is a relevant and useful factor in answering my research question: Who bears responsibility in case of accidents with autonomous vehicles? (see section 4.2)

4.1 Trolley-problems for AV and the missing factor speed

As I have shown above (see section 1.3), the factor speed is ignored in almost all of the ethical literature on AV, which in turn is highly focussed on so-called trolley-discussions. If speed is explicitly mentioned, it is mostly set as "given" at a sufficiently high level to reliably kill everyone involved in the subsequently discussed crash scenarios. I hold, that this is a serious shortcoming within the ethical debate on AV.

On the one hand the development and aspired further introduction of AV is all about traffic safety (see section 1.3) and on the other hand speed is the single most important factor in terms of severe crashes and road traffic fatalities (see section 3.2) – a relationship between those two topics therefore should be quite intuitive. As speed inevitably is a relevant factor for traffic safety, it necessarily plays a role in every trolley-problem, regardless of whether it is explicitly mentioned or not. For trolley-cases in which drivers as well as pedestrians reliably are killed in every possible alternative (i.e. the baseline scenario in most AV-quanderies), we have to (explicitly or implicitly) assume vehicle speeds of around 100 kmh and more (see section 3.2). Arguably, such high speed levels are rather unusual for the urban context with typical posted speed limits of around 50 or 30 kmh, although this is exactly where many trolley-scenarios (sometimes implicitly) take place – i.e. cases in which we find intersections, crossings, pavements and different pedestrians such as children and elderly. For crashes at around 50 kmh fatality risk for drivers is about 3 % compared to 10 % for pedestrians, i.e. three times as high. Possibly we might conclude, that in such circumstances AV should rather seek to avoid crashes with pedestrians, even if this meant to self-sacrifice the AV and crash against a wall – since its driver and occupants still have significantly higher chances to survive. Abrupt evasive manoeuvres with a high risk of skidding are better avoided in such cases, as side-impacts expose occupants to an even higher fatality risk of about 25 %. At a speed level of around 60 kmh fatality risk for pedestrians and drivers are much closer together at about 20 % and 17 % respectively – and with 85 % absurdly high for side-crashes. If we're into "trolleyology", we should also not forget that age plays an important role. So at a speed level of around 50 kmh, there's good reason to especially spare elderly pedestrians, since their fatality risk at this speed is significantly higher compared to other pedestrians as well as drivers. Strictly following the insights from such a speed-perspective, another problem in terms of "targetting" (see section 1.3.1) arises: should AV be aimed at younger pedestrians, because their fatality risk is significantly lower?

What's more, speed also determines what options for action drivers have. As pointed out above (see section 1.3.4), swerving is not an option at any speed. Especially in the usual trolley-scenarios (i.e. where everyone is reliably killed upon impact), we are implicitly talking about speed levels at which swerving is not a sensible option at all, as this probably leads to skidding, which maintains a higher and therefore more dangerous speed compared to braking, makes cars uncontrollable and unforeseeable for fleeing pedestrians, and risks side-impacts including very high fatality risks for the car's occupants. At lower speeds, when swerving (from a car dynamics-perspective) is a viable option, we're more likely faced with alternatives of the type "material damage versus fatality" – since occupants, different than pedestrians, will most likely survive these crashes. From this we could infer, that at lower speed levels AV as a general rule should be programmed to perform self-sacrificing avoidance manoeuvres, since human health arguably trumps material damage. Following this line of thought brings up the question, how we then could prevent pedestrians (e.g. prankster kids) of intentionally taking advantage of this kind of programming, i.e. blocking and thereby rendering AV rather unattractive. Fines for false alarms could do. However, what is important here is that taking into account speed significantly changes the discussion. The typical options (i.e. swerving or braking) in standard AV trolley-cases are not really options in high speed-scenarios, whereas in low speed-scenarios the typical outcomes (i.e. everyone reliably is killed) do not hold.

As already mentioned above, some scholars point to the relevance of "statistical trolley dilemmas" and "crash-optimization" instead of clear-cut trolley-situations (see section 1.3.1 and 1.3.2). In doing so, Lin (2016) rejects braking as a dominant strategy in potential crash-situations, stating that a wet road or tailgater could make braking more dangerous than swerving. This doesn't convince me as a rejection of braking as a valid strategy, and – more importantly for the purpose of this thesis – again hints at the importance of speed also from a probabilistic perspective.[23] Crash-optimization and statistical trolley dilemmas are about risk management, because in the long run even minor adjustments in the software or "behaviour" of AV will sooner or later lead to harmful outcomes. My argument is, that deliberations in this regard also have to take speed into account. The mentioned tailgating is a good example: the higher the speed, the more safety distance between vehicles is needed. Bonnefon et al. (2019: 503f) discuss the lateral distances of an AV between a cyclists on the one and a heavy truck on the other side in terms of

23 Although beyond of the scope of this thesis, I doubt that swerving is a more viable option on wet roads when braking is not. In case of aquaplaning it is certainly not – swerving is the worst you could do. Slamming the brakes neither, however slowing down (eventually supported by soft braking and an anti-lock brake system) is advised. In other wet road scenarios I suppose that a high skidding-risk also speaks against swerving compared to braking in most of the cases. Tailgating arguably is a human vice. So, if the following vehicle is an AV too, it doesn't hold as an argument against braking. If the following vehicle is manually driven, it might be more cautious when following an AV (knowing that it could abruptly stop with a higher chance compared to human driven cars) but anyway is legally and morally to blame – which is why, in my opinion, tailgating is not a strong argument against braking neither.

statistical safety risks. For the sake of argument, we could rearrange this example using the insights from part three (see section 3.2). Suppose an AV is driving at around 40-50 kmh on a narrow urban road. On the left side it encounters a group of children, on the right side a group of elderly. Knowing that at this speed fatality risk is significantly higher for elderly, we might conclude that the car should rather maintain a greater safety distance to them and therefore tend towards the left side of the road. At a speed of 100 kmh (however unrealistic for narrow urban roads), we could draw a different conclusion and – following the results of the "moral machine" (see section 1.3.1) – try to spare the kids by telling the AV to tend towards the right side of the road.

Whatever you think about these questions or trolley-discussions in general, my point is simple: speed matters and it is a significant shortcoming within the ethical discussion on AV, that it is barely ever mentioned. With the perspective of "functional safety" in mind (see section 1.3.4), one might even dare to ask if road traffic trolley-discussions are not rather for human drivers? Of course will AV be involved in crashes (see section 1.2.2). But the crashes that have happened so far were exclusively caused by different software errors or human errors, instead of any real-life manifestation of a trolley-problem. It is at least questionable if properly functioning AV should get in those kind of trolley-situations usually discussed in the first place. Lin for example describes a trolley-case in which an AV is driving on a narrow rural road. "No one and no technology could foresee that a school bus with 28 children would appear around the corner [...]" (Lin 2016: 76). Well, if you approach a corner that you can't see round – you have to adapt your speed in advance and AV will hopefully better comply with this obvious safety requirement than many human drivers do. Be this as it may, Lin holds that it is no valid objection to trolley-cases to claim that they are unrealistic – since they "are still useful in drawing out certain ethical intuitions and principles we want to test." (Lin 2016: 76) I accept this position but want to add, that if we detach trolley-discussions from reality in order to find and test general ethical principles, the anchoring of these discussions in the field of AV is completely arbitrary and it doesn't help us find practical solutions for real-life problems within this specific field. If we do wish to ground trolley-discussions in AV-reality, the factor speed shouldn't be overlooked so easily – as I have tried to argue above.

4.2 Speed as allocation-tool for shared responsibility in AV-crashes

In this section I want to come back to my initial research interest and discuss the relationship between the factor speed and responsibility for crashes in context of AV. As I have shown above (see section 2.2), virtually none of the discussed scholars is willing to grant AV the capacity of fully bearing responsibility by themselves. So, in search of responsible agents after a crash "the car" is (at least for current and foreseeable AV) not a satisfying answer for most scholars. Many scholars try to overcome "moral individualism" by exploring and describing collaborative, joint, extended, hybrid etc. forms of agency, having in common that besides one or more human agents, also one or more other, non-human or artificial agents are included into the picture – something others describe as responsibility network (see section 2.2.3). Only List (2021) conceptualises a direct form of AI-responsibility, similar to group agency for big organized collectives. Although I consent to much of this argumentation, I doubt that current AV or AV in the near future can be said to have sufficient deliberation capacities necessary to reach the status of "moral agency" List thinks of (see section 2.2.1). As Nyholm (2018c) points out, this would involve agents at eye level with other moral agents – e.g. an AV rejecting the orders of its programmers, based on "its own" deliberations (however those might look like). As long as AV are some form of tools or servants – although widely independent from direct interventions and sometimes unpredictable – and at least indirectly obey to the wishes, orders and rules given by others, they cannot be seen as fully moral agents themselves (see section 2.2.3). Nyholm, in line with Sio/Hoven (2018) and their tracking-/tracing-approach towards "meaningful human control" (see section 2.2.1), suggest that we have to look for responsible humans in the loop – since an AV doesn't act on its own behalf. As already suggested above (see section 2.4), List could combine his accounts of group and artificial agency in this regard, in order to get at a sort of human influenced (or "principled", as Nyholm says) moral artificial agent – e.g. a practically widely independent AV, bound back to human deliberations within something like ethics committees, successively developing and improving the AV's software. This way, an AV possibly could "be responsible" and "learn" from critical situations, similar to big companies being responsible and learning from harmful outcomes of their activities – as List has in mind with his conception of group agency. Anyway, a certain doubt remains as to whether we would accept such conceptual arrangements as adequate loci of moral blame too, in addition to their function of bearers of legal responsibility,. Think of the disastrous example of the

