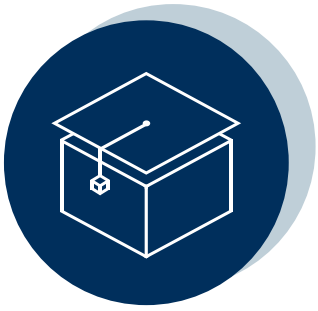

BISE Student

https://bise-student.io

MASTER'S THESIS

Fairness, Accountability, Transparency, Explainability: A Qualitative Approach towards Trustworthy AI in Autonomous Vehicles

Publication Date: 2022-02-22

Author
Florian GROSS
Karlsruhe Institute of Technology
Karlsruhe, Germany
florian.gro112@me.com
0x5d50c27401e84A2641049312Ca7d193dB8Ce2052

Abstract

A successful market launch of autonomous vehicles (AV) is only possible if users trust the AV and thus the artificial intelligence (AI) powering the vehicle. To conceptualize trust in AI, researchers recently started using so-called FATE characteristics (Fairness, Accountability, Transparency, Explainability). Until now, the FATE characteristics have not been contextualized by AV specific FATE attributes. This work aims to answer how to establish trust with the FATE characteristics in AVs by conducting a content analysis of 33 AV provider websites and 5 expert interviews. The findings suggest that in the context of AVs, the C-FATS characteristics (Certifiability, Fairness, Accountability, Transparency, Safety) better conceptualize trust. Applying the results of the analysis, a framework for TAI in the context of AVs encompassing 91 C-FATS attributes is developed. In addition, differences between providers and experts in the conceptualization of TAI in AVs are highlighted and interdependencies that need to be considered...

Keywords: autonomous vehicles, Trustworthy artificial intelligence
Methods: content analysis, expert interviews

Submission Date: 2022-02-22
Submission Contract: 0xb8fB968B01708a55371B34710F1bb748210F2baB

Fairness, Accountability, Transparency, Explainability: A Qualitative Approach towards Trustworthy AI in Autonomous Vehicles

Master Thesis

by

Florian Groß

Industrial Engineering and Management M.Sc.

Institute of Applied Informatics and Formal Description Methods (AIFB)

KIT Department of Economics and Management

Advisor:	Prof. Dr. Ali Sunyaev
Second Advisor:	Prof. Dr. Andreas Oberweis
Supervisor:	M.Sc. Maximilian Renner
Submitted:	February 5, 2022

Abstract

A successful market launch of autonomous vehicles (AV) is only possible if users trust the AV and thus the artificial intelligence (AI) powering the vehicle. To conceptualize trust in AI, researchers recently started using so-called FATE characteristics (Fairness, Accountability, Transparency, Explainability). Until now, the FATE characteristics have not been contextualized by AV specific FATE attributes. This work aims to answer how to establish trust with the FATE characteristics in AVs by conducting a content analysis of 33 AV provider websites and 5 expert interviews. The findings suggest that in the context of AVs, the C-FATS characteristics (Certifiability, Fairness, Accountability, Transparency, Safety) better conceptualize trust. Applying the results of the analysis, a framework for TAI in the context of AVs encompassing 91 C-FATS attributes is developed. In addition, differences between providers and experts in the conceptualization of TAI in AVs are highlighted and interdependencies that need to be considered in establishing TAI are identified. Since the interdependencies between trust-building attributes challenge the distinction between "trust in technology" and "trust in organizations", researchers are tasked to generalize extended trust concepts in the context of AI systems to include transfer of trust.

Contents

List of Abbreviations

AI Artificial Intelligence.

AV Autonomous Vehicle.

C-FATS Certifiability, Fairness, Accountability, Transparency, Safety.

DGA Data Governance Act COM/2020/767 of the European Commission.

EAIA European Artificial Intelligence Act COM/2021/206 of the European Commission.

FATE Fairness, Accountability, Transparency, Explainability.

GDPR General Data Protection Regulation 2016/679 of the European Union.

HMI Human-Machine-Interface.

IS Information System.

ML Machine Learning.

OEM Original Equipment Manufacturer.

TAI Trustworthy Artificial Intelligence.

XAI Explainable Artificial Intelligence.

List of Figures

List of Tables

1 Problem and Aim of the work

In recent years, Artificial Intelligence (AI) has become a driver of innovation and a much-discussed topic in society. AI is defined as hardware and software components executing tasks, which usually require human intelligence (Poole, Mackworth, & Goebel, 1997; Russell & Norvig, 1995, pp. 4–8). As the performance of algorithms advance, their impact, and importance for companies, users, and the society as a whole increases rapidly (Makridakis, 2017, pp. 47–53; Shin & Park, 2019, p. 277). One much-discussed application of AI, whose market launch is on the horizon, is autonomous driving. Vehicle manufacturers, tech companies and startups around the world are working to make Autonomous Vehicle (AV) (i.e. self-driving passenger cars) a reality for the masses. For example, Crunchbase Inc. lists over 200 companies world-wide under the search term "autonomous driving" (2022). Moreover, Tesla announced its vehicles are supposed to be ready for full autonomous driving at the end of 2021 (Hyatt, 2021).

There are numerous ways AVs are beneficial for society. Enabling mobility for disabled people, increasing the safety of traffic and cutting greenhouse gas emissions (Andersson & Ivehammar, 2019, pp. 125–127). To realize these benefits as soon as the technology is available, rapid adoption by users is desirable.

A successful market launch of AVs is only possible, if providers are able to build acceptance in their technology among consumers (Nastjuk, Herrenkind, Marrone, Brendel, & Kolbe, 2020, pp. 11–13, Renner, Lins, Söllner, Thiebes, & Sunyaev, 2021, Renner, Lins, Söllner, Thiebes, & Sunyaev, 2022. In order to achieve user acceptance and ultimately the adoption and use of technology, trust in the technology must be established, as shown by Mcknight, Carter, Thatcher, and Clay (2011, pp. 14–16) and in the Trust-TAM concept by Gefen, Karahanna, and Straub (2003, pp. 72–76). Additionally, providers have to take the nature of algorithmic technologies into account when establishing trust (Benbasat & Wang, 2005, p. 79).

Therefore, providers search for ways to foster trust in their new AI-based technologies. Specifically, in the context of AVs (Othman, 2021, p. 3-5). Unfortunately, research has shown that providers of products, services and technologies are faced with limited trust into AI-based technologies by consumers (Thiebes, Lins, & Sunyaev, 2020, pp. 458–459; Ipsos, 2018). Reasons for this hesitation towards AI-based technologies are recent issues and famous incidents regarding AI's fairness, accountability, transparency and explainability (Barredo Arrieta et al., 2020, pp. 103–108; Thiebes et al., 2020, pp. 456–458; Shin, 2020, p. 278–279).

Trust in AI is therefore a hot topic for researchers in the last years. Various stakeholders, such as scientists, authorities, governments, and companies developed guidelines for Trustworthy Artificial Intelligence (TAI). Thiebes et al. (2020, pp. 452–456) define five prin-

ciples for TAI (beneficence, non-maleficence, autonomy, justice and explicability). The Independent High-Level Expert Group on Artificial Intelligence of the European Commission (2019) proposed seven requirements for TAI (human agency and oversight, technical robustness and safety, privacy and data governance, diversity – non-discrimination and fairness, societal and environmental well-being, transparency, accountability). In his work, Floridi discusses the requirements proposed by the High-Level Expert Group (2019, pp. 261–262). Whereas Barredo Arrieta et al. (2020, pp. 103–108) summarize six principles for responsible AI (fairness, privacy, accountability, ethics, transparency, security and safety). Likewise, research recently examined antecedents of TAI (e.g. Explainable Artificial Intelligence (XAI) (Markus, Kors, & Rijnbeek, 2020)) and analyzed the impact of trust on user perceptions. For example, Shin (2020, pp. 550–557) investigates user perceptions of algorithmic decisions regarding Fairness, Accountability, Transparency, Explainability (FATE). Moreover, Shin and Park (2019, pp. 278–279) analyze the impact of fairness, transparency, and accountability on the use and adoption of algorithmic services.

Fueled by the work of Shin (2019; 2020), the efforts to conceptualize trust in AI-based systems have recently focused on the so-called FATE characteristics. But in the context of AVs, the research currently neglects the perspective of the FATE characteristics. Despite the need for use-case specific trust conceptualizations, described by Shin and Park (2019, pp. 281–282), Jarvenpaa, Shaw, and Staples (2004, p. 264) as well as M. K. Lee (2018, p. 13), current research fails to contextualize the FATE characteristics through context specific FATE attributes for AVs.

To close this gap, this thesis answers the following research question: How to establish trust with the FATE characteristics in AVs? To answer this research question, the thesis has the following objectives.

1. Assessment of how AV providers make use of the FATE characteristics to establish trust in their technology.
2. Derivation of a framework for various FATE attributes used by AV providers to establish trust in their technology.
3. Discovery of the interdependencies of FATE attributes in building TAI in the context of AVs.

The aim of the thesis is the provision of a framework of the FATE attributes which help to establish TAI in the context of AVs. Moreover, interdependencies between the FATE attributes in establishing TAI are described.

2 Background

2.1 Autonomous Vehicles

Autonomous driving is a technology field with the goal of developing vehicles which safely take over acceleration, deceleration, and navigation from a human driver to ultimately provide vehicles that drive on their own. In the context of this thesis, only self-driving passenger cars are considered. Technology system for autonomous driving, consisting of hardware and software components, can be classified by the level of autonomy they provide. SAE J3016 ranks systems from no automation (Level 0) to full autonomy without human interaction required (Level 5) (SAE International, 2021b). An overview of all SAE Levels can be found in Figure 1 (SAE International, 2021a).

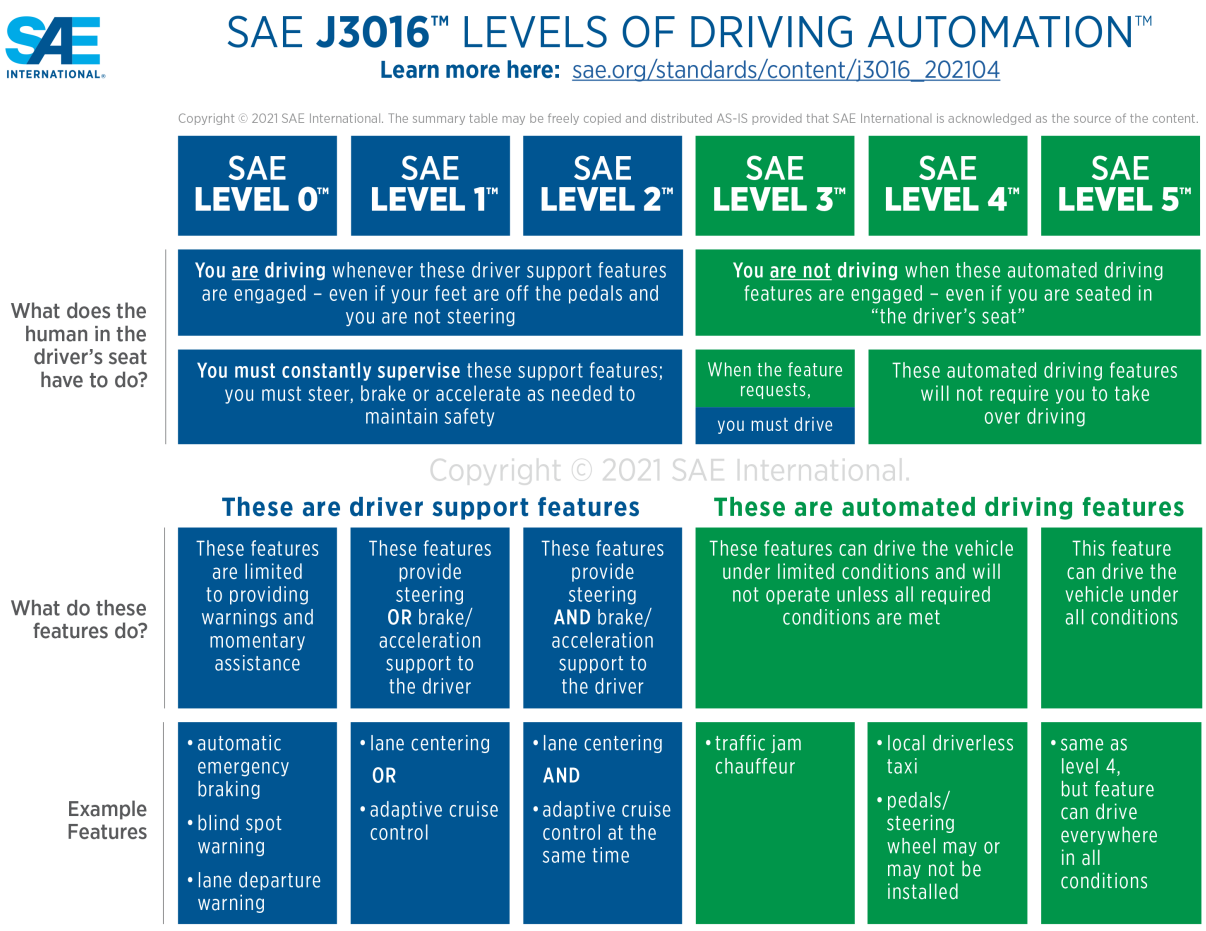

Figure 1: Levels of Driving Automation (SAE International, 2021a)

2.2 Trust

Trust is of great importance for human interaction. It is the cornerstone for cooperation between people and necessary for the use of new technologies, as shown by (Mcknight et al., 2011, pp. 1-25; Gefen et al., 2003, pp. 51-90; Jarvenpaa et al., 2004, pp. 250-267). Rousseau, Sitkin, Burt, and Camerer (1998, pp. 394–395) define trust as an individual's willingness to depend on another party because of the counterpart's characteristics. Research in recent decades examined trust in different contexts and from different perspectives. Because of the manifold perspectives on the concept of trust, there is no commonly accepted definition (Thiebes et al., 2020, p. 450; Lansing & Sunyaev, 2016, p. 61; Rousseau et al., 1998, p. 394), but the need for a contextualization of trust concepts (Jarvenpaa et al., 2004, p. 264).

In Information System (IS) research, trust is conceptualized from a dualistic perspective. Research finds that there are distinguishing factors between trust in people and trust in technology (Thiebes et al., 2020, p. 449–551; Mcknight et al., 2011, pp. 1–25). An overview of the duality of trust is given in Table 1.

Trust in persons	Trust in technology	
(e.g., Mayer et al. 1995; McKnight et al. 2002).	Trust in IT artifacts based on system characteristics (e.g., McKnight et al. 2011; Thatcher et al. 2010)	Trust in automation technology and autonomous systems (e.g., J. D. Lee and See 2004)

Table 1: Overview of the Duality of Trust by Thiebes et al. (2020, p. 449)

In their integrative model of organizational trust Mayer, Davis, and Schoorman (1995), and McKnight, Choudhury, and Kacmar (2002) in their work on trust measures for e-commerce, deliver a set of trusting beliefs related to persons. Whereas, Thatcher, McKnight, Baker, Arsal, and Roberts (2011), Mcknight et al. (2011) as well as J. D. Lee and See (2004) deliver trusting beliefs related to technology.

2.3 FATE Concept

FATE stands for Fairness, Accountability, Transparency and Explainability. These are concepts proposed by research which need to be fulfilled to achieve TAI. In the following, the four concepts will be introduced briefly.

Generally, fairness can be defined as impartial or equitable treatment of individuals or demographic groups (Yang & Stoyanovich, 2017, p. 1). But there is no generally accepted definition of fairness or algorithmic fairness (Shin & Park, 2019, p. 278). Instead, fairness is always context specific. In the case of an AI system, preventing discriminatory or biased treatment are two ways to achieve greater perceived fairness (Shin & Park, 2019, p. 281).

As the consequences of unintended actions of AI systems can be severe, the question of

accountability for the actions of an AI system arises (Shin & Park, 2019, p. 278). To date, however, there is no comprehensive legal regulation on the accountability of AI systems. Some argue that a comprehensive regulation should hold providers, not individual developers, accountable for the actions of their algorithmic services (Diakopoulos, 2016, pp. 56–62).

Algorithmic transparency is defined as the requirement that users can understand how a decision is made by a specific AI system (Shin & Park, 2019, p. 278). In a more detailed definition, transparency is described as the combination of input and algorithm that must be visible and comprehensible for the user (Diakopoulos & Koliska, 2017, p. 813).

Closely related to the concept of transparency is explainability. Explainability in algorithms refers to methods and techniques that provide justification of an AIs output behavior, which can be conceived by humans (Ehsan & Riedl, 2019, p. 2). Explainability therefore is the degree to which an instance's feature values are linked to its (machine learning) model prediction in a way that is humanly accessible (Rai, 2020, pp. 137–141; Shin, 2020, pp. 7–8).

2.4 Artificial Intelligence

AI is demonstrated by machines, whereas natural intelligence is found in humans (Russell & Norvig, 1995, pp. 4–8). Computer science defines AI as studying the design of intelligent agents (Poole et al., 1997). Intelligent agents are software and hardware components executing tasks, that usually require human intelligence (Crayton, 2019). Closely related to the term AI is the term Machine Learning (ML). ML is the science of making computers learn without explicitly programming them (Samuel, 1959). In AVs, AI is used for perception, localization, planning, vehicle control, and system management (Omeiza, Webb, Jirotka, & Kunze, 2021).

2.5 Trustworthy Artificial Intelligence

As mentioned in Section 2.2, user trust in a new technology is a key element for adoption. In order to exploit the full potential of a technology for individuals, organizations and societies, trust must be created among users. This is especially true for applications of AI technology. TAI is one approach to create trust in the development, deployment, and use of AI technologies (Thiebes et al., 2020, p. 447). Scientists, authorities, governments, and organizations define different criteria for TAI. An overview of some approaches can be found in Table 2.

Concept	Trustworthy AI	Trustworthy AI	Responsible AI
Author	Thiebes et al. (2020)	Independent High-Level Expert Group on Artificial Intelligence of the European Commission (2019)	Arrieta et al. (2019)
Elements	beneficence non-maleficence autonomy justice explicability	human agency and oversight technical robustness and safety privacy and data governance diversity – non-discrimination and fairness societal and environmental well-being transparency accountability	fairness privacy accountability ethics transparency security and safety

Table 2: Trustworthy Artificial Intelligence Concepts in Literature

3 Methodology

To achieve the outlined goals, a two-step empirical research approach is chosen. In the first step, a content analysis of AV providers' websites is carried out, the results of which are supplemented in a second step with the findings of expert interviews.

3.1 Content Anlysis

As the first step, a content analysis of prominent AV providers in Europe and North America is conducted following the methodology proposed by Krippendorff (2004). First, a list of automotive Original Equipment Manufacturer (OEM), Tier-1 suppliers, startups and tech companies in the AV space is created. Since AVs level 4 and 5 are not yet available on a large scale, manufacturers who announced the development of such systems are also being studied. Included in the first step of the analysis are 33 companies which already provide AVs or committed to actively develop such technologies. The analysis is limited to providers from Europe (including Israel) and North America in order to perform an analysis without the help of translation services. This limitation does not pose a major risk to the validity of the content analysis because only one of the global top ten providers currently developing AV technology is located outside of Europe or North America (Guidehouse Inc., 2020). The analyzed companies and related websites can be found in Table 3.

In the second step of the content analysis, the providers' websites are searched for information on TAI in the context of AVs. Of particular relevance is any information on guidelines or standards related to the development of AI systems or the properties to be achieved by the AI. Websites of third parties are also included in the analysis, if the providers' website or subsidiaries refer to the third party. Since not all AV providers published specific information or guidelines about their AI systems related to trust, only the ten providers identified as relevant to TAI in Table 3 are included in the content analysis. The references stated in Table 3 link to the specific content relevant for the content analysis.

Before the coding process, the content of the websites is prepared. Especially, irrelevant content on the website is excluded from the analysis (e.g. introductory sentences or information on the company). The relevant content is stored in a table in the same structure as on the website, by paragraphs. In the following step of coding, the paragraphs are divided as needed to allow for more accurate coding. The coding process is conducted according to Corbin and Strauss (2014). As a first step, open coding is performed in order to identify measures and concepts, proposed by providers, to increase trust in their AI technology. The open coding is followed by an axial coding to align codes and group codes

in higher ranking categories. For example, the codes "Internal Audit" and "External Audit" are combined to form the code "Auditability". By comparing codes, 6 characteristics of TAI in AVs are created on the highest level that form the results presented in Section 4. Overall, 67 codes are derived from 215 text segments related to 10 providers. During the coding process, constant comparison and memoing is used to retrieve as much information as possible.

Company	Type	Location	Website	TAI Relevance	References
Aptiv plc	Tier-1	Ireland	www.aptive.com		
Argo AI LLC	Startup	USA	www.argo.ai		
AUDI AG	OEM	Germany	www.audi.com	YES	(AUDI AG, 2018a), (AI4People, 2020)
Aurora Innovation Inc.	Startup	USA	www.aurora.tech		
BMW AG	OEM	Germany	www.bmw.com	YES	(BMW AG, 2020)
CARIAD SE	Tech	Germany	www.cariad.technology		
Continental AG	Tier-1	Germany	www.continental.com	YES	(Continental AG, 2020)
Cruise LLC	Startup	USA	www.getcruise.com		
Daimler AG	OEM	Germany	www.daimler.com	YES	(Daimler AG, 2021)
Dr. Ing. h.c. F. Porsche AG	OEM	Germany	www.porsche.com	YES	(Dr. Ing. h.c. F. Porsche AG, 2019)
Faurecia SE	Tier-1	France	www.faurecia.com		
Ford Motor Company	OEM	USA	www.ford.com	YES	(Ford Motor Company, n.d.)
General Motors Company	OEM	USA	www.gm.com		
Infineon Technologies AG	Tech	Germany	www.infineon.com		
Jaguar Land Rover Ltd	OEM	UK	www.jaguarlandrover.com		
Magna International Inc.	Tier-1	Canada	www.magna.com		
Mahle GmbH	Tier-1	Germany	www.mahle.com		
Mobileye Vision Technologies Ltd	Tech	Israel	www.mobileye.com		
Motional, Inc.	Tech	USA	www.motional.com		
NVIDIA Corporation	Tier-1	USA	www.nvidia.com		
Polestar AB	OEM	Sweden	www.polestar.com		
Qualcomm Technologies, Inc.	Tech	USA	www.qualcomm.com		
Renault S.A.	OEM	France	www.renaultgroup.com	YES	(Villani Mission on artificial intelligence, 2019a), (Villani Mission on artificial intelligence, 2019b)
Robert Bosch GmbH	Tier-1	Germany	www.bosch.com	YES	(Robert Bosch GmbH, 2020)
Stellantis NV	OEM	Netherlands	www.stellantis.com		
Tesla, Inc.	OEM	USA	www.tesla.com		
Valeo S.A.	Tier-1	France	www.valeo.com	YES	(Valeo S.A., 2019), (Villani Mission on artificial intelligence, 2019b)
Veoneer, Inc.	Tier-1	Sweden	www.veoneer.com		
Volkswagen AG	OEM	Germany	www.volkswagenag.com	YES	(Volkswagen Group Machine Learning Research Lab, 2020b), (Volkswagen Group Machine Learning Research Lab, 2020a)
Volvo Personvagnar AB	OEM	Sweden	www.volvocars.com		
Waymo LLC	Startup	USA	www.waymo.com		
ZF Friedrichshafen AG	Tier-1	Germany	www.zf.com		
Zoox, Inc.	Startup	USA	www.zoox.com		

Table 3: List of Analyzed Websites about TAI from AV Providers

3.2 Expert Interviews

To enrich the information gathered from the content analysis, expert interviews are conducted. To gain a deeper understanding of how companies in the automotive space try to build trust in their AI technology, employees of OEMs, Tier-1 suppliers and software companies are selected to be interviewed. The goal of the interviews is to obtain background information that has not been published, to ultimately extend the framework of FATE attributes which help to establish TAI in the context of AVs. The employees to be interviewed should work in the automotive industry in the field of autonomous driving, software or AI. For the acquisition of interviewees, a flyer with information about autonomous driving and the topic of research is created (see Appendix A.1). The flyer is shared with personal contacts and on the social network LinkedIn. In addition, experts for (Trustworthy) AI, who are identified as interesting interview partners, are targeted via email or LinkedIn messages. This leads to contact with 10 potential interviewees. Out of these 10 potential interviewees, eight are selected for an interview based on their role in the automotive sector and their involvement in autonomous driving and software development. Interviews with five of these selected experts are conducted during the analysis period from October to November 2021.

The interviews are conducted in a semi-structured nature, following Myers (2013, pp. 121–133) approach. For this purpose, an interview guideline is created with questions regarding measures and principles for TAI in AVs (see Appendix A.2). The experts are first asked to name possible measures to achieve TAI. Afterwards they are questioned specifically about the characteristics of fairness, accountability, transparency and explainability. Interviews are conducted and recorded via Microsoft Teams to allow for an uninterrupted interview atmosphere, and accurate transcription. The transcription rules are adapted from Dresing and Pehl (2018). The following rules are applied in the transcription process.

- The interviewing person is indicated by an "I:", the interviewee by an "R:" with each speaker's contribution as an own paragraph
- The spoken word is transcribed literally including colloquial phrases
- Stuttering, word slurring and punctuation are smoothed in favor of readability
- Reception signals that do not interrupt the other person's flow of speech are not transcribed
- Incomprehensible words are marked by "(Incomprehensible)"
- Pauses from approx. 1 second are marked by "(.)", from approx. 2 seconds are marked by "(..)", from approx. 3 seconds are marked by "(...)"

- Time marks for the paragraphs are not displayed
- The lines of the transcripts were numbered for easier coding

Since AVs are a technology that differentiates competitors from one another, the interviewees are granted anonymity with regard to their company and themselves. Therefore, information that would reveal the interviewees' employer or their own identity is removed from the transcripts and replaced by a more general term (e.g. "We at [company name] use..." becomes "We at the company use...") or censored. In addition, it is possible for companies to check their employees' responses for classified information, if desired. This is possible because no corruption of the results by the companies is expected. The transcripts were reviewed in only one case and approved for use without modification. An overview of the conducted interviews can be found in Table 4.

Interview	Date	Duration	Transcript
Interview 1	10/07/2021	41 minutes	Appendix A.3
Interview 2	10/19/2021	63 minutes	Appendix A.4
Interview 3	10/20/2021	39 minutes	Appendix A.5
Interview 4	10/20/2021	51 minutes	Appendix A.6
Interview 5	11/18/2021	29 minutes	Appendix A.7

Table 4: Interviews Conducted and Analyzed

In order to achieve better comparability of the results, the transcripts of the interviews are coded in the same way as in the content analysis (Corbin & Strauss, 2014). First, an open coding is conducted, followed by theoretical coding using the codes derived in the content analysis. A round of axial coding is used to identify and structure the FATE attributes in the context of AVs. In total, 73 codes are derived, with 5 codes on the highest level.

3.3 Development of Framework

To derive a common framework from the results of the content analysis and the interviews, the two frameworks are compared and combined. This results in a combined framework of measures and principles, called attributes, establishing TAI in AVs. In the process of comparing and combining the two frameworks, the following steps are taken iteratively, starting at the top of the hierarchy.