"Deepwater Horizon" oil spill in the Gulf of Mexico in 2010 (List 2021: 1222). It is questionable, if we accept "the company" as one and only blameworthy agent in this context – or if we further wish to identify one or more human representatives to blame in a retributive fashion. A similar objection can be raised against Champagne/Tonkens' (2015) approach of "blank check" responsibility (see section 2.2.1). Although I totally accept this line of argument and concede that, to some degree, it involves not only the legal but also moral side of responsibility, I claim that it's applicability for avoiding "retribution gaps" depends on the context, also information/instructions play an important role (see section 2.4). In discussing legal aspects of this issue (see section 2.2.2), I already pointed to the role of user instructions as an important aspect in the context of product safety – i.e. users have to know about the involved risks, how to properly use certain products and what is expected from them in critical situations. In this regard and for the context of using AV it arguably makes a huge difference for your responsibility, how much you know about and influence you have on the driving (see section 4.1). From today's perspective it is still unclear to which degree and in which type of utilisation AV will enter into our everyday life: as private cars, as robotaxis or as part of public transport. This differentiation runs in parallel with the degree of influence on the driving, as well as constitution of traffic risk (captured with strict liability; see section 2.2.2): a "driver" of a private car/AV is responsible for the driving and the constitution of risk; a "user" of a robo-/taxi doesn't control the driving but is the main reason for the taxi driving around; a "passenger" of a public AV-/bus does neither control the driving nor is she the main cause of the vehicle driving about in a certain region or city (because it would do this even without any passengers).[24] This arguably has implications for the question of responsibility and goes along perfectly with the assessment of many legal scholars, that liability presumably will shift from drivers to manufacturers in the context of AV (see section 2.2.2).

It also matches nicely with my argumentation for the factor speed as playing a decisive role in this regard. In what follows and based on the evidence gathered so far, I present several arguments to support the central claim of this thesis, i.e. speed matters for discussions on responsibility in the context of AV. More specifically I try to point out that the decision on speed can be used as an

24 Obviously one could object, that taxis sometimes roam around idly (i.e. users do not fully constitute this risk) and ever-empty public bus-lines would soon be cancelled (i.e. passengers do at least partly constitute this risk). Seen this way, differentiation becomes harder – but nevertheless, I think this prototypical categorisation helps to examine my line of argumentation.

allocation-tool for responsibility within ethical frameworks referring to forms of shared responsibility or responsibility-networks in the context of road traffic crashes involving AV.

(1) Speed is directly related to traffic safety: As elaborated in detail above (see part 3), speed is the single most important factor for sever and fatal crashes. Traffic safety on the other hand is the most important and generally recognised goal of introducing AV into our traffic systems (see part 2). It is therefore quite obvious that speed should be considered when talking about traffic safety.

(2) Speed becomes even more important in the context of AV: As other potential crash causes, especially human errors such as inattention, distraction, misjudgements, speeding, fatigue or disregarding of traffic laws will (hopefully) lose importance with AV taking over ever more driving activities, the factor speed will even gain in importance and – due to physics – will never go away.

(3) Speed can serve as indicator within traffic responsibility networks: As I have mentioned above (see part 2), many scholars within the discussion around moral and legal responsibility in the context of AI are moving towards different conceptions of shared responsibility or responsibility networks. I generally support this line of thinking. Nevertheless I claim that despite or even because of this, we need to find indicators to adequately allocate responsibility within these networks. One indicator is negligence – e.g. if a car owner or mechanic fails to properly maintain a car and a crash happens as a result, they surely have to bear some of the blame within a responsibility network on traffic safety. The factor speed can serve as another powerful indicator or allocation-tool, since it is clearly connected to traffic safety and some agents (e.g. human drivers or users, software developers etc.) always have to be responsible for choosing a certain speed level.[25] Even if "nobody" took an explicit decision and an AV just strictly follows the local speed limits – we still can think of the lawgiver as relevant agent, since posted speed limits (at least in the context of manually driven cars) definitely influence actual travel speeds and are results of political deliberations (see section 3.1). When talking about traffic safety and crashes, the factor speed therefore helps to single out relevant agents within an alleged traffic responsibility network.

25 An objection to this could be to deny that "some agents always have to be responsible for choosing a certain speed level", in the sense that the actual driving speed itself could be seen as non-decipherable outcome of many different influential factors (e.g. software, speed limits, weather conditions, the car user's stated preferences etc.). Essentially we'd be thrown back to the notion of "responsibility gaps" (see section 2.2.1) – this time for the "responsibility for speed". However, I hold that it is much easier to answer the question "Who is responsible for the travel speed of an AV?" compared to "Who is responsible for the AV in general?". Therefore speed can be seen as practical indicator within the context of responsibility networks and AV (also see Cases B/C in section 4.3).

(4) The focus on speed is useful for a wide range of cases in the context of traffic safety: Both in complex cases with many partial factors and potentially responsible agents, as well as in very simple cases – tracing back the agents responsible for the travel speed will always be a good starting point for answering the question: Who is responsible for this crash? Imagine a crash-scenario with an AV-driver choosing individual settings, a negligent mechanic, a software-bug, bad weather conditions, a blackout in internet coverage and a temporary change in posted speed limits in the area of a big construction site. Although not an easy task, identifying the causes and agents responsible for the AV's pre-impact speed level certainly will help shed light on the responsibility-question. Note that vehicle speed can furthermore appear as a hidden or reinforcing cause for other defects, such as bursting tyres or overheated brake pads due to high speeds. In case of a crash, the agent responsible for the braking system could try to shift the blame to the agent responsible for the improperly high speed. The same holds true for rather simple cases as well as cases including manually driven cars – e.g. a speeding human driver in a perfectly well maintained car under perfect external conditions will be a straightforward agent to blame for any crash.

(5) Speed is an intuitive allocation-tool for crash-responsibility: Above I discussed different proposals for (personal or mandatory) "ethics settings" (see section 1.3.3). To me it seems rather unclear, how such "settings" could be practically operationalised and implemented. How exactly should one imagine a "dial" in terms of an interdependent scale between the dimensions "mobility" and "other interest" when using an AV, as Himmelreich (2022) suggests? How should an interactive test to adapt an AV's settings to one's moral attitude before buying the car, combined with a sort of counselling interview look like in practice, as Arfini et al. (2022) suggest? And how exactly could these preferences and configurations, such as acceleration profile, braking characteristics etc. practically be translated into responsibility in case of a crash? Using the factor speed instead is very intuitive and exactly the same as we are already used to in the context of manually driven cars: how aggressive or cautious, how egoistic or altruistic we drive is to a very large extent related to our decision on how fast we drive in certain situations – as simple as that.

(6) Speed is a practical allocation-tool for crash-responsibility: Besides being intuitive, speed is a very clear and practically identifiable factor when it comes to crash causes and related responsibilities. It can also serve as a useful proxy for other factors more difficult to observe and compare (e.g. vague forms of "more aggressive driving preferences", or "egoistic ethics settings" aiming at reaching one's destination faster). Compared to manually driven cars and the above-mentioned challenges in reliably estimating pre-impact speed levels (see section 3.2), in case of AV potentially equipped with "Event Data Recorders" it will become even more practical in the future. Furthermore it goes without saying that speed is a well-established factor in legal frameworks on traffic for already more than a century – e.g. defining the difference between rightful and wrongful driving or as a minimum criterion for strict liability (see section 2.2.2).

(7) The "decision on speed" covers the legal, as well as moral side of responsibility: As discussed above, boundaries between the legal and moral side of responsibility in the context of AI are fuzzy and some of the proposed solutions to bridge alleged "gaps" fail to cover both sides (see part 2). I claim that using the factor "decision on speed" in order to answer the responsibility-question in the context of AV-crashes can be applied on both sides. An AV-driver able to enforce a very high and risky travel speed arguably will not only bear the Lion's share of liability, but also be an adequate goal of retributive blame for resulting crashes. An AV-owner without any possibility to individually change the car's speed-settings might be liable as well as morally responsible to a very low degree for introducing a risk to society in the first place, but we'd rather be reluctant to put all the moral blame on her – especially if there were other agents and factors (e.g. an unchangeable standard setting or software error) causing irresponsible high vehicle speeds. So, I do not claim that this approach easily gets rid of any complication and grey areas between liability and retribution gaps – but at least it serves as an orientation guide on both sides.

(8) Speed serves as a bridge to politics: As I have shown above, some scholars raise doubts that the strong focus on trolley-discussions in the context of AV distract from more relevant topics, e.g. crash-avoidance (see section 1.3.2). Furthermore we have seen that speed and speed limits are highly political issues (see section 3.1) and that some ethical scholars argue that public debates and politics influence our understanding of responsibility (see section 2.2.4). The short excursion on "jaywalking" is a good example: At the beginning of the 20th century, cars were normally seen and sentenced as dangerous intruders into public spaces. Just a few years later and until today, pedestrians are considered guilty for carelessly walking on the street. In the context of traffic, traffic safety and the development of AV, speed therefore serves as an intuitive bridge to politics and public debates. Also from the perspective of "hermeneutic responsibility" (see section 2.2.4), speed is a good anchor for discussion: Given the further development of AV, what kind of future do we envision? A future in which pedestrians and kids win back a part of their lost public space because AV promise to better comply with traffic laws combined with a drop in crashes? Or a future depicted in many trolley-cases, with AV at uncontrollably high speeds – i.e. a copy of the present with manually driven cars?

4.3 Discussion of case studies

Following the basic idea of this thesis, speed matters for the discussion of traffic safety and AV. In the previous section I submitted some arguments why the factor speed is well suited to discuss and analyse questions of responsibility in this regard. To elaborate these thoughts a little bit further, I now want to discuss different scenarios and the resulting variations in terms of (shared) responsibility. This means that I move away from "moral individualism" trying to identify single individuals fully responsible for a harmful outcome, and instead assume some kind of "responsibility network" with different, relevant and potentially responsible agents in place (see part 2). I do not claim that the focus on speed gives answers to all possible questions in the context of responsibility, AV and traffic safety and I'm aware that the following discussion has to remain rather superficial – nevertheless I hereby want to point out, that speed is an intuitive, practical, relevant and insightful indicator when discussing about responsibility for AV-crashes.

Case A: "driver with full control"

Imagine a scenario in which a "driver" uses his privately owned AV, having extensive control over all possible settings, from acceleration profile, braking characteristics, energy consumption, manoeuvre preferences and – most importantly – travel speed. For the purpose of this argument it is of little importance how exactly the driver executes his orders – i.e. sitting behind the wheel and pressing an accelerator pedal, changing the actual settings on a display or by voice commands from the backseat. What matters is that the driver is the one (mainly) responsible for the actual travel speed of the AV. If – without any other relevant or obvious causes – a fatal crash happens due to the high speed (e.g. 100 kmh in an urban context), I suggest that the driver will be the first candidate for liability (i.e. legal responsibility) as well as blame (i.e. moral responsibility). Now, suppose there was a technical defect – e.g. the driver wanted to slow down, but the AV's command interface for some reason didn't work. As the driver in this case is also the owner of the AV and was the one to expose society to a risk in the first place, he arguably should bear some liability – although in this case he rightfully will try to regress with the manufacturer for the technical defect (who could try to regress with relevant suppliers etc.). Since the driver didn't mean to drive this fast in the latter variation, since the crash was "not his fault" – I suppose he doesn't deserve all the moral blame. If the command interface didn't work due to negligence on the part of a supplier, this agent (or its representatives) within the responsibility network will have to bear most of the (moral and potentially also legal) responsibility. If the manufacturer failed to provide adequate instructions to change travel speed (i.e. the driver could not know how to use the AV properly), it also has made itself guilty. If the driver didn't take the time to learn how to properly use the AV, the picture changes. If (for whatever reason) the AV's software recommended a very high speed level to the driver and he consented to it, despite fearing that it might be too fast for the given surrounding – maybe we'd want to split moral responsibility 50:50, given that the AV's advices had so far been reasonable and adequate. The sharing of liability still could be different, e.g. if the driver explicitly and voluntarily consented in advance, that he'd bear legal liability no matter what the AV suggests. On the other hand suppose a cautious driver commanding his AV to never drive faster than 20 kmh on residential streets. On an uphill section of the road, he meets a cyclist who is racing down the hill at a far too high speed of 60 kmh, dying after crashing into the AV. As we have seen above, at such a low travel speed the AV-driver eventually would even be relieved of his strict liability for introducing a risk to society (see section 2.2.2). It seems clear that we give most or all of the responsibility to the cyclist, who was driving way beyond her safe braking distance. If both were driving at a very low speed of 10 kmh, speed arguably would no longer count as a decisive factor, since involved risks for both vehicles are negligibly low. If both were driving at a rather high speed of 50 kmh, I claim that the AV-driver should have to bear a bigger responsibility-share, since at this speed level his vehicle poses a significantly greater danger to others than is true the other way round for the cyclist.

So, if the driver of an AV controls or mainly influences the travel speed, his or her responsibility in case of severe crashes can be assessed on this basis due to involved risk levels. Needless to say that other factors (e.g. infringement of traffic laws, technical defects, awareness of the critical situation etc.) can also play an important role – but speed nevertheless is relevant in most cases and always a good starting point. Obviously this applies also to cases for manually driven cars. The potential risk, as well as moral and legal responsibility increase along with the travel speed.

Case B: “user with partial control”

Imagine a scenario in which the “user” of a robotaxi is not a driver any more, but still able to impose some of her decisions – e.g. travel destination and the selection of one of various possible routes. Suppose this user isn’t able to directly influence or order a certain travel speed, but she is given different options in terms of estimated travel time for specific routes – e.g. 20 minutes at an average driving behaviour, 25 minutes at a more cautious configuration and 15 minutes at a slightly more aggressive performance, but still obeying to traffic laws. This way speed gets into play indirectly. Nevertheless it seems fair to assume, I claim, that liability (potentially realised in form of higher travel fees) as well as moral responsibility increase by choosing faster options. However, the user’s responsibility is balanced by the responsibility of the robotaxi company, with the duty of providing safe journeys for all given options. Now, suppose the user must rush to hospital due to a life-or-death emergency and the robotaxi offers an “at your own risk”-option, disregarding stop signs and other barriers and exceeding posted speed limits by far. Obviously in this case the user – voluntarily – puts legal responsibility on her shoulders. I think it also seems fair to assume the same for moral responsibility. Note that this variation is very similar to the driver-case (see case A), because the user imposes an individual decision on the AV. Obviously we can still add more aspects, e.g. if she can credibly demonstrate that she did and could not know how fast the robotaxi really would go (e.g. she expected about 100 kmh, but it went absurd 150 kmh), maybe we’d take away some of the blame of her and shift it to some other agent – e.g. to the manufacturer due to insufficient user instructions or to software suppliers for allowing the AV to go irresponsibly fast. If the user refrains to make any decision related to travel speed (e.g. she simply specifies her destination of travel and leaves everything else up to the AV), the scenario is similar to the passenger-case (see case C).

In cases of “partial control”, for example when using robotaxis, it can be challenging to identify which agent actually decides how fast to drive. Suppose the user of a robotaxi selects one of the offered routes including estimated travel times, the AV flexibly adapts its travel speed to the overall traffic situation, the lawgiver (for some unknown reason) sets a posted minimum speed and bad weather conditions (e.g. fog) deteriorate the performance of the AV’s sensors, forcing it to slow down a bit. Who made the decision on speed? As already stated above (see section 4.2, fn 3), such complex situations might resemble those insoluble riddles we were confronted with in the context of responsibility and AI, were some scholars feared "responsibility gaps" (see section 2.2.1). I hold that in case of the decision on speed, the riddle is easier to solve – for example by relying on "Event Data Recorders". Contrary to questions concerning somewhat abstract concepts like responsibility, it is way more practical and technically feasible to trace back how a certain travel speed of an AV came about without any "gaps". And if this is not possible for current AV, perhaps we should demand it from manufacturers for the next generation of AV – referring to our "narrative responsibility" (as discussed in section 2.2.4). However, I hold that this challenge and complexity in terms of reconstructing how pre-crash travel speeds came about is not a problem for my argument, but actually a strength. Since this riddle is based on concrete, physical and digital facts, it is always soluble. If we don't succeed to figure out the agents responsible for the travel speed of an AV – we'll have an even much harder time to settle disputes on abstract concepts like responsibility. Therefore, the decision on speed serves as a starting point and orientation guide, especially for tricky cases. In fact the decision on speed actually makes a big part of the difference between driver-case (see Case A) and passenger-case (see Case C) – and this obviously makes a big difference in terms of responsibility: as a driver I have a lot more responsibility than as a passenger.

Case C: "passenger with no control"

Imagine a scenario in which a "passenger" waits at a bus stop and after a few minutes a public AV-bus with several other passengers turns around the corner. In contrast to the cases discussed above, the passenger here has no strict liability to bear – i.e. he didn't introduce the risk to society, since the public bus is driving around anyway. We can think of other passenger-scenarios including strict liability as well – e.g. a privately owned AV where the "passengers" only specify their travel destination and leave everything else up to the AV. In detail it might get more complicated. In case of a robotaxi, strict liability arguably lies with the taxi-company (also think of the "ubi commoda ibi incommoda"-principle from section 2.2.1). If we think of an exclusive customer, aimlessly driving around all day for "frivolous reasons" (see section 3.1) – maybe we would want this passenger to bear at least a little bit of moral responsibility too, if (for whatever reason) a tragic crash happens. Although the taxi-company was strictly liable and the passenger did not intervene in the driving performance at all, she was still the reason causing a heavy and potentially dangerous vehicle to roam around all day for no good reason and therefore exposing others to a certain degree of risk. Put differently, ownership as well as exclusivity are relevant factors to discuss strict liability and moral responsibility for generally exposing others to a hazard (irrespective of any specific fault), whereas – following my line of argument – the decision on speed makes the major difference between the morally relevant categories of drivers, users and passengers.

Especially in the case of a "passenger with no control" the question arises: Who actually decides on the travel speed of an AV? In the previous "user-case" (see case B) I already touched upon this question, focussing on a complex mixture of different influential factors. Now I want to address the simple version of this question and ask, which agent in the passenger-scenario mainly determines the travel speed level under normal conditions? If we approach this question with manually driven cars in mind, we might point to the manufacturer (as the agent with overall responsibility for everything related to an AV's performance). Just as human drivers are accountable for the travel speed of their car, manufacturers are accountable for the travel speed of their AV. What about the influence of posted speed limits? As we have seen above, posted speed limits have a significant influence on average travel speeds (see section 3.2). Against this background, it seems fair to ask whether the decision on speed in case of human drivers is not actually more influenced by the lawgiver than the human drivers. Although traffic law usually obliges drivers not to endanger anyone and to adapt their speed accordingly, it also forbids to drive so slowly to hinder traffic without a compelling reason (RIS 2024b: §20). Seen from this perspective the question arises, how much responsibility the lawgiver should have to bear for fatal crashes? Think of a case where a human driver, followed by many other cars (i.e. she is forbidden to drive slowly and hinder traffic), sticks to an arguably way too high posted speed limit of 70 kmh in the surrounding of a school and a little kid darts out on the street… I think it is fair to argue that in this case the lawgiver has to bear a significant share of (at least moral) responsibility for severe crashes. If we go back to the passenger-case the question arises, how significant posted speed limits are for AV? On the one hand we could argue that they become even more relevant because AV, unlike human drivers, adhere to them perfectly. On the other hand and based on the perspective of "functional safety" (see section 1.3.4) we could argue that AV don't really need speed limits any more, since they should be programmed to tactically adapt their travel speed to their operational capabilities – or put differently: AV are never tempted to speed anyway and posted speed limits would only be an unnecessary relic from the old days.