1. Comparison of the two frameworks on each level of hierarchy

2. Identification of differences originating from synonyms or conceptual relationships
3. Identification of differences based on conceptual differences in the data used
4. Decision in case of differences based on the consideration of arguments and findings of third parties

One example of this process is the attribute explainability. To combine the results of content analysis and interviews, this attribute was assigned to the attribute transparency. The resulting hierarchy of codes is visualized as a tree for better comprehension.

3.4 Identification of Interdependencies

For the derivation of interdependencies, coded attributes are analyzed regarding their conceptual interdependencies. To this end, provider websites and interview transcripts are searched for evidence of interdependencies. Furthermore, information obtained during coding by memoing is used. In the interviews, explicit questions are asked about possible interdependencies between the attributes discussed. All information is gathered and compared to identify interdependencies. Identified interdependencies between attributes are documented and verified through third party findings, if possible. To show the dependencies, connecting lines are entered into the visualization of the framework.

4 Trustworthy Artificial Intelligence in Autonomous Vehicles

This chapter presents the results of the content analysis of the provider websites (see Table 3) and expert interviews (see Table 4). In the following, all identified attributes will be briefly described and presented in the derived framework. The highest level of hierarchy consists of **Certifiability**, **Fairness**, **Accountability**, **Transparency** and **Safety**, short **C-FATS**. Each of the attributes will be further explained in the following chapters. The concepts on the highest hierarchical level are described by underlying concepts in the subchapters. If the underlying concept in the subchapter is in turn described by several concepts, these are marked bold. Concepts subordinate to the bold marked concepts in the hierarchy are marked in italic. An overview over the full framework can be found in Figure 2. All the attributes described have been published by AV providers or were mentioned in the interviews with experts, and have an impact on user trust in AV technology.

4.1 Fairness

Fairness of an AI system can be decomposed into several attributes, which help to create trust in such a system. These attributes are **Lawfulness**, **Data Protection & Privacy**, **Ethic Guidelines** as well as ensuring fairness in **Traffic Flow**.

4.1.1 Lawfulness

Lawfulness in this context means that the AI system in an AV is complying with all regulatory requirements. Compliance with regulatory requirements must be ensured to gain users trust in the technology. Bringing vehicles onto the market with AI systems that do not meet the legal requirements is likely to result in severe, lasting losses of trust. The loss of trust due to non-compliant technology was observed, for example, during the so-called “Dieselgate” emissions scandal (Statista, 2016). Therefore, the compliance of technology with governing laws is a driver of trust. Furthermore, in an interview, an expert explains that it is not enough to do what is legally required (A.4 l. 483–486). In addition to regulatory compliance, adherence to more stringent corporate rules, where applicable, aid in providing trust in the AI system. For example, Continental AG demands “Usage of AI that Complies with Laws, Regulations, and Continental Corporate Rules, Standards and Instructions” (Continental AG, 2020, p. 2). The adherence to company guidelines and international standards is even more important as long as there are no laws and ultimately court rulings on how to regulate AVs and underlying AI technology (A.5 l. 121–128; A.6 l. 50–57).

4.1.2 Data Protection & Privacy

Data Protection and Privacy is an attribute located within the context of fairness, because the violation of privacy in most cases does not pose an immediate safety risk, but is perceived "unfair" by users. Experts mention that considering privacy is important in the context of AI systems (A.7 l. 183–187; A.3 l. 148–152). Due to high profile incidents of data misuse, the handling of data by companies is not perceived as fair (Public Affairs Council, 2021; NBC News, 2018). By designing AI systems that protect data and privacy, users trust in an AI system can be increased. Daimler writes on its website that data protection is a quality indicator for the company (Daimler AG, 2021). By doing so, they are trying to increase the perceived fairness of their systems.

One method of signaling that an AI system protects data and ensures privacy is to build a system compliant with **Standards**. Data protection and privacy standards can be laws like the General Data Protection Regulation 2016/679 of the European Union (GDPR) (European Union, 2016) or company guidelines and industry standards like the IEEE P7002[1] (IEEE Standards Association, 2016; AI4People, 2020, p. 22). Both can help to guide the development of AI systems in a way that increases user's perceived fairness. Different **Technical Solutions** can be applied to implement privacy and data protection into the AI system. First, *access control* can be established to avoid unauthorized access to data. Second, *on-device processing* can be used so that sensitive data does not have to be transmitted from the system via potentially vulnerable channels. In the case of unavoidable transmission of data, for example over a network, *encryption* can be utilized to protect the data. In order to use as little sensitive data as possible for training of the algorithms, *anonymized data* or *synthetic data* is applicable. Furthermore, the method of *differential privacy* aids in protecting collected data and the privacy of individuals when sharing datasets.

4.1.3 Ethic Guidelines

For an AI system to be fair, it needs to encompass ethical considerations. Therefore, the consideration of ethics is required when developing a trustworthy AI system. Companies in multiple instances give themselves and their AI systems ethic guidelines based on their **Company Values** and the recommendations of a company or public **Ethics Committee** (Robert Bosch GmbH, 2020, p. 1; Volkswagen Group Machine Learning Research Lab, 2020a, p. 17; Villani Mission on artificial intelligence, 2019b, p. 9; A.5 l. 104–106, 109–113). Either way, ethical considerations and decisions need to be codified in the algorithms of an AV to build trust with potential users and the public.

[1] IEEE P7002 – Standard for Data Privacy Process

Non-Discrimination is a notion regularly coming up on providers websites and in expert interviews (BMW AG, 2020; Villani Mission on artificial intelligence, 2019b, p. 8; Continental AG, 2020; A.5 l. 86–88; A.7 l. 108–123). One proposal is that the "[...] IEEE P7003 standard can serve as a baseline to address and eliminate issues of bias in the creation of algorithms."[2] (AI4People, 2020, p. 25). Non-discrimination supports trust-building in that customers have no interest in being at the mercy of a system that puts them at a disadvantage compared to others. In addition, discrimination against others might be perceived negatively by potential users, as it contradicts social and ethical standards in society. In broader terms, ethical considerations should always take into account **Human Rights** (BMW AG, 2020; Robert Bosch GmbH, 2020, p. 1). An AI system must not violate human rights in order to be considered trustworthy by potential users and the public.

Moreover, the concept of **Accessibility** is an important determinant for users trust. Accessibility means that a technology should be usable by all people who wish to. This goes beyond just the AI system and places requirements on the design of AVs as well as the areas of application. AI4People states "[...] AVs should be accessible by design and as inclusive as possible (e.g., disabilities included)" (AI4People, 2020, p. 25). One expert points out that accessibility is not just about *equal access* to an AV, but it may also include equal access to the *safest AV* possible (A.3 l. 134–143). After all, the benefits of better AVs should not be reserved for individuals, but made available to society as a whole. Access to the best available AV technology likely helps increase trust in AVs and ultimately adoption.

The existence of **Socioeconomic Benefits** for potential users and society are considered a fundamental part of building trust in AVs. That is why companies are writing about using AI systems for the good of society (Daimler AG, 2021; BMW AG, 2020). AI4People argues that "[...] providing benefits of increased public health and mobility, better traffic flow and decreased carbon emission" is necessary for automotive companies that want to build trust in their AI systems (AI4People, 2020, p. 28). But the benefits are also on a much more relatable everyday basis, as one expert points out: "[...] some people never [...] open up to AVs until they start getting their pizza delivered in an AV, or they see their friends taking one" (A.6 l. 174–180, 221–223). The analysis shows, the socioeconomic benefits must outweigh the shortcomings of the AV and its AI system to build trust (A.4 l. 450–453, 529–531, A.6 l. 247–252).

Benefits of AVs include not only economic factors but also ecological ones, as different companies and experts suggest (AUDI AG, 2018a; AI4People, 2020, pp. 25, 28; BMW AG, 2020; Robert Bosch GmbH, 2020, p. 1; Continental AG, 2020, p. 3; A.6 l. 300–304). **Ecological Sustainability** of technological solutions has increased in importance as the

[2]IEEE P7003 – Standard for Algorithmic Bias Considerations

topic itself becomes more salient for the global population. AVs therefore should not only be electric vehicles but need to act sustainably. AI4People mentions so-called eco-drive modes as an example, which protect the environment through lower energy consumption (AI4People, 2020, p. 28). Furthermore, the prevention of wildlife accidents by specific algorithms in the AV could pose a contribution to environmental sustainability.

4.1.4 Traffic Flow

The perceived fairness of an AVs' AI system is related to the AVs behavior in the traffic flow. Which is dependent on the driving characteristics and the handling of traffic rules by the AV.

The **Driving Characteristics** are determining how the AV is perceived by passengers and other road users. In an AV, driving behavior no longer depends on the driver but on the vehicle's AI. For example, the AI can be trained in offensive driving behavior, which is then also manifested in the behavior of the AV and might offend other road users (A.4 l. 145–156, 178–184). Offensive driving behavior may lead to negative sentiment toward AVs and lower trust. Therefore, it is proposed to train AVs to drive in a defensive way (A.4 l. 152–156). In addition, the driving characteristics of a vehicle can also influence the user's perception of the vehicle. Driving characteristics are not only an important criterion in the purchase decision for sports cars, but also for AVs, since they have an influence on user's trust in the technology (Jayaraman et al., 2019). For example, if the vehicle drives very cautiously or, as one expert described, lets everyone else pass first at an intersection, such a behavior can damage the user's trust that the system is taking the driver's needs into account (A.3 l. 123–126). In order to adapt the driving behavior to the preferences of each user, manufacturers can think about offering different driving modes in AVs as well (A.4 l. 361–373).

The adherence to **Traffic Rules** by AVs is expected at this point. But as human drivers do break traffic rules sometimes, the strict following of these rules by an AI system can put the AV in a disadvantage, as there might be cases where breaking traffic rules prevent an accident. Some companies virtualize traffic rules to make the AI system's decisions more flexible, accepting that the vehicle violates traffic regulations to a minor extent (A.4 l. 161–175). As described in chapter 4.1.1, compliance with laws and rules is relevant for perceived fairness. Consequently, the handling of traffic rules by the AI system is of particular importance for trust in the system.

4.2 Accountability

Accountability in the context of AVs means taking responsibility for the consequences of the AV's decisions. In order to establish trust in AI systems, especially in the context of AVs, the issue of accountability is particularly important. In the event of errors or failure of the AI in an AV, the consequences can be significantly more severe than in other applications of AI systems. Therefore, it must be possible to identify those responsible for errors and hold them accountable. Since technology itself cannot take responsibility for its behavior, there are two approaches to establishing accountability in AI systems. The accountability of humans and the accountability of organizations for the outcome of AI technology in the context of AVs. In the following, the concepts of human accountability, organizational accountability and redress are explained in more detail.

4.2.1 Human Accountability

German OEMs and Tier-1 suppliers emphasize the importance of human accountability and control over AI systems, especially in autonomous driving (AI4People, 2020, pp. 12, 15; BMW AG, 2020; Robert Bosch GmbH, 2020, pp. 1-2; Continental AG, 2020, pp. 1–2; Daimler AG, 2021). Placing the liability for the outcome of the AI system on humans. Either the user of the AV or the developer. Three methods of accountability regarding an AI system are distinguished by the industry. Not all are equally applicable to AVs, but for the sake of completeness and the varying relevance of the methods for the different levels of autonomous driving, all three methods are explained and put into context.

The first method, **Human-in-command**, means "[...] the AI product is used as a pure tool, where the human constantly decides on the deployment and use of the results, as is the case, for example, when a machine supports the human in classification tasks" (Robert Bosch GmbH, 2020, p. 2). This is the least relevant method of the three for AVs. Due to the need of making decisions in real time, the AI system in AVs cannot be run as a human-in-command system. In fact, all situations in which the driver himself controls the car can be classified as human-in-command situations. In such situations, the driver would be considered accountable. This is also true for assisted driving systems level 2 like lane and distance keeping systems in modern premium brand cars (A.4 l. 277–284, 295–312).

Human-in-the-loop describes AI systems, where a human can influence decisions made by an AI system before or while decisions are executed (Robert Bosch GmbH, 2020). In the case of an AV, one example for this would be the implementation of an override option (Continental AG, 2020; AI4People, 2020, p. 15). In this scenario, the driver is accountable in case he or she overrides the decision of the system, otherwise the system remains accountable for the outcome. In level 3 autonomous driving, the driver's attention

is continuously monitored to ensure that he or she can intervene when the system reaches its limits. Meaning, the human driver regains accountability for the system's actions when he or she overrides the system or the system is handing back control to the user. In systems of higher level automation, on the other hand, the override option has no safety reasons, but gives the user trust in the system, as he can regain control at any time. Continuous attentiveness assessment can then be seen more as a means of ensuring that the use of the override option does not happen accidentally and accountability is always correctly assigned. In cases where the accountability is not with the driver but with the vehicle, human accountability for the AI system is governed by the third method.

The third method **Human-on-the-loop** refers to all AI systems who can not be influenced by a human in runtime, but the decision-relevant parameters are set by humans during development (Robert Bosch GmbH, 2020, p. 2). Developers must limit the operation domain of the AV to the extent that they are confident they can bear the responsibility (A.4 l. 207-213). This describes AVs level 3 to 5 where decisions are made by the AI system without human approval for each concrete decision. Human-on-the-loop is the most important method in the context of AVs, because the nature of AVs is inherently about executing decisions without the interference of humans. Human drivers already make a large number of decisions during travel (40 per two minutes (Network of Employers for Traffic Safety, unknown) or 200 per mile driven (U.S. Department of Labor Occupational Safety and Health Administration, 2006)). AVs, on the other hand, have to make even more decisions, to correctly classify the environment. The classification of the environment usually happens subconsciously for a human driver. Unfortunately, having the driver review every decision made by the AI system would lead to infinitely slow vehicles and at the same time ultimately mean that the vehicle would not drive autonomously, but the steering decision would again rest with the driver. Human-on-the-loop puts the responsibility on the developers of AI systems and makes them responsible for the outcome of their systems. However, since it is not practical to assign full accountability with all legal consequences for the outcomes of AI systems to developers, the legal responsibility lies with organizations developing the AI systems. To that end, the method Human-on-the-loop ultimately leads to the concept of organizational accountability described in the following chapter.