So, even if the user of an AV is merely a passenger without any influence on speed and driving performance, the factor speed still works as an orientation guide to think about other potentially relevant agents – which in turn helps to discuss questions of responsibility in case of harmful outcomes.

5. Conclusion

The dream of self-driving, autonomous vehicles (AV) has been dreamt for decades, today it has become true. From vague futurism, to deceptively real-looking science-fiction, to first pilot projects and prototypes, to commercially used reality in several regions of the world: AV are here to stay. Even if it is not yet completely foreseeable which type of utilisation will prevail – private cars, robotaxis or means of public transport – all types of AV share common hopes for the future: easy, comfortable and efficient mobility accessible for all, economic opportunities, and most importantly a drastic improvement in traffic safety. So far, the ethical discussion that has evolved over the last decade has centred primally on so called trolley-dilemmas of the kind: *imagine your AV encounters this terrible choice in which it is forced to either swerve left and kill an old women or swerve right and kill an innocent boy…* Interestingly, this discussion usually takes little notice of the practice. Taking the perspective of vehicle dynamics reveals that swerving would be an absurdly risky option in many of the assumed situations. The perspective of product development points to international standardisation for hazardous products, indicating that manufacturers would get into big trouble if they sold such life-threatening AV as assumed in trolley-cases due to liability obligations. Representatives of the automotive industry consider trolley-discussions to be a distraction from real-world problems. And any code of ethics for AV forbids discrimination based on personal characteristics as usually assumed in trolley-problems – information, moreover, which cannot reasonably be assumed for an AV to be accessible and if it is, it arguably better shouldn't be from a data privacy perspective. Even though many interesting aspects were undoubtedly brought up in this discussion, the question of sensibility and practical relevance of the focus on trolley-dilemmas remains – and is also discussed within the ethical community itself. One common response from the proponents is to detach the discussion from any claims of practicality: trolley-discussions, so the argument goes, don't have to be realistic and nevertheless bring about meaningful insights into ethical challenges. If this is the case, I claim, the anchoring in the context of AV is rather arbitrary and the insights gained have more to do with general ethical principles than with traffic safety and AV – as arguably is the case with the main findings of the empirically unrivalled trolley-project "moral machine": human victims trump animals, less victims trump more victims and young victims trump old ones. If, on the other hand, trolley-discussions are supposed to be related to AV in real-life traffic, practically relevant factors should be taken into account. In this thesis I have mainly focussed on one such practical factor, which has largely been ignored in the ethical discussion so far: speed. I claim that speed implicitly already is present in any trolley-scenario whether desired or not and that options for action as well as the outcomes of potential AV-crashes are significantly influenced by the involved speed levels – e.g. at urban speed levels (at which most trolley-scenarios take place) AV should tend to self-sacrificing manoeuvres, since their occupants have reasonably high chances of surviving these crashes anyway.

Turning to the question of responsibility for traffic safety at heart of this thesis it became clear, that traditional conceptions of responsibility reach their limits in the context of AI. In case of a crash caused by a manually driven car, things seem pretty obvious: a conscious and capable driver in control of his vehicle is the natural responsible subject. On closer inspection we saw that public opinion and the political framework conditions also play important roles for the question of responsibility, since implicit as well as explicit norms define its boundaries. Before the massive proliferation of cars, pre-motor age speeds stemming from horse-drawn carriages were the norm on public streets and speeding cars were seen as unwelcome intruders – morally and legally responsible for any harmful encounter. Some decades and a jaywalker-campaign later, things have turned upside down and pedestrians now are held responsible for "causing" crashes and fatalities. Political deliberation brought about legal speed limits and adaptations in the course of time – rather according to political logic than traffic safety insights. However, speed limits significantly influence actual travel speeds which raises the question of whether the lawgiver should also bear some of the responsibility for crashes that we usually put on the shoulders of drivers?

Be this as it may, the emergence of evermore powerful, independent and adaptable AI has evoked a broad and profound discussion on the basics of our conception of responsibility. How can someone be responsible for artificial agents which are the products of hundreds of hands, performing self-determined actions under no direct control and even involved experts cannot predict or explain their behaviour? Soon, alleged "responsibility gaps" were feared and thoroughly discussed from a moral as well as legal perspective. One fairly common ground in the ethical discussion is that artificial agents themselves cannot and should not be allowed to fully bear moral responsibility – at least for the present and foreseeable future. This notion is reinforced by the argument and current reality that artificial agents are always designed for humans and fulfil tasks on behalf of humans, who at least indirectly command or influence their behaviour. Therefore artificial agents cannot be seen as on par with their principals. In terms of responsibility this means: we always need to look for humans in the loop. Many proposals have been made to cope with this challenge in the legal realm, from voluntarily and in advance assuming liability, to creating specific legal entities for artificial agents. Relating to the legal practice in the transport sector, adaptations of well-established concepts such as strict liability and product liability have been suggested as viable solutions for liability challenges posed by the foreseeable introduction of more and more AV into the traffic system. With regard to the moral side of the discussion, many different concepts of extended or collaborative agency, combined with joint or shared responsibility as well as responsibility networks have been proposed – with the common feature that responsibility is distributed to more than one agent, whereas "moral individualism" has been overcome.

Against this background, the main focus and research interest of this thesis was to bring together the discussion on responsibility in the context of AV with the factor speed. Based on a detailed discussion of historical, theoretical and empirical aspects I claim that speed is a powerful and inevitably relevant factor for discussing responsibility in context of AV for several reasons. Speed is the single most important factor related to severe and fatal crashes. Since other, human-related risk factors such as inattention, distraction, misjudgement or overconfidence become less important in the advent of AV and the factor speed cannot lose its relevance for physical reasons alone, it therefore will become even more important compared to manually driven cars. Since the allocation of responsibility will foreseeably become more and more challenging due to the many, interwoven contributions of different agents (e.g. driver/user, manufacturer, software developer, lawgiver etc.) to certain harmful outcomes (e.g. a crash), the factor speed can serve as a practical, well-established, versatile and intuitive orientation guide to discuss questions of responsibility in the context of traffic safety and AV. Furthermore it is applicable to the legal (i.e. who has to pay?) as well as moral (i.e. who is to blame?) side of responsibility and, last but not least, emphasises the political dimension of the topic by building a bridge to historic debates on the question: which speed is normal? In some case studies I tried to develop my argument further. I do not claim that the factor speed completely solves the distribution of responsibility in case of any AV-crash. I simply hold that it is an inevitably useful factor to start with.

Policy recommendations

Most of the discussions around AV are highly abstract and theoretical. In this section I want to share some thoughts on how to practically implement some of the findings of this thesis.

First of all I want to argue, that we should talk more about speed in traffic and its societal consequences in general, as well as in the context of AV in particular. As the excursion on "Jaywalking" (see section 3.1) revealed, in the course of the shifting normality from horse to car speed levels, there was a lot of public discussion about speed in traffic at the beginning of the 20th century. Since then, speed and speeding cars have become pretty normal to us – a normality that still costs millions of human lives every year and has been transferred very uncritically to the recently emerged discussion on AV: AV travelling at high or even deadly speed levels are commonly set as given starting points for any further ethical considerations. We could and arguably should use this turnaround in transportation – from manually driven cars to AV – to resume the discussion on speed in traffic and its societal consequences, since it is closely related to socio-politically relevant topics such as physical integrity and personal freedom of drivers as well as bystanders, social participation, the distribution of public spaces and others more. In the last century, we seem to have dedicated more and more public roads to car traffic and allowed human drivers to drive beyond their capabilities, causing millions of accidents based on human errors. Presumably, one reason for accepting speeding despite its ridiculously high death-toll is that human car drivers are the perfect moral and legal culprits. Whereas we as a society accepted human drivers intentionally violating traffic laws, it is not clear if we should grant AV the same right. This obviously makes a huge difference. Say everyone knew that AV are obliged to drive safely and carefully, ready to hit the brakes anytime – pedestrians as well as kids could partly reclaim the streets by forcing AV to slow down. In this scenario we would have to discuss policies that prevent wilful obstruction of traffic. If we otherwise allow AV to run at posted speed limits no matter how risky and unclear the traffic situation might be (as many human drivers routinely do) – pedestrians as well as kids probably wouldn't dare to set foot on any driving lane (as they don't do now). Put differently one could ask: which speed is considered the norm – the speed cars are designed to go or the speed at which casualties are a negligible risk? Or more bluntly: are streets for cars only (like in the past century) or for everyone (as before)? Regardless of where you stand on this issue, my point is that we should remind ourselves of this important question and – in the wake of AV gaining more and more importance – have an open discussion, instead of setting high speed levels in traffic as given. This discussion also ties in with the context of advertising, film and media. As we have learned above, the representation of speeding cars as uncritical, exciting and cool has a major influence on the real-life driving behaviour of many spectators (see section 3.1). Similar to regulations for cigarette advertising and the depiction of alcoholism in films, we could and possibly should discuss rules so that reckless speeding is no longer seen as heroic in our mass media.

As a second takeaway I want to emphasise that we as a society should make sure to prevent responsibility and liability gaps to arise when introducing AV into traffic. To begin with, I would like to recall that we should be very mindful of how and why we are having this discussion in the first place – since talking about alleged responsibility gaps might already be a loaded expression, put into circulation with strategic ulterior motives (see section 2.2.4). In order to close and avoid responsibility gaps when using AV from a moral perspective, my proposal is to use the factor speed as a responsibility distribution tool together with working out a set of ethical standards in this context (see below). As to the legal side of it (i.e. liability gaps), we as a society need to elaborate an adequate insurance regime – i.e. finding a fair balance between strict liability of the vehicle owners as well as users and product liability of the AV manufacturers (see section 2.2.2). If manufacturers had to assume all the liability, it would possibly prohibit the proliferation of AV altogether, due to unbearable high costs. On the other hand it is fair to assume, that AV manufacturers have strong economic interests in shifting as much liability as possible onto the users – which they potentially would have to assume, even if they had no fair chance of understanding and controlling their AV. As I have argued above (see section 4.2), speed could serve as a good starting point in order to settle disputes on liability in case of AV-accidents. What we have also seen above (see section 2.2.2) is that expert testimony and "Event Data Recorders" presumably will play important roles in allocating liability-shares when it comes to AV-crashes. Additionally to elaborating a functional and waterproof liability insurance scheme before introducing AV into everyday traffic, we also could further develop the idea of creating a specific legal entity for artificial agents (and their developers and owners) similar to corporations (see section 2.2.1). One way to get there, as I have argued above, is to understand artificial agents as part of some specific form of group agency and define minimum criteria (e.g. an ethics department to steadily improve an AI's ethical behaviour) to qualify for such a legal personhood.

Now I want to come back to the moral side of alleged responsibility (and retribution) gaps. I agree with all the authors quoted above that AV cannot currently assume moral responsibility by themselves, as they are always tools for the objectives of other agents (see section 2.2.3). As this thesis tries to show, speed plays an important role when discussing responsibility issues in the context of AV and can be used as a tool to better distribute responsibility within a responsibility network of AV. I suggest that regulatory bodies should elaborate ethical standards for coping with the challenge of fairly distributing responsibility when using AV in traffic – using the factor speed as powerful organizing principle. To begin with, responsibility in general should closely be related to the agent who decides how fast the AV drives: the more risk this agent takes by (directly or indirectly) commanding the AV to go faster, the more responsibility she has to assume. This decision-making power over speed also separates the different use-cases of AV mentioned above: if you (mainly) impose travel speed on your AV – you're a driver; if you only jump into a driverless shuttle-service – you're a passenger (see section 4.3). This obviously has implications for questions of responsibility: a driver has more responsibility to bear than a mere passenger. Besides this, also the factors ownership (who owns the AV?), exclusivity (who can use it?) and profitability (who benefits from it?) should play a role in distributing responsibility in case of accidents with AV. Furthermore such ethical standards could build upon the idea of "ethics settings" (see section 1.3.3). In terms of "personal ethics settings", the (direct or indirect) configuration of travel speed could serve as a rather intuitive proxy and could therefore be demanded as an explicit duty for any individual AV user: the more speed and aggressive driving behaviour you allow, the more responsibility you have to bear. Transparency and adequate information as well as practical introduction is very important here – if you are ready or forced to deciding on average travel speed levels of an AV, you should know exactly what you're getting yourself into and the relevant information shouldn't be hidden in endless terms and conditions. Seen from the other side: it shouldn't be allowed that you are asked to chose one of several route and travel time options – and thereby unwittingly all the responsibility is transferred to you. Using speed as a proxy for your personal ethics settings, by the way, is pretty much following the current path with manually driven cars: speeding drivers tend to take greater risks and value their individual freedom higher than conservative slow drivers. In this vein such ethical standards on responsibility when using AV should also adress the role of driving licenses. One practical solution could be that you'd only be allowed to set faster and therefore riskier configurations for any AV if you have a driving licence (however the training for this will look like in the context of an AV-dominated traffic system).

When talking about "mandatory ethics settings" the question arises, what role the lawgiver plays in the context of using AV. As we have seen above, posted speed limits are results of political debates and starkly influence average travel speed levels (see section 3.2). When it comes to AV, two conflicting aspects emerge: on the one hand, AV (hopefully) obey strictly to posted speed limits – which strengthens their importance; on the other hand, following the perspective of functional safety (see section 1.3.4), posted speed limits become rather superfluous if AV never drive faster than they safely can manage; In the context of mixed traffic with normal and self-driving cars it is rather straightforward that posted speed limits remain highly relevant, since big differences in travelling speed increase the risk of accidents (see section 3.2). Nevertheless I think we should use this opportunity of emerging AV to start a broad, fact-based and open discussion on how much speed we want to allow in specific contexts – as already mentioned above. I suggest that we'll need some form of compromise between personal (i.e. individual decision-making) and mandatory (i.e. posted speed limits) ethics settings – just as we have now with manually driven cars. As there are other socially relevant goals connected to car traffic besides traffic safety (e.g. noise, air pollution, energy consumption, spatial segmentation of neighbourhoods), I hold that posted speed limits should still play a role even in an "AV only" future. When thinking of such a future, I suggest that dynamic posted speed levels become worth thinking about – e.g. lower speed levels when more people are out on the streets and in case of risky weather conditions and vice versa.

What if we agreed that speed has to play an important role in distributing responsibility in cases of AV accidents – but due to complexity and the problem of many hands (see section 2.2) we fail to identify the decisive agents responsible for the actual travel speed of an AV? As I have argued above (see section 4.2), this riddle is easier to solve than the one about responsibility in general and "Event Data Recorders" as well as expert testimony will hopefully help us. To make sure that we can use speed as a helpful decision-tool in this context, we also could proactively demand that manufacturers design AV in such a way that it is transparent and clear, how actual travel speed levels came about and who is in charge of deciding it – right from the start, but also in retrospect. In the driver case, the AV could simply ask the responsible driver before starting, if she prefers to use her own speed and driving settings, potentially also including the duty to take over control in risky situations, with the benefit of arriving faster in exchange for bearing a larger part of (moral and legal) responsibility in case of an accident – or if she wants to stick with the standard settings defined by the manufacturer (or software company or lawgiver), with the benefit of bearing less responsibility in exchange for driving slower. However this is achieved, the point is to design and use AV in a way that helps us to identify the responsible agents for the actual travel speed of an AV – and to use this information to inform disputes about responsibility for AV-accidents in general.

Last but not least and as I have said before, talking about AV also means constructing expectations, ideas, dreams and hopes and thus ultimately the future reality itself – and we should be aware of this "narrative responsibility" (see section 2.2.4). If we focus on the discussion of trolley-dilemmas, we skip most of the transformative potential of AV and assume that these cars should be allowed to run as fast and risky as manually driven cars. Even if we prescribe the functional safety perspective (see section 1.3.4) and allow AV to only drive as fast as they can safely manage – this could backfire if we imagine a strong car lobby demanding access bans and fencing for roads in order to enable faster AV-traffic. On the other hand, if we break free of our car-centric thinking of the past century and consider the (human) vice of speeding as obsolete thanks to the introduction of AV – we could imagine AV as safe mobility service tools, opening up unimagined possibilities in terms of boosting traffic efficiency as a whole while reducing the number of vehicles (via ride pooling and sharing), improving social inclusion (because everyone can use them) and giving back some of the lost freedom and access to public spaces to our kids. However the story about AV is told – it will influence further developments and contribute to the upcoming normality in traffic.

Outlook

As could be expected, this thesis left open much more relevant questions than it could address, let alone answer – which seems pretty natural, since we are only at the beginning of this new chapter in the history of transportation. In regard to the ethical discussion on AV, I suggest to turn away the focus from abstract trolley-problems and to better ground our reasoning on practically relevant facts, insights from traffic safety research as well as ethical principles for product development.

The discussion on responsibility in the context of AI doubtlessly needs further research as well. As I have shown in this thesis, boundaries between the legal and moral side of responsibility are often blurred. Therefore the discussion on responsibility in this context needs further clarification. Whereas questions on liability when using AI, robots or AV are already well advanced, concepts to settle questions of moral as well as other (such as social or hermeneutic) forms of responsibility need to be developed further. Probably the most often discussed question in this context is: Can robots ever be full-blown moral agents, worthy of fully bearing moral responsibility? Of course this is an epochal question. Nevertheless it has little to do with today's and tomorrow's practical challenges, where robots are certainly not on par with their human counterparts and forced to do what humans want from them – even if humans give up direct control and cannot predict their actions. In order to avoid confusion and delusion in the context of AI-ethics we should sharpen our focus on the most relevant questions at present and leave the unsolvable ones for later.

Regarding the discussion on AV, speed and traffic safety, questions on different levels arise. On a rather practical level one might ask, which role the possession of a driving licence will play in an AV-future, as already briefly mentioned above. The whole idea of learning how to "drive" is obviously at odds with the promise of AV. Maybe this opens up room for other forms of traffic education, e.g. giving a broad overview of all possible means of transportation, how to properly use them and which legal and moral duties come along with them.