4.2.2 Organizational Accountability

Organizational accountability is based on the concept that companies are responsible for the technology they bring to market (A.4 l. 217–221). Thus, to establish trust in an AV, not only the technical characteristics of the system are relevant, but also the trustworthiness of the organization providing the vehicle itself. Since the focus in this paper is on trust in technology, trust building in organizations is only touched upon.

The **Definition of System Boundaries** is important for companies to establish accountability. Companies can only take responsibility for AVs which are utilized in use cases where the company is sufficiently confident that the vehicle can handle all potential situations. For any situation the vehicle's AI system can not handle, the functionality must be limited by the developing company to counteract irresponsible decisions (A.4 l. 223–225).

One element that can be applied by providers to increase trust in the AVs offered is the **System Performance Monitoring**, which addresses frequently occurring errors (Continental AG, 2020, p. 2; A.7 l. 53–56). One way to assess the performance of deployed AI systems is to ensure traceability by recording the environment and decisions of the AI around particular events (e.g., disengagements of the system) to detect shortcomings of the system (A.7 l. 151–158).

The data and insights gathered can be used in an **Iterative Development Process** which ensures the continues improvement of the AI system. The AI can either be self-learning or be improved by developers themselves. For AI systems in AVs, there is no uniform benchmark for when they are ready to be deployed (A.4 l. 493–503). The lack of a definitive benchmark is based on the generally high number of test kilometers required to validate an autonomous driving system, but also AI-specific challenges such as concept drift [3] (Wachenfeld & Winner, 2016, pp. 439–442; A.7 l. 88–98). Experts from the industry therefore advocate for an iterative development process which enables continues improvement of the AV and gradual expansion of the use cases (A.4 l. 261–268, 493–503, 539–544; A.5 l. 141–145; A.6 l. 134–138, 170–174; A.7 l. 68–84, 124–125, 137–142, 159–160). A learning, constantly improving AI system is perceived by potential users as a better and more trustworthy system (A.4 l. 346-352).

Stakeholder Involvement in the development and deployment process is considered to build trust by experts (A.6 l. 114–122, 142–147, 291–300). Stakeholders can be spokespersons of the communities, affected interest groups, but also regulators and local governments.

For trustworthy companies, this makes **Risk Management** a crucial tool (AI4People, 2020, pp. 25, 31; BMW AG, 2020; Continental AG, 2020, p. 3). The correct definition of system boundaries, the monitoring of deployed AVs and the continuous development of the AI system serve to minimize the risks of the technology.

In connection with monitoring the AI system and the management of risks, the providers mention **Reporting** about the *system performance* and *negative impact* (BMW AG, 2020; AI4People, 2020, pp. 28, 31). In this context, the reporting of negative incidents is con-

[3] Concept drift is a term coined in ML to describe the change over time in unforeseen ways of the distribution of target variables that a model attempts to predict.

sidered particularly important (AI4People, 2020, p. 31). The aim is to avoid potential customers learning about negative incidents through press reports, while vendors do not comment on them. Such a situation would undermine user's trust already build in the technology. Moreover, communicating the performance of the technology in comparison with other means of transport can help rationalize fears and build trust (A.3 l. 199–206; A.4 l. 529–531). By communicating system performance and negative events, companies can claim accountability for their AI system. The associated public commitment of companies to their systems helps to build trust in the long term. In the short term, however, communicating negative incidents also runs the risk of damaging trust.

To ensure the compliance of all technological components and development processes with **Company Guidelines**, Continental AG and an expert point out that all *third-party vendors* need to comply with the "Guidelines For The Ethical Usage Of Artificial Intelligence" (Continental AG, 2020, pp. 1-2; A.5 l. 332–340; A.6 l. 79–82, 84–88). Additionally, Continental AG and Valeo S.A. argue that the *training of staff* on guidelines and processes is important to ensure compliance and therefore improve trustworthiness (Continental AG, 2020; Villani Mission on artificial intelligence, 2019b).

4.2.3 Redress

In both cases, human and organizational control, the question of accountability goes hand in hand with the question of redress (AI4People, 2020, p. 28; A.3 l. 165–166, 189–194). Taking a look at the relationship between the development and deployment of AVs, one can observe that AVs could often be technologically more advanced, but this technological progress is significantly delayed on the roads (KPMG, 2020). Delay is due to various reasons, such as the long average lifetime of vehicles and complex legislation. In especially, one point is the lack of clarity regarding accountability and liability and the associated redress. In both cases, human and organizational accountability, the one that has been identified as responsible would be obliged to redress the damage caused. Potentially facing high claims for compensation discourages manufacturers from taking risks. The resulting lack of accountability of providers is a deterrent for potential users to trust the technology. After all, it does not create trust that one might be responsible for a system whose behavior one does not know, but which can potentially cause great damage. Moreover, it is equally untrustworthy if a manufacturer does not take on this responsibility and the associated risks for redress. The user may ask himself why he should trust the system if even the manufacturer does not. Hence, a legally sound definition of accountability and redress and a technological implementation is important for building trust in AVs. Companies should also set aside reserves to symbolize to potential users their willingness to make amends and have these funds available in the case of an incident (A.4 l. 234–239).

4.3 Transparency

Transparency is described in multiple dimensions by the providers of AV technology. One suggests using the international standard IEEE P7001[4] or SAE J3197[5] for the transparency of autonomous systems (AI4People, 2020, p. 31).

4.3.1 Disclosure of AI Use

The first measure to create trust in AI systems in AVs through transparency is to disclose the use of AI. The disclosure of AI use is proposed by BMW AG and Continental AG to let users know what kind of technology they are interacting with. Knowing the type of technology one is interacting with may be particularly relevant for users who do not understand how AVs and especially AI technology work (BMW AG, 2020; Continental AG, 2020, pp. 3–4). One expert also suggests labeling all systems containing AI with a recognizable badge (A.3 l. 262–270).

4.3.2 System Documentation

Further, providers state that the documentation of the autonomous system should be transparent, to help users understand how the technology works (Continental AG, 2020, p. 3). Especially, the **System Logic** should be documented transparently, including the training data, algorithms and development methods as well as users data application (AI4People, 2020, p. 31; A.3 l. 221–226; A.7 l. 183–187). A transparent system logic helps potential users to gain trust in the AV technology. Either the users read and evaluate the documentation themselves, if they are able to do so. Or independent experts have the possibility to read the transparent documentation of the system logic and provide others with an evaluation. One major challenge remaining is "[...] explaining the various edge cases that could still remain with an algorithm like this [...]" (A.6 l. 242-243).

The **Handover Window** is defined as the time period in which the user of the vehicle must take back the wheel after a level 3 or 4 vehicle has detected its system limits. In addition to the general system logic, the particular handover window should be documented and communicated transparently, with a justification for its length (AI4People, 2020, p. 15; A.4 l. 61–66).

[4]IEEE P7001 – Standard for Transparency of Autonomous Systems
[5]SAE J3197 – Standard for Automated Driving System Data Logger

4.3.3 System Status and Actions

The system should communicate its current status and its actions to the **Passengers** for two reasons. On the one hand, users need transparent information on whether the AI system they want to use is operational, active, or malfunctioning. This information is especially important for systems level 2-4, since the user may have to take control of the vehicle again permanently or in certain situations (A.3 l. 63–66; A.4 l. 128–134). One expert sees it as "[...] one of the biggest tasks to somehow make people aware of this. Especially in the transition phase" to AVs (A.4 l. 124–128, 295–312). On the other hand, even when people have relinquished control, they still want to know many things about the system and its status (A.3 l. 21–24). Moreover, information about the decisions and actions of the AI system can be provided (A.4 l. 321–324; A.7 l. 161–163). This information can help users prepare for the consequences of the decision. For example, they will not be surprised by the vehicle making a sharp turn or initiating an overtaking maneuver, if they are told a few seconds beforehand (A.5 l. 239–243). Being informed about the decision of the AV can increase trust in the technology.

The transparency of the actions of an autonomous system for **Third Parties** is a measure to help build trust in the system not for users themselves, but for other traffic participants. For AI4People an autonomous system, "[...] should clearly communicate about the vehicle's motion intention and awareness of other traffic participants to humans outside the vehicle" (AI4People, 2020, p. 15). The communication can be facilitated, for example, by extending the existing turn signals and brake lights.

In addition to displaying the systems status and the systems actions, displaying the **Detected Environment** can help build trust in the AVs AI system. Even though the system might not yet be capable of actions based on the detected environment, the user gets the feeling that the vehicle sees the same as he or she does (A.4 l. 45–52, 430–435; A.5 l. 213–219). In cases where the system cannot detect certain parts of the environment with sufficient certainty, the object can be represented by a blurred area. Blurred object prevent misidentified objects from undermining user's trust in the vehicle (A.4 l. 430–435). Displaying the detected environment is successfully applied by Tesla, Daimler, and others (A.4 l. 45–52; A.6 l. 205–210).

4.3.4 Human-Machine-Interface

AI systems can create transparency for the user through the vehicle's Human-Machine-Interface (HMI). Today, the human machine interface in a vehicle is usually a combination of (touch) displays, head-up displays, voice control, gesture control as well as light signals and warning sounds (A.3 l. 338–345). In the future, interactions between the user and the vehicle will continue to take place via all these channels and presumably even more.

In an AV, these interfaces can be used to show users the system status and actions, as described in 4.3.3. The transparency helps users better understand and ultimately trust the system (A.5 l. 248–251). Building "intelligible user interfaces" for AVs which increase the transparency of AI systems is a prerequisite for explainability (Villani Mission on artificial intelligence, 2019b, p. 8; A.3 l. 348–349). In the following, possible HMIs are described.

The most common are **Visual Stimuli** used to communicate with users (A.3 l. 226–232; A.4 l. 45–52, 318–321, 328–334, 418–422; A.5 l. 228–238). These can be graphics in displays, for example, as they are also known from modern navigation systems. In addition, the use of simple light pulses is conceivable, such as those used in the VW ID.3 to indicate the direction of travel (Volkswagen AG, 2019).

In addition, the use of **Acoustic Stimuli** to interact with the user is possible, such as those used in parking assistants (A.4 l. 418–422).

Less frequently used today are **Haptic Stimuli**. Up until now, these have only been used as input feedback for touch displays (AUDI AG, 2018b). In the future, however, these could also be used to alert users to the takeover of control in level 3 systems by means of vibration (A.4 l. 418–422). Moreover, these stimuli can be used to communicate decisions made by the AI system. For example, a short vibration of the seat before the vehicle decelerates sharply.

Combining the possible stimuli with the detected environment, **Augmented Reality** technologies can be used to create transparency about the actions of the AI system (A.4 l. 318–322, 328-334). To help a user understand why the vehicle is braking, the AI system can, for example, mark the target object in augmented reality and provide metrics related to the decision (A.4 l. 324–326; A.7 l. 165–169).

The **Personalization** of such HMI systems can be another building block in establishing trust. Trustworthiness of an AI system depends on how transparent it is. However, more transparency does not necessarily lead to higher trust. Providing too much information might be perceived as not rational by users or highlight the limitations of the AI system, which in turn damages trust (A.3 l. 91–98; A.4 l. 425–430; A.3 l. 312–319). Moreover, each user group may have a different level of demand for transparency and explainability (A.4 l. 391–394, 418–422; A.5 l. 213–219). For some people, knowing that the system works is enough; others may want detailed explanations of how the system arrives at its decisions. In addition, the need for information is likely to change with the length of time the AV is used. At the beginning, more information is required, which is later no longer relevant for building trust (A.4 l. 398–404). Providers should therefore offer options for individualization so as not to overwhelm any user with information, but to be able to offer sufficient information when needed (A.5 l. 210–219).

4.3.5 Explainability

The explainability of algorithms and related behavior of an AV is an important part in companies efforts to build trust in AV technology (A.7 l. 56–65). Even though one expert argues that potential users are not interested in the explainability of systems, as long as the AV works in the expected manner, the topic is perceived important by most experts (A.6 l. 216–221; A.7 l. 29–32). Companies distinguish multiple ways of making their systems more explainable or helping users to understand the behavior of an AV.

One way to enable explainability is the usage of **Explainable Models** (Villani Mission on artificial intelligence, 2019b, p. 8). Explainable models help developers to identify errors in the development process and to find better solutions. Commonly used methods are surrogate functions and heatmaps for Convolutional Neural Networks (A.5 l. 180-185). Developers can use these tools to better understand algorithms themselves, but also to build functions that improve explainability to users.

Predictable Behavior of the AV is particularly relevant for several reasons. On the one hand, it is important that the vehicle behaves predictably so that other road users do not feel endangered. Accelerating in front of a crosswalk so that the vehicle can pass before the pedestrian may be efficient to avoid congestion, but would trigger negative emotions and undermine trust. Ford Motor Company argues, that "[...] people tend to gain trust in something when they can predict what it will do" (Ford Motor Company, n.d., p. 19). This is not only true for other road users interacting with an AV but also for passengers of the vehicle. Occupants also do not want the vehicle to behave in an unusual or inconsistent way, as this would mean that the vehicle behaves differently than they think is appropriate (A.4 l. 337–339; A.5 l. 52–54; A.7 l. 44–46, 142–144). Different behavior than anticipated may be justified in individual cases, but it can lead to users not trusting the system if it happens on a regular basis. Additionally, unpredictable behavior of the vehicle more likely leads to motion sickness resulting in unpleasant driving experiences which disrupts the formation of trust in the AV. Therefore, the predictability of the vehicle's behavior appears to be a key element in making AVs more explainable to strengthen trust.