On a rather abstract level we furthermore could discuss what such developments tell us about ourselves, about our reasoning and our "making sense" of the world. How, for example, does the perception of freedom change in the face of a transformation from manually driven cars to AV? Since their invention, cars were associated with the promise of freedom – i.e. going wherever, whenever and how fast you want. The driving experience itself, a rush of speed and adrenaline with wind in your hair, was hailed as sense of freedom in thousands of adverts and movies. AV also promise a kind of freedom, but a very different one: you are no longer responsible, you don't have to pay attention while driving, you don't have to figure out the best route, you don't have to look for a parking space – maybe you don't even have to worry about owning a car at all, if AV-enhanced mobility turns into an encompassing public service.

Last but not least, the introduction of AV should also be examined more detailed from the perspective of traffic and urban planning. As I have mentioned above, it is not yet clear which form of AV-usage will proliferate – exlusively used private AV or publicy used shared AV (or any other form in between)? Instead of leaving this question to the free market, it should be answered politically in advance – with a view to the question of how the quality of life of society as a whole is best served. To give an example: although shared AV might have the potential to reduce the number of vehicles necessary to guarantee a certain level of mobility – the opposite could also happen if empty AV drive around idly to avoid parking fees. As traffic makes up a large part of urban planning, the rise of AV essentially means nothing less than asking the question: What kind of city do we want for the future?

In this thesis I have tried to shed light on one important factor usually omitted from the discussion on AV, responsibility and traffic safety: speed. Nevertheless, many more has to be done to finally reach what AV promise: stop killing millions of people in traffic every single year.

Michael Schwendinger

6. Bibliography

6.1 Academic articles and books

Aarts, L., & van Schagen, I. (2006). Driving speed and the risk of road crashes: A review. *Accident Analysis and Prevention*, *38*, 215–224.

Allen, C., & Wallach, W. (2012). Moral Machines: Contradiction in Terms or Abdication of Human Responsibility? In L. Patrick, A. Keith, & B. George A. (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT Press.

Andrade, J. A. (2021). The ethics of the ethics of autonomous vehicles: Levinas and naked streets. *South African Journal of Philosophy*, *40*(2), 124–136.

Arfini, S., Spinelli, D., & Chiffi, D. (2022). Ethics of Self-driving Cars: A Naturalistic Approach. *Minds and Machines*, *32*(4), 717–734.

Asaro, P. (2007). Robots and Responsibility from a Legal Perspective. *Proceedings of the IEEE Conference on Robotics and Automation, Workshop on Roboethics, Rome, April 14, 2007*. https://peterasaro.org/

Asaro, P. (2016). The Liability Problem for Autonomous Artificial Agents. *AAAI Symposium on Ethical and Moral Considerations in Non-Human Agents, Stanford University, March 21-23, 2016*. https://peterasaro.org/

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.

Berendt, B. (2020). (De)constructing ethics for autonomous cars: A case study of Ethics Pen-Testing towards "AI for the Common Good". *International Review of Information Ethics*, *28*(6), 1–21.

Bhargava, V., & Tae, W. K. (2017). Autonomous Vehicles and Moral Uncertainty. In P. Lin, R. Jenkins, & K. Abney (Eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (pp. 5–19). Oxford University Press.

Birck, L. (2024). Freie Fahrt für freie Volksgenossen. *Merkur*, *78*(896), 28–40.

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.

Bonnefon, J., Shariff, A. F., & Rahwan, I. (2015). Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars? *ArXiv*, *abs/1510.03346*. https://api.semanticscholar.org/CorpusID:16519064

Borenstein, J., Herkert, J. R., & Miller, K. W. (2019). Self-Driving Cars and Engineering Ethics: The Need for a System Level Analysis. *Science and Engineering Ethics*, *25*, 383–398.

Champagne, M., & Tonkens, R. (2015). Bridging the Responsibility Gap in Automated Warfare. *Philosophy & Technology*, *28*(1), 125–137.

Coeckelbergh, M. (2016). Responsibility and the Moral Phenomenology of Using Self-Driving Cars. *Applied Artificial Intelligence*, *30*(8), 748–757.

Coeckelbergh, M. (2020). *AI ethics*. MIT Press.

Coeckelbergh, M. (2023). Narrative responsibility and artificial intelligence. *AI & Society*, *38*, 2437–2450. https://doi.org/10.1007/s00146-021-01375-x

Coeckelbergh, M., & Loh, J. (2020). Transformations of responsibility in the age of automation: being answerable to human and non-human others. In B. Beck & M. Kühler (Eds.), *Technology, Anthropology, and Dimensions of Responsibility* (Vol. 1, pp. 7–22). J. B. Metzler'sche Verlagsbuchhandlung & Carl Ernst Poeschel GmbH.

Combs, T. S., Sandt, L. S., Clamann, M. P., & McDonald, N. C. (2019). Automated vehicles and pedestrian safety: exploring the premise and limits of pedestrian detection. *American Journal of Preventive Medicine*, *56*(1), 1–7.

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, *18*(4), 299–309.

Davnall, R. (2020). Solving the Single-Vehicle Self-Driving Car Trolley Problem Using Risk Theory and Vehicle Dynamics. *Science and Engineering Ethics*, *26*, 431–449.

De Freitas, J., Anthony, S. E., & Censi, A. (2020). Doubting Driverless Dilemmas. *Perspectives on Psychological Science*, *15*(5), 1284–1288.

de Jong, R. (2020). The Retribution-Gap and Responsibility-Loci Related to Robots and Automated Technologies: A Reply to Nyholm. *Science and Engineering Ethics*, *26*(2), 727–735.

Dubljevic, V., List, G., Milojevich, J., Ajmeri, N., Bauer, W. A., Singh, M. P., Bardaka, E., Birkland, T. A., Edwards, C. H. W., Mayer, R. C., Muntean, I., Powers, T. M., Rakha, H. A., Ricks, V. A., & Samandar, M. S. (2021). Toward a rational and ethical sociotechnical system of autonomous vehicles: A novel application of multi-criteria decision analysis. *PloS One*, *16*(8), e0256224–e0256224.

Elvik, R. (2005). Speed and Road Safety. Synthesis of Evidence from Evaluation Studies. *Transportation Research Record*, *1908*, 59–69.

Elvik, R. (2013). A re-parameterisation of the Power Model of the relationship between the speed of traffic and the number of accidents and accident victims. *Accident Analysis and Prevention*, *50*, 854–860.

Elvik, R., Vadeby, A., Hels, T., & van Schagen, I. (2019). Updated estimates of the relationship between speed and road safety at the aggregate and individual levels. *Accident Analysis and Prevention*, *123*, 114–122.

Etienne, H. (2021). The dark side of the "Moral Machine" and the fallacy of computational ethical decision-making for autonomous vehicles. *Law, Innovation and Technology*, *13*(1), 85–107.

Evans, L. (2008). Death in Traffic: Why Are the Ethical Issues Ignored? *Studies in Ethics, Law, and Technology*, *2*(1), 1–11.

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, *14*, 349–379.

Foot, P. (1978). The Problem of Abortion and the Doctrine of the Double Effect. *Virtues and Vices and other essays in moral philosophy,* 19.

Fossa, F. (2022). Operationalizing the Ethics of Connected and Automated Vehicles: An Engineering Perspective. *International Journal of Technoethics*, *13*(1), 1–20.

Franssen, M. (2015). Design for values and operator roles in sociotechnical systems. In J. van den Hoven, P. E. Vermaas, & I. van den Poel (Eds.), *Handbook of ethics, values, and technological design. Sources, theory, values and application domains* (pp. 117–149). Springer Reference.

Gogoll, J., & Müller, J. F. (2017). Autonomous Cars: In Favor of a Mandatory Ethics Setting. *Science and Engineering Ethics*, *23*, 681–700.

Goodall, N. J. (2016). Away from Trolley Problems and Toward Risk Management. *Applied Artificial Intelligence,* 30(8), 810–821.

Gurney, J. K. (2013). Sue my car not me: products liability and accidents involving autonomous vehicles. *Journal of Law, Technology & Policy*, *2013*(2), 247–277.

Hanson, A. F. (2009). Beyond the skin bag: on the moral responsibility of extended agencies. *Ethics and Information Technology*, *11*(1), 91–99.

Heidbrink, L. (2017). Definitionen und Voraussetzungen der Verantwortung. In L. Heidbrink, C. Langbehn, & J. Loh (Eds.), *Handbuch Verantwortung* (pp. 3–34). Springer VS.

Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis. *Science and Engineering Ethics*, *21*, 619–630.

Hey, T. (2019). *Die außervertragliche Haftung des Herstellers autonomer Fahrzeuge bei Unfällen im Straßenverkehr*. Springer Fachmedien Wiesbaden Imprint: Springer Gabler.

Himmelreich, J. (2018). Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations. *Ethic Theory and Moral Practice*, *21*(3), 669–684.

Himmelreich, J. (2022). No wheel but a dial: why and how passengers in self-driving cars should decide how their car drives. *Ethics and Information Technology*, *24*(45), 1–12.

Husak, D. (2004). Vehicles and Crashes: Why is this Moral Issue Overlooked? *Social Theory and Practice*, *30*(3), 351–370.

Hydén, C. (2020). Speed in a high-speed society. *International Journal of Injury Control and Safety Promotion*, *27*(1), 44–50.

Inagaki, T. (2010). Traffic systems as joint cognitive systems: issues to be solved for realizing human-technology coagency. *Cognition, Technology & Work*, *12*(2), 153–162.

International Transport Forum [ITF] (2018). Speed and Crash Risk. Research Report. Paris.

JafariNaimi, N. (2018). Our Bodies in the Trolley's Path, or Why Self-driving Cars Must Not Be Programmed to Kill. *Science, Technology, & Human Values*, *43*(2), 302–323.