User Training is another way to increase the explainability suggested by AI4People. Either in *Tailored Trainings* for different demographic groups with curriculums focussed on explaining the functioning and limitations of the technology or through a *Training Mode* implemented in the AV that trains users on the AI system (AI4People, 2020, p. 15; A.4 l. 328–334, 404–411; A.6 l. 286-290). In addition to specific training, a thoughtful deployment strategy can achieve user training. For example, the general functioning of an AV can already be tested and learned by the user in a level 2 system in rudimentary fashion. Subsequently, a knowledge and understanding advantage is available when using a level 3 system (A.3 l. 105–116; A.4 l. 39–45, 134–141, 443–447).

Post-hoc Explainability describes the explainability of decisions of the AI system after they have been made (post-hoc). Post-hoc explainability is of particular relevance for improving the algorithms, resolving errors and ensuring accountability. By explaining to users on what basis and why a decision was made, trust can be build (A.5 l. 40–44, 58–61; A.7 l. 36–43). One prerequisite for post-hoc explainability is to have a *Logging System* that keeps records of the decisions made by the AI system (Continental AG, 2020, p. 2; AI4People, 2020, p. 31; A.4 l. 221–223; A.7 l. 53–56, 151–158). The logging system can be build according to SAE J3197, the standard for automated driving system data loggers, and should keep record of the *Considered Parameters* and make them transparent (Daimler AG, 2021; A.3 l. 293–301; A.4 l. 383–388; A.5 l. 176–178; A.7 l. 161–169; SAE International, 2020). The parameters taken into account by the algorithm for decision-making can help explain why an algorithm acted as observed. Parameters like other traffic participants, their speed, and behavior create a virtual version of the *Considered Environment* that was sensed by the AV and the underlying algorithms (A.4 l. 383–388; A.5 l. 263–268). Information on these parameters, considered by the AI system for driving-related decisions, can increase the perceived trustworthiness of AVs. Furthermore, the *Considered Algorithms* are relevant since they give the considered parameters meaning (A.5 l. 178–180, 260–263, 268–271).

To improve the experience and explainability for users, a *Feedback Loop* can help. Users can provide feedback to the system in situations that have felt unintuitive to them or where they can not understand the vehicle's decision (A.3 l. 251–256, A.4 l. 257–259, 263–268, 352–361). Then the system itself or the developers can address this and improve the explainability of the AI system.

4.4 Safety

The safety of AVs is the most mentioned attribute of TAI to increase trust in the technology (A.4 l. 346–352; A.5 l. 28–30, 36–37; A.6 l. 46–48, 67–69). Building AI systems which are safe to use is the goal of all providers. To achieve safe AVs, providers propose technical and procedural measures to ensure software and hardware are capable of delivering the required safety. But in the first place, suppliers guide the development of AVs with the overarching principle of minimizing harm.

4.4.1 Minimization of Harm

Providers set themselves the goal of ensuring that their system causes as little harm as possible (AI4People, 2020, p. 25; Continental AG, 2020, p. 3). If harm is unavoidable, it should be as small as possible. The extent of the damage can be minimized, for

example, by analyzing the speed of road users or the potentially resulting collision angles (AI4People, 2020, p. 25). If providers develop their systems in such a way, they could be perceived as particularly safe and trustworthy by potential users.

4.4.2 Adequacy

One measure to ensure safe AVs is to develop AI systems that are suitable for the use case of autonomous driving (Continental AG, 2020, p. 3; AI4People, 2020, p. 19). One example of the importance of the adequacy of algorithms is the recognition of different animals for emergency braking assistants. For example, kangaroos in Australia could not be reliably detected with the algorithm for the detection of moose (A.4 l. 464–469).

In more detail, that means an AI system needs to be adapted to different environments via **Localization** (Continental AG, 2020, p. 3; Ford Motor Company, n.d., p. 19). Each geographical or cultural region may have different requirements for the AI system (A.4 l. 184–191, 367–373). Be it through different traffic rules, different driving behavior or cultural differences that influence the systems performance and ultimately users trust and acceptance of such a technology.

Moreover, the **Business Case** an AV is deployed in might affect the adequacy of the AI system and its configuration (Continental AG, 2020, p. 3). The requirements placed on the system by different situations can be tremendous. An autonomous shuttle which transports passengers between the aircraft and the terminal at an airport must meet different requirements than a shuttle in inner-city traffic or an autonomous fuel truck on the airfield. The AI system of an autonomous shuttle on the airfield must recognize aircraft correctly and behave accordingly. In addition, it must follow different traffic rules than a shuttle in inner-city traffic. But its driving behavior also needs to be different from that of a fuel truck on the airfield. As passengers should be provided with a smooth experience, the fuel truck might not be required to break smoothly and can take sharper turns, as there are no passengers negatively affected by such driving. The business case is therefore significant for developing and deploying a trustworthy AI system.

A trustworthy AV is only ever trustworthy in its **Operation Domain**. By restricting the operation domain to use cases in which the vehicle safely fulfills all requirements, it is possible to generate trust among users (A.4 l. 69–78). The operation domain can then be extended step by step with every evolution of the system or reduced with every negative incident that occurs (A.4 l. 134–141, 207–213).

4.4.3 Reliability & Accuracy

The reliability of an AV is of paramount importance for user trust. That is why this topic takes up a lot of space in the manufacturers' publications (AI4People, 2020, p. 16, 19; Robert Bosch GmbH, 2020, p. 1; Daimler AG, 2021; Ford Motor Company, n.d., p. 16). Daimler AG even states, "Safety and reliability are quintessential parts of our brands. Therefore, we apply our high standards of quality to AI applications as well. [...] In AVs, reliable AI is [...] safety relevant" (Daimler AG, 2021). Accidents of AVs that are not prevented by the AI system make them appear unreliable. As the example of the Uber incident involving the death of Elaine Herzberg shows (BBC, 2019). After such events, the public's perception of the technology worsens (Penmetsa, Sheinidashtegol, Musaev, Adanu, & Hudnall, 2021). Worsened public perception can include decreasing trust in the AI system of the AV. To avoid such incidents and build trust in their AI technology from the get go, providers of these systems try to build highly reliable and accurate products. Accuracy refers to how precisely the AI system in the AV is able to detect situations. In more technical terms, the accuracy of an algorithm is a performance metric that shows what percentage of cases an algorithm predicts correctly. This number alone is not sufficient to conclusively evaluate an algorithm, but it does give an indication of how well the data used to train the algorithm represent reality (A.5 l. 189–192; A.7 l. 43–44).

To increase the accuracy of algorithms, it is important to train them with a **Representative Dataset**. The training data must cover the reality in which the vehicle is used as exactly as possible (A.4 l. 78–88, 97–98, 259; A.5 l. 75–84; A.7 l. 134–137). That includes local differences and the adaption to changes in the environment (A.4 l. 191–196; A.7 l. 46–50, 98–101). Representative training data needs to be collected in the field and possibly enriched with known edge cases which were not part of the collected data. Keeping the training data representative can be done through the iterative development process, described in Section 4.2.2, where models are constantly retrained to account for changes in the environment, like new means of transportation (A.7 l. 88–101). Dr. Ing. h.c. F. Porsche AG (2019) is boiling it down to "Without representative training data, there will be no representative output."

Moreover, **Edge Case Detection** is mentioned a lot, as everybody in the industry fears to fail in detecting and handling such an edge case. Edge cases are situations an AV encounters that it is not trained to handle, because they were not observed before. Edge cases will always exist, the question is how vehicles deal with them and whether they are designed to continuously learn from such cases (A.4 l. 259–261, 539–544; A.6 l. 126–131). Continuously improving the system's performance in detecting edge cases makes the systems safer and creates trust (A.7 l. 46–52, 179–183).

(Durability) Testing is one method to prove the system's performance and ensure

reliability of an AV and its AI system. Today, vehicles are tested according to standardized procedures like ISO 15037[6] or ISO 26262[7] during the development process (International Organization for Standardization, 2019; International Organization for Standardization, 2018). First, components and functional groups are checked for reliability and compliance with requirements. Subsequently, the entire vehicle is tested with regard to a large number of criteria. One of it being durability or fatigue testing. AVs must also go through this testing process, but the procedure must be extended to include testing of the AI system for reliability (AI4People, 2020, p. 19; Daimler AG, 2021; Ford Motor Company, n.d., p. 16; Continental AG, 2020, p. 3; A.4 l. 521–529; A.5 l. 145–148). Testing an AI system means proving its adherence to predefined standards to ensure that AVs are reliable and detect edge cases.

Moreover, companies use **Simulations** which run the AI system of the AV by real world data to evaluate the performance of the system (A.5 l. 160–162). The simulations are a great help in validating the systems, since fewer functional prototypes need to be built in order to drive the large number of kilometers needed to validate AVs. In simulations, extremely rare cases such as wrong-way drivers can be simulated and trained (A.4 l. 88–93; A.5 l. 162–165).

To ensure safety, companies should perform scientific validation methods to put the algorithms used to the test and ensure they are statistically validated to be safe to use (Daimler AG, 2021; A.4 l. 55–61, 73–78). These new validation and test procedures, required to ensure the safety of AVs, vary significantly from historical methods of testing and simulation (A.4 l. 521–529; Wachenfeld & Winner, 2016, pp. 428–447; Aptiv Services US, LLC et al., 2019, pp. 72–97).

4.4.4 Robustness

The robustness of an AI system is closely related to its reliability, but opens an additional dimension of safety – A dimension about the resilience and consistency of such system (A.7 l. 43–44; 125–130).

The first thing that contributes to the robustness of AI systems is the **Failure Behavior**. The failure behavior describes how an AI system behaves in the event of an error and what fall back plans are in place (AI4People, 2020, p. 16). The first method of protection is *redundancy* of technical components that are relevant for the functioning of the AV (A.4 l. 233–234, 250–254; A.5 l. 140–141). The second method is to have *fallback options* in place to ensure that the vehicle remains in a safe state (A.3 l. 162–165; A.4 l. 61–66). In the event of a detected error, for example, it could begin signaling its current state to

[6]ISO 15037 — Road vehicles — Vehicle dynamics test methods
[7]ISO 26262 — Road vehicles — Functional safety

other road users and come to an appropriate stop at the side of the road. In addition to the actual behavior in case of an error, the communication of errors plays an important role in building trust in the system. Appropriate messaging is needed to not scare users and undermine trust by pushing incomprehensible error messages to the HMI.

Another part of robustness is the **Resilience to Attack**. In the case of AVs, attacks feared by experts and the public are typically cyberattacks which aim at influencing the behavior of the vehicle. Companies propose to build systems compliant with cybersecurity standards as the SAE J3061[8] (AI4People, 2020, p. 16; Continental AG, 2020, p. 2; SAE International, 2016). In addition, the system must be able to handle deliberately manipulated traffic signs in order to be robust (A.7 l. 179–183).

Continental AG brings up the principle of **Safeguarding** in their guidelines (Continental AG, 2020, p. 3). Meaning that providers should "[...] ensure that the AI has safeguards against any uncontrolled behavior (especially for fully or semi-automated physical robots) that may cause harm." (Continental AG, 2020, p. 3). Safeguarding in this case means, that the AV and its AI system should be designed in a way that does not violate its boundaries under any circumstances. For the AI, this implies that it must adhere to basic rules, even if it is allowed to determine its own behavior within the limits. The need for such safeguards can best be demonstrated by a thought experiment. If the AI of an AV is designed to always take the fastest route to a destination, then it could decide, without generally applicable restrictions, that swerving the vehicle onto the footpath at high speed would lead to the destination faster than waiting at the traffic lights. Such behavior could be prevented, for example, by rules that prohibit the AI system from using the sidewalk under any circumstances. Another example would be that the vehicle is first at a red pedestrian light and predicts a rear-end collision. To avoid this, the vehicle accelerates through the light and runs over pedestrians. Such a scenario could be prevented, for example, by a rule that forces the AI system to brake for human obstacles under all circumstances. Obviously, these are contrived examples, but they show why safeguarding as a last resort can be helpful in limiting the decision-making power of AI. To avoid uncontrolled behavior and safeguard against it can foster trust in the technology.

4.5 Certifiability

The attribute certifiability describes the need to design an AI system in such a way that it can successfully pass an evaluation. A successful certification signals trustworthiness to potential users, as has already been shown in other contexts (Renner, Lins, & Sunyaev, 2021, p. 1; Sunyaev & Schneider, 2013, pp. 33–34; Özpolat, Gao, Jank, & Viswanathan, 2013, pp. 1100–1111; Lins, Grochol, Schneider, & Sunyaev, 2016, p. 68). The certification

[8]SAE J3061 – Cybersecurity Guidebook for Cyber-Physical Vehicle Systems

should take place in all the attributes mentioned above, i.e. in the areas of safety, fairness, accountability, transparency and explainability (A.4 l. 477-483; A.6 l. 223–226, 231–233).

4.5.1 Official Standards

A prerequisite for certifiability and ultimately certification is the existence of official standards that the AI system and the organization developing such a system has to comply with. These standards should be defined by industry organizations or governmental institutions to establish a standard accepted industry-wide. AI4People proposes IEEE P7009[9] as baseline standard for autonomous systems, where specific guidelines can add on (AI4People, 2020, p. 19). An overview of official standards for AVs is given by Omeiza et al. (2021, p. 4).

4.5.2 Auditability

Official standards alone are not sufficient for certification. Typically, **internal and external audits** are conducted to ensure compliance with standards. Some companies and expert groups advocate for the certification by trusted third parties using external audits supplemented by internal audits (Volkswagen Group Machine Learning Research Lab, 2020a, p. 17; Villani Mission on artificial intelligence, 2019b, p. 8; AI4People, 2020, p. 31; A.5 l. 51–52). These audits can be part of the homologation process for the vehicle and should either include extensive testing of the AI system or rely on testing results during the development of the vehicle.

[9]IEEE P7009 – Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems

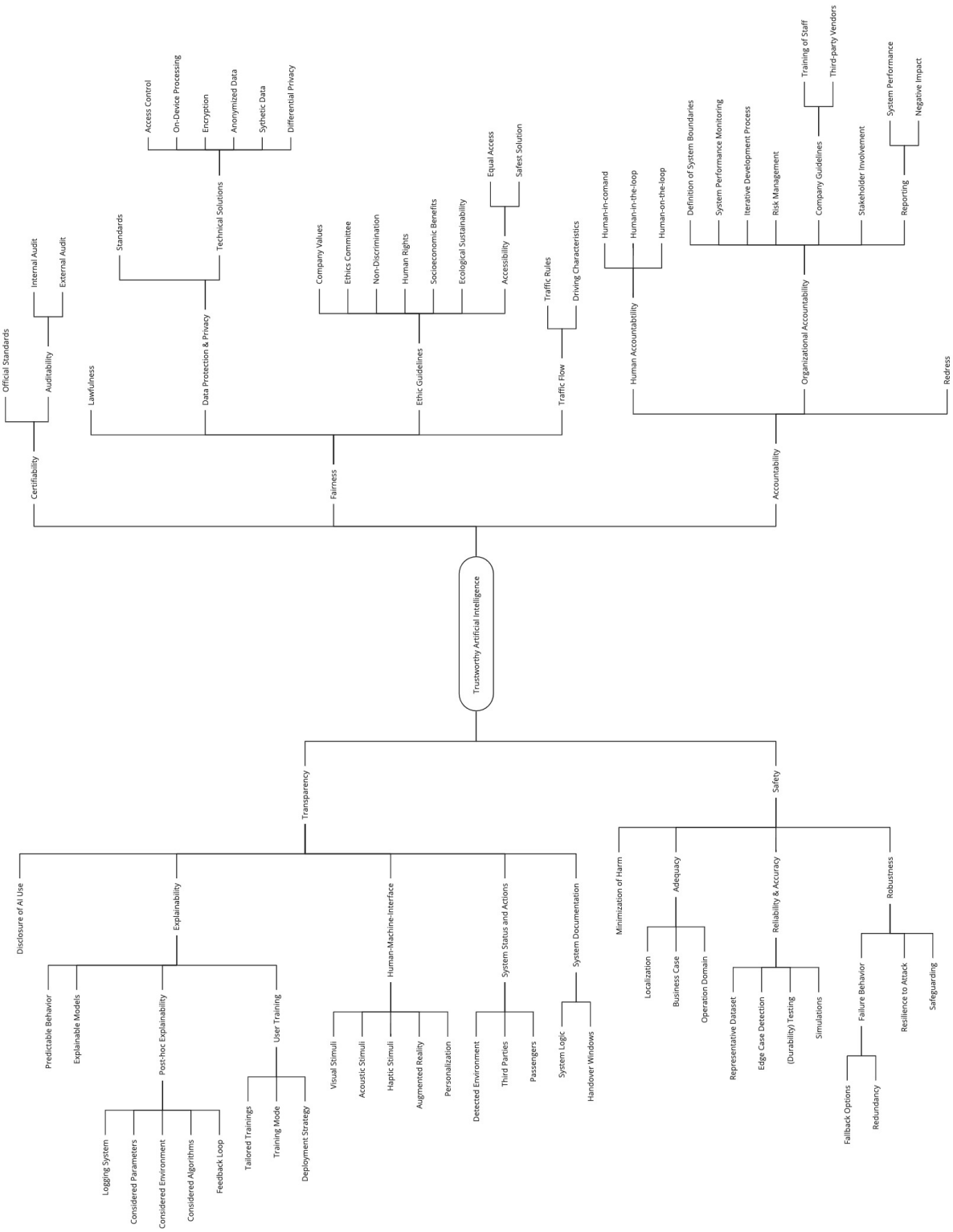

Figure 2: Framework for Trustworthy AI in Autonomous Vehicles

5 Discussion

5.1 Discussion of Principal Findings

The findings suggest that TAI in the context of autonomous driving cannot be generated only by the FATE characteristics described in literature. In addition, important characteristics in the context of AVs identified are certifiability and safety. Furthermore, explainability plays a subordinate role to transparency in the findings. This thesis rather shows that Certifiability, Fairness, Accountability, Transparency, Safety (C-FATS), form a framework for TAI in AVs. The importance of certifiability in the context of AVs is due to the complexity of AI systems of an AV. Not all C-FATS characteristics can be observed, understood, and verified by non-experts to make a decision about the trustworthiness of the technology. Therefore, certification by a trusted third-party can signal the trustworthiness of a technology to potential users. The other characteristic that is particularly significant in the context of AVs is the safety of AI systems. The importance of safety can be explained mainly by the severity of the potential consequences of AI systems failures. While errors in a recommendation engine are unlikely to have serious consequences, an AV error can lead to a potentially fatal accident. Users therefore expect a safe and reliable system as a basis for the development of trust.

The results show that it is difficult to assign individual FATE attributes to one of the two concepts of "trust in technology" and "trust in organizations" due to an organization's accountability for AI technology. For example, identified attributes such as an iterative development process create trust in both the organization and the technology itself. Assignment to one of the two trust concepts mentioned thus depends on the standards applied. Moreover, some FATE attributes, such as organizational accountability, suggest that it is not only the characteristics of the technology itself that are important for building trust. Rather, the attributes and actions of the providing organization are also important for building trust in AVs.

5.2 Comparative Analysis of Providers' and Experts' Perspectives

Generally speaking, the results of the content analysis and the interviews complement each other. The perspectives of providers and experts on TAI in AVs differ only on some of the described attributes. In the following, major identified differences are described.

Overall, technological issues are more prominent in the interviews than on the providers websites, except for technological solutions for data protection and privacy. In turn, ethical topics are not as heavily mentioned in the interviews as on the providers websites.

The difference in focus is presumably due to the technical background of the experts and the purpose of the publications on the providers' websites. The providers' websites are marketing tools that serve to build up a reputation in certain technology fields. The providers therefore design the content to be as comprehensible and general as possible so that a large target group receives a positive image of the company. In the context of AI systems, a positive image in society is particularly important after scandals involving Facebook and others surrounding data protection (NBC News, 2018).

The first difference identified relates to the perspective on fairness. Based on the manufacturer websites, "ethics" is identified as an attribute to be assigned to "fairness". From the experts' point of view, ethics alone do not improve user's trust, but the definition of public-facing ethic guidelines for an AI system do. These guidelines encompass all ethical considerations, and therefore all elements identified to be encompassed in the attribute "ethics" are subordinate to the attribute "ethic guidelines" in the derived framework.

The results of the content analysis show that organizational accountability is enabled by the training of staff and third-party vendor accountability, among other things described in Chapter 4.2.2. Due to the high degree of division of labor and complexity of technologies and processes, the collaboration with suppliers and employees is of great importance in the automotive industry. The accountability of third-party vendors, as well as the successful training of staff, must be ensured to credibly take accountability for AI technology as a company. The results of the expert interviews suggest that it is useful to define corporate policies so that employee training can be conducted in a standardized manner and third-party vendors know what standards they must adhere to (A.5 l. 332–340; A.6 l. 79–82, 84–92). Therefore, the attributes "training of staff" and "third-party vendors" are subordinate to the attribute "company guidelines" in the framework.

The major difference in perspectives identified is the connection of explainability and transparency. In every interview, experts exhibit difficulties in differentiating between transparency and explainability. Practitioners – in line with AI4People – see transparency as an underlying fundamental of trustworthiness (A.5 l. 256-260). In two interviews, the close relationship between explainability and transparency is even mentioned explicitly (A.3 l. 263–269; A.4 l. 473–486). Moreover, experts tend to view explainability as less important. One train of thought, unique to the interviews, is that users only need a functioning vehicle and positive experiences with the technology to adopt it (A.6 l. 231–233). This train of thought suggests that transparency and explainability are not that relevant in building trust as long as the vehicle is safe to drive and the benefits of AVs can be observed (A.6 l. 216–221). One approach to explaining this view can be taken by comparing AVs with the rapid spread of other modern technologies. For example, the spread of smartphones is supported by the virtual absence of accidents and a positive benefit for users, as well. Moreover, Koul and Eydgahi (2018) were able to show that

perceived usefulness has the strongest influence on the intention to use an AV, in the TAM model. The findings by Koul and Eydgahi support the theory that the benefit of an AV to users is of paramount importance to adoption. However, the impact of trust in AI technology on the intention to use AVs has not been studied. Moreover, it remains unclear whether users will accept little explainability and transparency for AVs, as they do for other technologies, when flaws in the technology are potentially lethal. Because of the uncertainty described above and the need for transparency and explainability to ensure accountability and certifiability, transparency and explainability are important components for establishing trust in AI systems of AVs. Although explainability and transparency are coded on the same hierarchical level in the content analysis, the results of the interviews lead to the conclusion that explainability is subordinated to transparency in the context of AVs. Therefore, explainability is assigned as an attribute of transparency in the framework.

Table 5 and 6 summarize the results of the content analysis and interviews. A heat map shows which parts of the framework are covered by which provider and where experts see the main focus in establishing TAI in AVs. In the following, major identified differences are discussed.

In the first step of the content analysis, it was noticed that especially US providers do not write about principles and approaches to achieve TAI on their website. US providers often publish safety and disengagement[10] statistics to prove their vehicles are safe, but do not explicitly relate this to building trust in AI systems in the vehicle. The different approach to communication on TAI in AVs between U.S. and European firms can be illustrated by comparing columns (B) BMW and (F) Ford in Table 5 and 6. Ford's published information focuses strongly on safety, while BMW provides significantly more information on fairness and accountability. The approach of communicating the safety of AVs can be linked to the statement on building trust via benefits and safety discussed in the previous paragraph.

Even though the French suppliers Valeo and Renault set themselves the same standards through Villani Mission on artificial intelligence, for the most part providers give themselves different guidelines. The heterogeneity of approaches to establishing trust communicated by providers suggests that there is no consensus on how to establish trust in AVs. Furthermore, there are no agreed upon standards as of today that guide providers in their effort, which they feel comfortable sharing. There may be several reasons for manufacturers' reluctance to share information. Most likely, they are not interested in fully disclosing their approach to building trust with competitors at this time. It is well-known that corporations selectively reveal their knowledge to avoid falling behind their

[10]Disengagement occurs when a safety driver takes control of an AV during testing because the AV may be behaving unsafely or has stopped functioning.

competitors (Alexy, George, & Salter, 2013). In addition, TAI is an evolving topic in academia that has not yet made many inroads into the day-to-day work of AV providers, as this thesis shows.

Even though accessibility of AV technology is mentioned more often, equal access to the safest AV technology is only mentioned by one expert. Whether the safest AV technology must be accessible to everyone is an intriguing ethical question, since even today, not every vehicle meets the same safety standards. Either because of the age of the vehicle and the technological advances that have taken place since it was manufactured, or because of the different levels of safety created by design, materials, and layout. The importance of AI algorithms to the safety of an AV can enable safety improvements even during the lifecycle of the vehicle, and theoretically across brands. Offering particularly safe vehicles that go beyond the required safety standards may also be a way to differentiate from the competition in the future and drive safety innovations that are subsequently adopted by the whole industry. According to the results of this work, it seems clear that a safe AV creates trust. However, it is not evident whether the equal access to the safest AI technology for AVs provides an additional benefit to users' trust in the technology.

The attribute "traffic flow" is only mentioned in the expert interviews, but not on the websites of providers. Configuring the AI in an AV in a way that its behavior does not offend other road users and passengers, does not seem to be an issue providers want to communicate. Providers may not want to imply that AVs can theoretically also be designed to push traffic laws or exhibit a more offensive driving style. Therefore, they do not communicate information about the importance of traffic flow of AVs. This observation is given further emphasis by the interviewed experts (A.3 l. 123–131; A.4 l. 145–156, 161–175, 178–184, 361–373).

Reporting the system performance is an attribute only mentioned by experts as a possible trust-building measure. The reluctance to report the system's performance can also be explained by possible negative consequences for providers. If, for example, a provider communicates the performance of its systems, this can have a negative impact on trust compared to the performance of other providers. The effect of communicating system performance and negative impact on TAI needs to be better understood to be a relevant mean for providers to build trust.

Table 5 and 6 report that Audi in particular provides detailed information on TAI in AVs and thus fulfills most of the attributes of the developed framework. This can be attributed to the nature of the source. Audi communicates that it supports the AI4People initiative and is guided by the proposed recommendations in its development (AUDI AG, 2018a). The AI4People's 7 AI Global Frameworks contain recommendations for the automotive industry to ensure ethical AI solutions, and therefore covers many aspects which help build users trust in AVs (AI4People, 2020, p. 3). The joint development of such guidelines

by researchers and practitioners can help foster consensus among providers of AVs on how to establish trust in their AI technology.