Johansson, R. (2015). The Importance of Active Choices in Hazard Analysis and Risk Assessment. Workshop CARS 2015 – Critical Automotive applications: Robustness & Safety, Septembre 2015, Paris, France. *HAL open science,* hal-01193028.
https://hal.science/hal-01193028

Johansson, R., & Nilsson, J. (2016). Disarming the Trolley Problem. Why Self-driving Cars do not Need to Choose Whom to Kill. Workshop CARS 2016 – Critical Automotive applications: Robustness & Safety, Septembre 2016, Göteborg, Sweden. *HAL open science,* hal-01375606.
https://hal.science/hal-01375606

Johnson, D. G. (2015). Technology with No Human Responsibility? *Journal of Business Ethics*, *127*(4), 707–715.

Kauppinen, A. (2021). Who Should Bear the Risk When Self-Driving Vehicles Crash? *Journal of Applied Philosophy*, *38*(4), 630–645.

Kröyer, H. R. G., Jonsson, T., & Várhelyi, A. (2014). Relative fatality risk curve to describe the effect of change in the impact speed on fatality risk of pedestrians struck by a motor vehicle. *Accident Analysis and Prevention*, *62*, 143–152.

Lawlor, R. (2022). The Ethics of Automated Vehicles: Why Self-driving Cars Should not Swerve in Dilemma Cases. *Res Publica*, *28*, 193–216.

Lin, P. (2016). Why Ethics Matters for Autonomous Cars. In: M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous Driving. Technical, Legal and Social Aspects* (pp. 69–85). Springer.

List, C. (2021). Group Agency and Artificial Intelligence. *Philosophy & Technology*, *34*(4), 1213–1242.

Loh, J. (2019a). Responsibility and Robot Ethics: A Critical Overview. *Philosophies*, *4*(58), 1–20.

Loh, J. (2019b). *Roboterethik: Eine Einführung*. Suhrkamp.

Loh, W., & Loh, J. (2017). Autonomy and Responsibility in Hybrid Systems: The Example of Autonomous Cars. In P. Lin, R. Jenkins, & K. Abney (Eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (pp. 35–50). Oxford University Press.

Lohmann, M. F. (2016). Liability Issues Concerning Self-Driving Vehicles. *European Journal of Risk Regulation*, *7*(2), 335–340.

Marino, D., & Tamburrini, G. (2006). Learning robots and human responsibility. *International Review of Information Ethics*, *6*, 46–51.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183.

Millar, J. (2014): An ethical dilemma: When robot cars must kill, who should pick the victim? *Robohub,* 11.6.2014.
https://robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim/

Millar, J. (2015). Technology as Moral Proxy: Autonomy and Paternalism by Design. *IEEE Technology & Society Magazine*, *34*(2), 47–55.

Millar, J. (2017). Ethics Settings for Autonomous Vehicles. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot Ethics 2.0* (pp. 20–34). Oxford University Press.

Miller, K. W. (2011). Moral Responsibility for Computing Artifacts: "The Rules." *IT Professional*, *13*(3), 57–59.

National Highway Traffic Safety Administration [NHTSA] (2008). *National Motor Vehicle Crash Causation Survey* [Report to Congress]. U.S. Department of Transportation.

National Highway Traffic Safety Administration [NHTSA] (2018). *Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey* [Brief Statistical Summary]. U.S. Department of Transportation.

Neuhäuser, C. (2013). Roboter und moralische Verantwortung. In E. Hilgendorf (Ed.), *Robotik im Kontext von Recht und Moral* (pp. 269–287). Nomos.

Nida-Rümelin, J. (2011). *Verantwortung*. Reclam.

Noothigattu, R., Gaikwad, S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., & Procaccia, A. (2018). A Voting-Based System for Ethical Decision Making. *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, *32*(1), 1587-1594.

Norton, P. D. (2007). Street Rivals: Jaywalking and the Invention of the Motor Age Street. *Technology and Culture*, *48*(2), 331–359.

Nyholm, S. (2018a). The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass*, *13*(7), e12506.

Nyholm, S. (2018b). The ethics of crashes with self-driving cars: A roadmap, II. *Philosophy Compass*, *13*(7), e12506.

Nyholm, S. (2018c). Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics*, *24*(4), 1201–1219.

Nyholm, S., & Smids, J. (2016). The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem? *Ethical Theory and Moral Practice*, *19*(5), 1275–1289.

Ochigame, R. (2019). The invention of "ethical AI". How big tech manipulates academia to avoid regulation. *The Intercept,* 20.12.2019.
https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/

Paulo, N. (2023). The trolley problem in the ethics of autonomous vehicles. *The Philosophical Quarterly*, *73*(4), 1046–1066.

Richards, D. C. (2010). Relationship between speed and risk of fatal injury : pedestrians and car occupants. In *Road Safety Web Publication No. 16*. Department for Transport, London.

Roe, M. (2019). Who's driving that car? An analysis of regulatory and potential liability frameworks foro driverless cars. *Boston College Law Review*, *60*(1), 317–347.

Rosén E., Sander U. (2009). *Accident Analysis & Prevention,* 41/3, 536–542.

Rosén, E., Stigson, H., & Sander, U. (2011). Literature review of pedestrian fatality risk as a function of car impact speed. *Accident Analysis and Prevention*, *43*(1), 25–33.

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI*, *5*.
https://libkey.io/10.3389/frobt.2018.00015

Santoro, M., Marino, D., & Tamburrini, G. (2008). Learning robots interacting with humans: from epistemic risk to responsibility. *AI & Society*, *22*(3), 301–314.

Sauvé Mayer, S. (2011). *Aristotle on Moral Responsibility: Character and Cause*. Oxford University Press.

Sauvé Mayer, S., & Hause, J. P. (2017). Der Begriff der Verantwortung in der Antike und im Mittelalter. In L. Heidbrink, C. Langbehn, & J. Loh (Eds.), *Handbuch Verantwortung* (pp. 87–110). Springer VS.

Shafi, S., & Gentilello, L. (2007). A nationwide speed limit ≤65 miles per hour will save thousands of lives. *The American Journal of Surgery*, *193*, 719–722.

Shinar, D. (2017). *Traffic safety and human behavior* (2nd ed.). Emerald Publishing Limited.

Shinar, D. (2019). Crash causes, countermeasures, and safety policy implications. *Accident Analysis and Prevention*, *125*, 224–231.

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, *24*(1), 62–77.

Sullins, J. P. (2006). When Is a Robot a Moral Agent? *Internatioal Review of Information Ethics*, *6*(12), 23–30.

Thomson, J. J. (1985). The Trolley Problem. *The Yale Law Journal*, *94*(6), 1395–1415.

Tigard, D. W. (2021). Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible. *Cambridge Quarterly of Healthcare Ethics*, *30*(3), 435–447.

Vadeby, A., & Forsman, A. (2018). Traffic safety effects of new speed limits in Sweden. *Accident Analysis and Prevention*, *114*, 34–39.

Wishart, J., Chen, Y., Como, S., Kidambi, N., Lu, D., & Yang, Y. (2022). *Fundamentals of connected and automated vehicles*. SAE International.

Woollard, F. (2023). The New Trolley Problem: Driverless Cars and Deontological Distinctions. *Journal of Applied Philosophy*, *40*(1), 49–64.

World Health Organization [WHO] (2017). Managing Speed. Geneva.

World Health Organization [WHO] (2023). Global status report on road safety 2023. Geneva.

Zankl C., Rehrl K. (2018). Digibus 2017. Erfahrungen mit dem ersten selbstfahrenden Shuttlebus auf öffentlichen Straßen in Österreich. Salzburg: Salzburg Research. www.digibus.at

6.2 Other sources

BBC (2019). Uber in fatal crash had safety flaws say US investigators. *BBC News*, 6.11.2019. https://www.bbc.com/news/business-50312340

BBC (2020). Uber's self-driving operator charged over fatal chrash. *BBC News*, 16.9.2020. https://www.bbc.com/news/technology-54175359

Bhuiyan, J. (2023). San Francisco to get round-the-clock robo taxis after controversial vote. *The Guardian*, 11.8.2023.
https://www.theguardian.com/us-news/2023/aug/10/san-francisco-self-driving-car-autonomous-regulation-google-gm

Block, T., & Uhlenbroich, B. (2023): Wissing im Interview: Auf der Straße drohen Zustände wie bei der Bahn. *Bild.de*, 22.1.2023.
https://www.bild.de/bild-plus/politik/inland/politik/wissing-im-interview-auf-der-strasse-drohen-zustaende-wie-bei-der-bahn-82633866.bild.html

Bundesamt für Justiz [BfJ] (2023a). Straßenverkehrsgesetz (StVG). Last modification November 21st 2023 (BGBl. 2023/315).
https://www.gesetze-im-internet.de/stvg/

Bundesamt für Justiz [BfJ] (2023b). Autonome-Fahrzeuge-Genehmigungs-und-Betriebs-Verordnung (AFGBV). Last modification July 20th 2023 (BGBl. 2023/199).
https://www.gesetze-im-internet.de/afgbv/

Bundesministerium für Digitales und Verkehr [BMDV] (2021). Gesetz zum autonomen Fahren tritt in Kraft. 27.7.2021.
https://bmdv.bund.de/SharedDocs/DE/Artikel/DG/gesetz-zum-autonomen-fahren.html

Bundesministerium für Inneres [BMI] (2024): 396 Verkehrstote auf Österreichs Straßen im Jahr 2023. Press release on 1.1.2024, Vienna.
https://www.ots.at/presseaussendung/OTS_20240101_OTS0011/396-verkehrstote-auf-oesterreichs-strassen-im-jahr-2023