Framework Part 1				(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(1)	(2)	(3)	(4)	(5)
Certifiability				■	■					■		■	■	■	■		■	
	Official Standards			■						■		■	■					
	Auditability			■														
		Internal Audit		■						■		■	■					
		External Audit		■	■	■		■			■		■	■	■	■	■	■
Fairness					■	■		■				■	■	■	■	■	■	
	Lawfulness			■	■			■		■	■	■		■				■
	Data Protection & Privacy			■	■			■		■		■						
		Standards		■				■		■	■	■						
		Technical Solutions		■			■			■	■	■						
			Access Control	■														
			On-Device Processing				■	■										
			Encryption				■				■							
			Anonymized Data					■										
			Synthetic Data								■							
			Differential Privacy								■							
	Ethic Guidelines			■	■	■		■		■		■	■		■	■		■
		Company Values																
		Ethics Committee								■		■						
		Non-Discrimination		■	■			■		■		■				■	■	■
		Human Rights			■	■												
		Socioeconomic Benefits		■	■	■		■		■	■	■			■		■	
		Ecological Sustainability			■	■											■	
		Accessibility		■										■		■	■	
			Equal Access											■		■	■	
			Safest Solution											■				
	Traffic Flow													■	■			
		Traffic Rules												■	■			
		Driving Characteristics												■	■			
Accountability				■	■	■	■	■		■		■	■	■	■	■	■	■
				■	■	■	■	■							■	■	■	
	Human Accountability					■												
		Human-in-comand				■	■								■			
		Human-in-the-loop		■		■												
		Human-on-the-loop																
	Organizational Accountability			■	■					■		■	■	■	■			■
		Definition of System Boundaries													■			
		System Performance Monitoring					■											■
		Iterative Development Process													■	■	■	■
		Risk Management		■	■		■						■	■	■			
		Company Guidelines					■									■	■	
			Training of Staff				■			■		■						
			Third-party vendors				■									■		
		Stakeholder Involvement					■										■	
		Reporting		■	■									■	■			
			System Performance											■	■			
			Negative Impact	■														
	Redress			■										■	■			

Table 5: Overview of the coverage of the framework by the individual providers and interviews – Part 1
(A) AUDI AG, (B) BMW AG, (C) Robert Bosch GmbH, (D) Continental AG, (E) Daimler AG, (F) Ford Motor Company, (G) Renault S.A., (H) Dr. Ing. h.c. F. Porsche AG, (I) Valeo S.A., (J) Volkswagen AG, (1) Interview 1, (2) Interview 2, (3) Interview 3, (4) Interview 4, (5) Interview 5

Framework Part 2				(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(1)	(2)	(3)	(4)	(5)
Transparency				■	■	■	■	■	■	■	■	■		■	■	■	■	■
	Disclosure of AI Use				■		■							■				
	Explainability			■		■		■	■	■	■	■		■	■	■	■	■
		Predictable Behavior							■						■	■		■
		Explainable Models								■		■				■		
		Post-hoc Explainability		■			■	■						■				■
			Logging System												■			
			Considered Environment													■		
			Considered Parameters					■						■	■	■		✕
			Considered Algorithms													■		
			Feedback Loop											■	■			
		User Training		■													■	
			Tailored Trainings	■														
			Training Mode	■														
			Deployment Strategy											■	■			
	Human-Machine-Interface									■		■		■	■	■	■	
		Visual Stimuli												■				
		Acoustic Stimuli													■			
		Haptic Stimuli													■			
		Augmented Reality													■			
		Personalization												■	■	■		
	System Status and Actions			■										■	■	■	■	■
		Detected Environment													■			
		Third Parties		■														
		Passengers												■	■	■		■
	System Documentation			■			■							■	■	■	■	■
		Handover Windows		■											■			
		System Logic		■										■		■	■	■
Safety				■	■	■	■	■	■		■		■	■				
	Minimization of Harm			■	■		■											
	Adequacy			■			■		■						■			
		Localization		■			■		■						■			
		Business Case					■								■			
		Operation Domain													■			
	Reliabiliy & Accuracy			■		■	■	■	■		■			■	■	■		■
		Representative Dataset		■			■	■			■				■		■	
		Edge Case Detection													■		■	■
		(Durability) Testing					■	■	■						■		■	
		Simulations													■		■	
	Robustness			■	■	■	■						■	■	■	■		■
		Failure Behavior		■										■	■			
			Fallback Options	■										■				
			Redundancy												■	■		
		Resillience to Attack		■			■											■
		Safeguarding																

Table 6: Overview of the coverage of the framework by the individual providers and interviews – Part 2
(A) AUDI AG, (B) BMW AG, (C) Robert Bosch GmbH, (D) Continental AG, (E) Daimler AG, (F) Ford Motor Company, (G) Renault S.A., (H) Dr. Ing. h.c. F. Porsche AG, (I) Valeo S.A., (J) Volkswagen AG, (1) Interview 1, (2) Interview 2, (3) Interview 3, (4) Interview 4, (5) Interview 5

5.3 Interdepending Characteristics of Trustworthy Artificial Intelligence

There are interdependencies between the individual attributes that establish TAI in AVs identified in Chapter 4. The interdependencies occur since properties of a technology can influence different attributes for trust building. Furthermore, the effect of a property can have different directions and different magnitudes on user trust. Resulting in interdependencies that may reinforce or undermine the effect of specific C-FATS attributes in establishing TAI. Hence, it is important for providers to address these interdependencies. The interdependencies discussed in this chapter are marked in Figure 3 by dashed connections.

One interdependency identified regards changes in laws. One expert describes how legal changes affect all other attributes (A.5 l. 289–298). In particular, changes in official standards as well as traffic rules affect the lawfulness of a system. As traffic rules are typically national laws that differ between states, localization of AI systems, to ensure adequacy of algorithms, is related to lawfulness as well. Lawfulness is also connected to standards for data protection, such as GDPR, because these standards are often defined by laws. Data protection standards are also part of the official standards used for certifying an AI system. Because organizational accountability demands of an organization to ensure an AI system is lawful and complies with all official standards, changes in laws also affect the organizational accountability. All of the dependencies described are important to consider, since law changes might have consequences for the technological setup of the AV. In changing the technological setup, the AV might be perceived differently by users. Tightening laws may also have a direct impact on user trust, as laws may impose higher safety standards or greater transparency of AI, both of which are considered to build trust. The European Artificial Intelligence Act COM/2021/206 of the European Commission (EAIA) and Data Governance Act COM/2020/767 of the European Commission (DGA) proposed by the Commission in 2020 and 2021 will, once enacted, likely be interesting case studies to observe the implications of legal changes regarding TAI in AVs (Commission, 2021a, Commission, 2021b).

Another interdependency connects non-discrimination and safety. In cases where people are not detected properly due to a discriminatory system, their involvement in an accident with an AV is more likely. Because unintentional discrimination can be contained by training the AI with representative data, non-discrimination and representative datasets are linked. This link describes why "[...] an AI is always as fair as the data with which I feed it" (A.4 l. 178–184). Moreover, non-discrimination is closely related to accessibility, because if certain groups of people do not have access to AVs, they are discriminated compared to other groups. This leads to the conclusion that withholding equal access to

AV technology negatively impacts user trust in two ways. Directly by withholding equal access and indirectly by discriminating certain communities. Although one could argue, that accessibility and equal access should be subordinate to non-discrimination in the framework. More generally, fairness is related to accountability, in that the company is responsible for fair and just AV behavior. Lastly, the concept relates to safety, since a safe system must deal with ethical dilemmas that may arise (A.3 l. 263–269; A.6 l. 271–280).

A strong dependency is observed between transparency and explainability, because transparency can be seen as a prerequisite of explainability (A.3 l. 263–269; A.4 l. 473–486). Without a transparent system, no explanation of the decisions by algorithms and the behavior of the whole AV is possible. Therefore, explainability is considered part of transparency in the framework as described in Chapter 4. Moreover, transparency can be seen as an enabler of all other discussed C-FATS characteristics. If companies are not transparent about the properties of their systems, potential users will have difficulties getting a reasonable perception of the technology. AI4People puts it as follows: "In the automotive sector, we contend that transparency is not a freestanding desideratum, but rather a key mechanism to realize the other principles or requirements." (AI4People, 2020, p. 22). Despite the great overarching importance of transparency, the concept is not superior to the other characteristics at the highest level of the C-FATS framework, as certifiability, fairness, accountability, and safety are concepts in their own right. However, particularly significant, is the connection of transparency to auditability. Transparency can be viewed as a prerequisite for auditability, as there needs to be transparency to certify an AI system, at least for the third-party conducting the external audit. Another interdependency with an attribute subordinate to transparency is that the system logic described in the system documentation needs to describe fallback options that are implemented to achieve robustness by an appropriate failure behavior. If there is a mismatch between failure behavior and described fallback options, users trust may be destroyed.

The feedback loop is a tool for building trust with users if post-hoc explainability measures do not provide a sufficient explanation for incidents. Moreover, providing users with a feedback loop fuels the iterative development process. In doing so, the organization takes accountability for its AV system and its shortcomings, which in turn increases safety and builds users trust. As the iterative development process is required to improve the safety level of the AV and its AI system, it is interdependent with safety. To ensure the reliability of the AI system, it is necessary to constantly retrain the AI to improve its performance and ensure the representativeness of training data. A system trained on non-representative data can consequently erode trust via the unreliability of the system and possible patterns of discrimination.

For achieving accountability in each individual case, it is important to have explainability via a logging system. The logging system helps to answer the question of why a vehicle

did something and who is responsible for it. Only through the combination of the post-hoc explainability via a logging system and organizational accountability, trust in the technology can be established. Therefore, when considering changes on the logging system, the effect on post-hoc explainability and ultimately on the organizational accountability need to be considered, in order to avoid losing users trust due to declining accountability.

The adequacy of algorithms exudes an influence on the reliability and accuracy of an AI system. Companies define the system boundaries of the AV as precisely as possible because they are accountable for the outcome of the AI system in that operation domain. Therefore, the operation domain and the definition of system boundaries are closely related. If companies define system boundaries which constitute an operation domain for an AV they feel comfortable with, the system can be perceived as more trustworthy by potential users.

Company guidelines on AI systems are considered to encompass ethical considerations and are therefore related to ethic guidelines. Therefore, one could argue, if ethical guidelines are not viewed as trust-building measures by users, company guidelines may not have much impact on user trust either. However, company guidelines could be so extensive in describing how the company ensures TAI that ethical considerations make up only a small part of them, possibly insignificant for users. Thus, the trust-building property of the company guidelines would be called into question by ethical guidelines that do not create trust. Until a reliable statement can be made, the two guidelines should be regarded as interdependent to avoid jeopardizing the establishing of TAI.

In general, there are many interdependencies with safety and accountability. These interdependencies can be attributed to the fact that safety and accountability are of particular importance for building trust with users in the context of AVs. Moreover, certifiability of AV technology, including the AI system, can be considered an overarching attribute required to give credibility to a provider's AV. Especially since non-professionals are not able to review all properties of a vehicle to gain trust in a system. Therefore, TAI in AVs requires certification by a trusted third party, such as regulatory authorities.

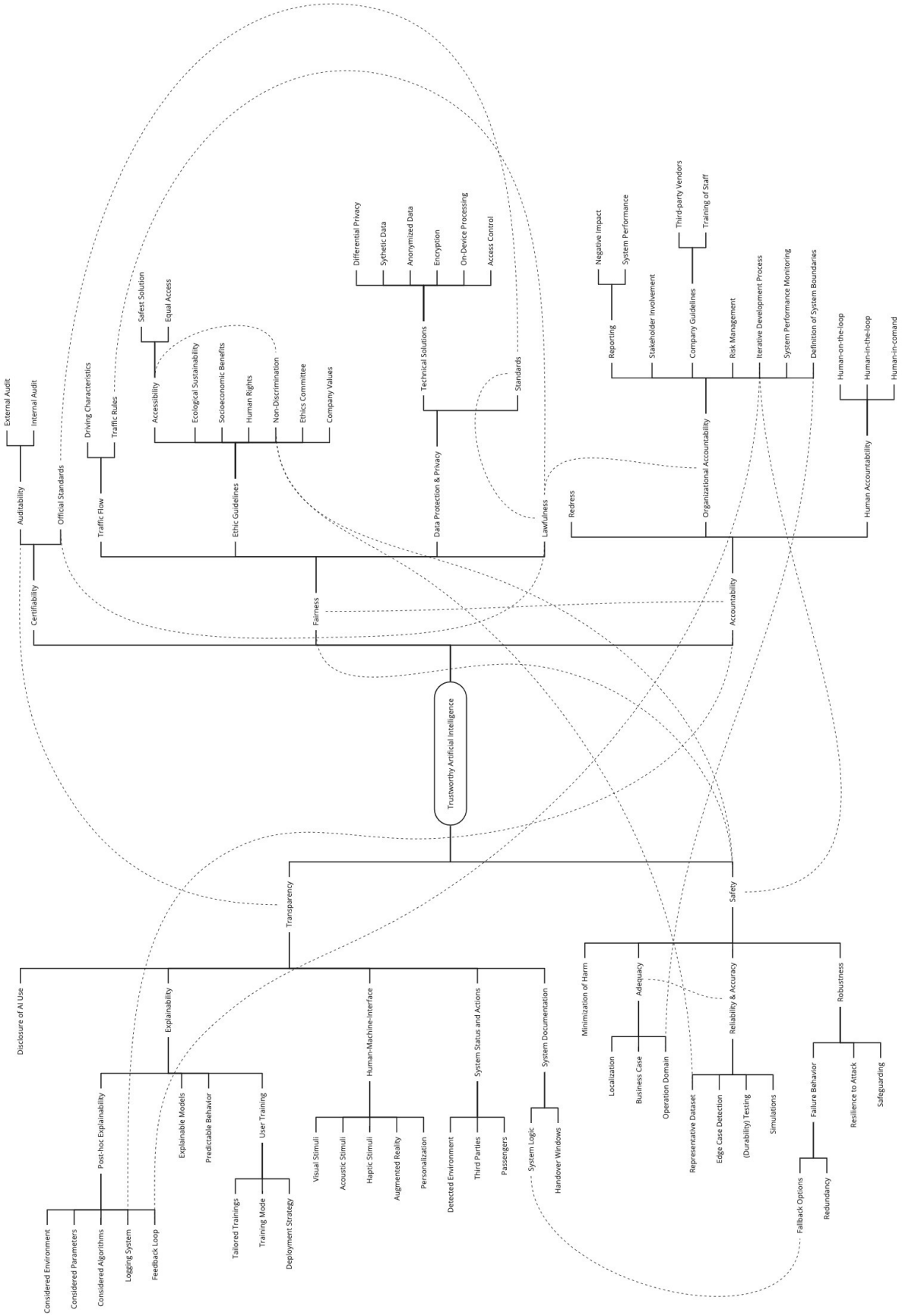

Figure 3: Framework TAI in AV with Interdependencies

5.4 Implications for Research

The framework developed is a starting point for researchers investigating which FATE attributes contribute to TAI in AVs. For the first time, the framework gives an overview of FATE attributes in the context of AVs. The proposed C-FATS attributes are a contextualization of the FATE characteristics described in the existing literature and are therefore a contribution to the conceptualization of trust in AI technology. Researchers can use the identified attributes to guide their further research in the field of AV technology.

The results confirm that contextualizing FATE characteristics is of particular importance for conceptualizing TAI. Therefore, researchers should more strongly contextualize their studies on trust formation in AI technologies to account for different technologies and use cases.

The attributes in the framework that cannot be assigned solely to "trust in technology" or "trust in organizations" and the interdependencies of the different attributes show that IS research needs to think more broadly to conceptualize trust in AI technologies. Moreover, trust transfer from organizations to technology seems of high importance in the context of AV technology and therefore should be part of trust conceptualizations.

5.5 Implications for Practice

The results suggest that establishing fairness, accountability, transparency, and safety is paramount to user trust in AV technology. Therefore, providers must ensure these attributes during development and deployment and emphasize them in communication with customers. In addition, certification of the aforementioned characteristics is an important tool that lends credibility and builds trust among users. The framework is also important for providers to review their own AI technology against the proposed C-FATS attributes. The attributes help providers identify potential white spots and might guide the development of AI technology in AVs to increase the perceived trustworthiness of AVs. If a high level of trust in the AV technology is achieved, many of the positive societal effects described in Chapter 1 are likely realized.

In addition, the framework enables a systematic comparison of providers and their AV systems with respect to TAI. This can help providers to benchmark their solutions against the competitive landscape.

5.6 Limitations and Further Research

The regional subset of the analysis (only North America and Europe) and the selection of a limited number of AV providers for the content analysis limit the ability to generalize

the findings. Due to the focus of the content analysis on the communication of providers in regard to building trust in AI in AVs, purely safety-focused content was not considered. Since the analysis revealed safety as an essential characteristic for building trust in AI systems of AVs, the content published by providers on AV safety should be included in the analysis even without an explicit link to the topic of trust, to complement the results of the analysis. Additionally, one could argue that due to the high competitiveness in the field of developing AVs, providers protect their knowledge and do not share as much information as they possibly could. Due to the limited resources available for obtaining interview partners and the scope of the work, the interviews conducted also do not constitute a representative sample. Any objective to generalize the findings should therefore be carefully considered and justified extensively. The websites and guidelines of providers included in the analysis are not only focused on TAI in AVs but on TAI in general and additionally include different business processes. Findings therefore might not be as AV specific as hoped. Furthermore, the information utilized for the analysis in this thesis are guidelines from providers and often only assumptions expressed by experts. Thus, the attributes summarized in the C-FATS framework are not collectively exhaustive properties with scientifically quantified influence on trust in AI, but provide an overview of attributes to be investigated in more detail.

Further research could seek to replicate the findings of this study through a larger number of interviews and analyzed providers. For example, different provider departments should be interviewed more systematically, and providers from Asia should be included in the analysis. Moreover, the focus of the analysis should be broadened to include provider information on the safety of AV technology in general. In addition, further studies could include the views of potential users on the derived attributes to demonstrate the contribution of individual attributes to trust building. These studies involve, for example, conducting experiments with potential users or analyzing the AI systems of AV providers, to ultimately find the best combination of attributes to create TAI in AVs. In particular, the role of certifications of AI systems in AVs should be investigated, considering the impact on user trust. In addition, further research can explore, for example, possible thresholds indicating how much information about the behavior of the system at different stages of use is needed to build and maintain sufficient trust in AI systems in AVs. The framework derived from the findings contains not only attributes that can be assigned to the concept of “trust in technology”, but also to the concept of “trust in organizations”. Moreover, as described, some attributes can be assigned to booth. To better investigate the contribution of C-FATS attributes to trust in AI systems in AVs, further research should distinguish between attributes promoting trust in technology versus organizations and carve out differences. In addition, the role of trust transfer in the context of AI technologies should be further explored to provide a better understanding of the trust-building process for AI technologies.

6 Conclusion

A successful market launch of AVs is only possible if users trust the AV and thus the AI powering these vehicles. To conceptualize trust in AI in the context of personalized AI systems, researchers recently started using so-called FATE characteristics. Until now, these characteristics have not been contextualized by AV specific FATE attributes. This paper answers the question of how to establish trust with the FATE characteristics in AVs by conducting a content analysis of AV providers' websites and experts' interviews. By analyzing 33 providers websites and conducting interviews with 5 industry experts, the thesis shows that in the context of AVs, the C-FATS characteristics better conceptualize trust. Applying the results of the content analysis, a framework for TAI in the context of AVs encompassing 91 C-FATS attributes is developed. In addition, differences between providers and experts regarding the conceptualization of TAI in AVs are highlighted and interdependencies that need to be considered in establishing TAI are identified. The findings suggest that establishing certifiability, fairness, accountability, transparency, and safety is paramount to user trust in AV technology. Therefore, providers must ensure these characteristics during development and deployment and emphasize them in communications with customers. The C-FATS attributes in the framework are a starting point for researchers and practitioners to better conceptualize TAI in AVs and support the need for contextualization of trust concepts. Since the interdependencies between trust-building attributes also challenge the distinction between "trust in technology" and "trust in organizations", researchers are challenged to generalize extended trust concepts in the context of AI systems to include trust transfer. Due to the limited scope of the analysis, the results of this work can only be a starting point for researchers and practitioners to further explore the trust-building C-FATS attributes of AI systems in AVs.

References

AI4People. (2020). *Ai4people's 7 ai global frameworks.* Retrieved 2021-12-8, from https://ai4people.eu/wp-content/pdf/AI4People7AIGlobalFrameworks.pdf

Alexy, O., George, G., Salter, A. J. (2013). Cui bono? the selective revealing of knowledge and its implications for innovative activity. *Academy of Management Review*, *38*(2), 270-291. doi: https://doi.org/10.5465/amr.2011.0193

Andersson, P., Ivehammar, P. (2019). Benefits and costs of autonomous trucks and cars. *Journal of Transportation Technologies*, *09*, 121-145. doi: https://doi.org/10.4236/jtts.2019.92008

Aptiv Services US, LLC, AUDI AG, Bayrische Motoren Werke AG, Beijing Baidu Netcom Science Technology Co., Ltd, Continental Teves AG & Co oHG, Daimler AG, ... Volkswagen AG (2019). *Safety first for automated driving.* Retrieved 2021-12-18, from https://www.daimler.com/documents/innovation/other/safety-first-for-automated-driving.pdf

AUDI AG. (2018a). *Audi und beyond-initiative setzen sich in globalem forum für verantwortungsvollen umgang mit ki ein.* Retrieved 2021-12-8, from https://www.audi-mediacenter.com/de/pressemitteilungen/audi-und-beyond-initiative-setzen-sich-in-globalem-forum-fuer-verantwortungsvollen-umgang-mit-ki-ein-10908

AUDI AG. (2018b). *User operation and displays – mmi touch response.* Retrieved 2021-12-18, from https://www.audi-mediacenter.com/en/technology-lexicon-7180/user-operation-and-displays-7182

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82-115. doi: https://doi.org/10.1016/j.inffus.2019.12.012

BBC. (2019). *Uber in fatal crash had safety flaws say us investigators.* Retrieved 2021-12-15, from https://www.bbc.com/news/business-50312340

Benbasat, I., Wang, W. (2005). Trust in and adoption of online recommendation agents. *Journal of the Association for Information Systems*, *6*, 72-101. doi: https://doi.org/10.17705/1jais.00065

BMW AG. (2020). *Sieben grundsätze für ki: Bmw group legt ethik-kodex für den einsatz von künstlicher intelligenz fest.* Retrieved 2021-12-8, from https://www.press.bmwgroup.com/deutschland/article/detail/T0318411DE/sieben-grundsaetze-fuer-ki:-bmw-group-legt-ethik-kodex-fuer-den-einsatz-von-kuenstlicher-intelligenz-fest?language=de

Commission, E. (2021a). *European data governance act.* Retrieved 2022-01-10, from https://digital-strategy.ec.europa.eu/en/policies/

data-governance-act

Commission, E. (2021b). *Regulatory framework proposal on artificial intelligence.* Retrieved 2022-01-10, from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Continental AG. (2020). *Ethikregeln für künstliche intelligenz.* Retrieved 2021-12-8, from https://www.overleaf.com/project/60fbfc2681a1706187be68d

Corbin, J., Strauss, A. (2014). *Basics of qualitative research : techniques and procedures for developing grounded theory.* SAGE Publications.

Crayton, E. D. (2019). *Redefining life sciences with artificial intelligence and blockchain.*

Crunchbase Inc. (2022). *Company search.* Retrieved 2022-01-05, from https://www.crunchbase.com/discover/organization.companies/c845b3b46045d71111b7365e1a643c1a

Daimler AG. (2021). *Two letters and four principles: How daimler uses artificial intelligence (ai).* Retrieved 2021-12-8, from https://www.daimler.com/sustainability/data/ki-guidelines.html

Diakopoulos, N. (2016, January). Accountability in algorithmic decision making. *Communications of the ACM*, *59*(2), 56–62. doi: https://doi.org/10.1145/2844110

Diakopoulos, N., Koliska, M. (2017). Algorithmic transparency in the news media. *Digital Journalism*, *5*(7), 809-828. doi: https://doi.org/10.1080/21670811.2016.1208053

Dr. Ing. h.c. F. Porsche AG. (2019). *Responsible and ethical use of artificial intelligence at porsche.* Retrieved 2021-12-8, from https://medium.com/next-level-german-engineering/responsible-and-ethical-use-of-artificial-intelligence-at-porsche-413acb3a1a93

Dresing, T., Pehl, T. (2018). *Praxisbuch transkription und analyse. anleitungen und regelsysteme für qualitativ forschende* (8th ed. ed.). Marburg.

Ehsan, U., Riedl, M. (2019). On design and evaluation of human-centered explainable ai systems. In *The acm chi conference on human factors in computing systems in glasgow, uk.*

European Union. (2016). *Regulation (eu) 2016/679 of the european parliament and of the council.* Retrieved 2021-12-8, from https://eur-lex.europa.eu/eli/reg/2016/679/oj

Floridi, L. (2019, 05). Establishing the rules for building trustworthy ai. *Nature Machine Intelligence*, *1*, 261–262. doi: https://doi.org/10.1038/s42256-019-0055-y

Ford Motor Company. (n.d.). *A matter of trust – ford's approach to developing self-driving vehicles.* Retrieved 2021-12-8, from https://media.ford.com/content/dam/fordmedia/pdf/Ford_AV_LLC_FINAL_HR_2.pdf

Gefen, D., Karahanna, E., Straub, D. W. (2003). Trust and tam in online shopping: An integrated model. *MIS Quarterly*, *27*(1), 51-90. Retrieved from http://www.jstor.org/stable/30036519 doi: https://doi.org/10.2307/30036519

Guidehouse Inc. (2020). *Guidehouse insights leaderboard: Automated driving vehicles.* Retrieved 2021-07-27, from https://guidehouseinsights.com/reports/guidehouse-insights-leaderboard-automated-driving-vehicles

Hyatt, K. (2021, Jan.). *Elon musk says tesla's full self-driving tech will have level 5 autonomy by the end of 2021.* Retrieved from https://www.cnet.com/roadshow/news/elon-musk-full-self-driving-tesla-earnings-call/

IEEE Standards Association. (2016). *Ieee p7002 – ieee draft standard for data privacy process.* Retrieved 2021-12-8, from https://standards.ieee.org/project/7002.html

Independent High-Level Expert Group on Artificial Intelligence of the European Commission. (2019). *Ethics guidelines for trustworthy ai.* Retrieved 2021-07-14, from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

International Organization for Standardization. (2018). *Road vehicles — functional safety* (Standard No. ISO 26262-1:2018). International Organization for Standardization. Retrieved 2022-01-07, from https://www.iso.org/standard/68383.html

International Organization for Standardization. (2019). *Road vehicles — vehicle dynamics test methods* (Standard No. ISO 15037-1:2019). International Organization for Standardization. Retrieved 2022-01-07, from https://www.iso.org/standard/70164.html

Ipsos. (2018). *Entrepreneurialism – the emergence of social entrepreneurialism to compete with business entrepreneurialism: An ipsos global advisor survey.* Retrieved 2022-01-07, from https://www.ipsos.com/en/entrepreneurialism-alive-and-well-and-taking-todays-social-challenges

Jarvenpaa, S., Shaw, T. R., Staples, D. (2004). Toward contextualized theories of trust: The role of trust in global virtual teams. *Information Systems Research*, *15*, 250-267. doi: https://doi.org/10.1287/isre.1040.0028

Jayaraman, S. K., Creech, C., Tilbury, D. M., Yang, X. J., Pradhan, A. K., Tsui, K. M., Robert, L. P. (2019). Pedestrian trust in automated vehicles: Role of traffic signal and av driving behavior. *Frontiers in Robotics and AI*, *6*, 117. Retrieved from https://www.frontiersin.org/article/10.3389/frobt.2019.00117 doi: https://doi.org/10.3389/frobt.2019.00117

Koul, S., Eydgahi, A. (2018, 12). Utilizing Technology Acceptance Model (TAM) for driverless car technology Adoption. *Journal of technology management & innovation*, *13*, 37-46. doi: https://doi.org/10.4067/S0718-27242018000400037

KPMG. (2020). *2020 autonomous vehicles readiness index.* Retrieved 2022-01-10, from https://assets.kpmg/content