Bundesministerium für Klimaschutz, Umwelt, Energie, Mobilität, Innovation & Technologie [BMK] (2019). Neuerungen zum Automatisierten Fahren. Was ist erlaubt? *BMK Infothek,* 14.3.2019.
https://infothek.bmk.gv.at/novellierung-der-automatisiertes-fahren-verordnung/

Bussgeldkatalog (2023). Level 3-autonomes Fahren: Definitionen, Autos, Gesetze. *bussgeldkatalog.org*, 27.9.2023.
https://www.bussgeldkatalog.org/level-3-autonomes-fahren/

Capoot, A., & Piazza, J. (2023). Uber begins offering rides in self-driving Waymo cars. *CNBC*, 26.10.2023.
https://www.cnbc.com/2023/10/26/uber-begins-offering-rides-in-self-driving-waymo-cars.html

Donath, A. (2023). Autonomes Fahren in Deutschland mit 130 km/h erlaubt. *Golem*, 3.1.2023.
https://www.golem.de/news/mehr-geschwindigkeit-autonomes-fahren-in-deutschland-mit-130-km-h-erlaubt-2301-170873.html

Ethik-Kommission (2017). Automatisiertes Und Vernetztes Fahren. Bericht: Juni 2017. Bundesministerium für Verkehr und digitale Infrastruktur

European Parliament [EP] (2016). Automated vehicles in the EU. Briefing, January 2016. European Parliamentary Research Service (EPRS)

Evers, A., & Bosa, D. (2023). Why San Francisco's robotaxi rollout has been such a mess. *CNBC*, 4.11.2023.
https://www.cnbc.com/2023/11/04/why-san-franciscos-robotaxi-rollout-has-been-such-a-mess.html

Goswami, R. (2023). Cruise robotaxi gets in accident with San Francisco fire truck, one injured. *CNBC*, 18.8.2023.
https://www.cnbc.com/2023/08/18/cruise-self-driving-car-in-san-francisco-fire-truck-crash-one-injured.html

Griffin, A. (2015). Self-driving cars will need to be programmed to kill their owners, academics warn, and people will have to choose who will die. *Independent,* 27.10.2015.
https://www.independent.co.uk/tech/selfdriving-cars-will-need-to-be-programmed-to-kill-their-owners-academics-warn-and-people-will-have-to-choose-who-will-die-a6710956.html

Holtermann, F. (2023). Kalifornien lässt "Drive Pilot" zu. *Handelsblatt*, 9.6.2023.
https://www.handelsblatt.com/mobilitaet/elektromobilitaet/mercedes-autopilot-kalifornien-laesst-drive-pilot-zu/29195664.html

Hubik, F., & Holtermann, F. (2022). Mercedes nennt erstmals Preise für autonomes Fahrsystem. *Handelsblatt*, 5.5.2022.
https://www.handelsblatt.com/unternehmen/industrie/autonomes-fahren-mercedes-nennt-erstmals-preise-fuer-autonomes-fahrsystem/28310938.html

Jeffs, J., & He, X. (2023): Autonomous Cars, Robotaxis and Sensors 2024-2044. *IDTechEx Research*.
https://www.idtechex.com/en/research-report/autonomous-cars-robotaxis-and-sensors-2024-2044/953

Kraftfahrbundesamt [KBA]. Gesetzgebung zum autonomen Fahren.
https://www.kba.de/DE/Themen/Marktueberwachung/Produktpruefungen/AutomatisiertesAutonomesFahren/Gesetzgebung/gesetzgebung_autonomes_fahren_node.html

Lee, L. (2023). Self-driving car company Cruise agreed to reduce its driverless fleet by 50% after a spate of recent crashes. San Francisco officials previously pushed for a slower rollout of robotaxis. *Business Insider*, 20.8.2023.
https://www.businessinsider.com/cruise-slashes-robotaxi-fleet-half-after-collision-incidents-san-francisco-2023-8

Lewis, A. (2014): Jaywalking: How the car industry outlawed crossing the road. *BBC News*, 12.2.2014. https://www.bbc.com/news/magazine-26073797

Lin, P. (2013). The Ethics of Autonomous Cars. *The Atlantic,* 8.10.2013.
https://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/

Lin, P. (2014). Here's a Terrible Idea: Robot Cars With Adjustable Ethics Settings. *Wired,* 18.8.2014. https://www.wired.com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings/

Marshall, A. (2023). GM's Cruise Loses Its Self-Driving License in San Francisco After a Robotaxi Dragged a Person. *Wired*, 24.10.2023.
https://www.wired.com/story/cruise-robotaxi-self-driving-permit-revoked-california/

Nikkei (2023). Honda and GM to launch driverless taxis in Japan. *Nikkei Asia*, 19.10.2023. https://asia.nikkei.com/Business/Automobiles/Honda-and-GM-to-launch-driverless-taxis-in-Japan

oesterreich.gv.at (2023). Automatisierte Fahrsysteme. Version as of 17.2.2023. https://www.oesterreich.gv.at/themen/mobilitaet/kfz/Seite.061910.html

ORF Wien (2019). Fahrerlose Busse nach Unfall gestoppt. *wien.orf.at*, 18.7.2019. https://wien.orf.at/stories/3004979/

ORF Wien (2021). Aus für autonome Busse der Wiener Linien. *wien.orf.at*, 29.6.2021. https://wien.orf.at/stories/3110594/

Pladson, K. (2021). Teslas Unfall und die großen Hürden der Branche. *DW Wirtschaft*, 23.4.2021. https://www.dw.com/de/autonomes-fahren-teslas-unfall-und-die-gro%C3%9Fen-h%C3%BCrden-der-branche/a-57314799

PR Newswire (2023a). Baidu Launches Commercial Fully Driverless Ride-Hailing Service in Shenzhen, Expanding Nationwide Operations. 16.6.2023.
https://www.prnewswire.com/news-releases/baidu-launches-commercial-fully-driverless-ride-hailing-service-in-shenzhen-expanding-nationwide-operations-301853087.html

PR Newswire (2023b). Baidu Becomes First in China to Offer Driverless Airport Rides. 25.8.2023. https://www.prnewswire.com/news-releases/baidu-becomes-first-in-china-to-offer-driverless-airport-rides-301910249.html

Rechtsinformationssystem des Bundes [RIS] (2023): Automatisiertes Fahren Verordnung – AutomatFahrV. Version as of 13.11.2023.
https://www.ris.bka.gv.at/Bundesrecht/

Rechtsinformationssystem des Bundes [RIS] (2024a). Eisenbahn- und Kraftfahrzeughaftpflichtgesetz – EKHG. Version as of 27.1.2024.
https://www.ris.bka.gv.at/Bundesrecht/

Rechtsinformationssystem des Bundes [RIS] (2024b). Straßenverkehrsordnung – StVO. Version as of 21.2.2024.
https://www.ris.bka.gv.at/Bundesrecht/

Rosevear, J. (2023). Volkswagen will start testing self-driving cars in Austin as it moves on from Argo AI. *CNBC*, 6.7.2023.
https://www.cnbc.com/2023/07/06/volkswagen-mobileye-autonomous-vehicle-tests-to-launch-in-austin.html

Salzburg Research (2023). Der Digibus ist wieder unterwegs: Vorbereitungen für den Pilotbetrieb 2023. *Salzburg Research*, 18.1.2023.
https://www.salzburgresearch.at/presseaussendung/der-digibus-ist-wieder-unterwegs-vorbereitungen-fuer-den-pilotbetrieb-2023/

Shepardson, D. (2017). Tesla driver in fatal 'Autopilot' crash got numerous warnings: U.S. government. *Reuters*, 20.6.2017.
https://www.reuters.com/article/us-tesla-crash-idUSKBN19A2XC/

Society of Automotive Engineers [SAE] (2021). SAE Levels of Driving Automation. Refined for Clarity and International Audience. Posted on 3rd May, 2021.
https://www.sae.org/blog/sae-j3016-update

Stromberg, J. (2015): The forgotten history of how automakers invented the crime of "jaywalking". *Vox*, 4.11.2015. https://www.vox.com/2015/1/15/7551873/jaywalking-history

SURAAA (2023). Statement zu Zwischenfall vom 9. August. Pressemitteilung vom 10. August 2023.
https://www.suraaa.at/wp-content/uploads/sites/6/2023/08/PM_SURAAA_10_08_23.pdf

SURAAA. Smart Urban Region Austria Alps.
https://www.suraaa.at/

Templeton, B. (2023). Cruise Robotaxi Drives Into Wet Concrete; Waymo Shows Off Same Route. *Forbes*, 17.8.2023
https://www.forbes.com/sites/bradtempleton/2023/08/17/cruise-robotaxi-drives-into-wet-concrete-waymo-shows-off-same-route/?sh=461813fe6bfd

Theile, G. (2022). Mercedes bringt automatisiertes Fahren auf den Markt. *Frankfurter Allgemeine Zeitung [FAZ]*, 6.5.2022.
https://www.faz.net/pro/d-economy/mercedes-bietet-als-erster-hersteller-autonomes-fahren-auf-level-3-an-18009871.html

Thompson, C. (2022): The Invention of 'Jaywalking'. *Marker/Medium*, 29.3.2022.
https://marker.medium.com/the-invention-of-jaywalking-afd48f994c05

UNECE (2022). UN Regulation extends automated driving up to 130 km/h in certain conditions. *UNECE Press Releases*, 22.6.2022.
https://unece.org/sustainable-development/press/un-regulation-extends-automated-driving-130-kmh-certain-conditions

Yadron, D., & Tynan, D. (2016). Tesla driver dies in first fatal crash while using autopilot mode. *The Guardian*, 1.7.2016.
https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk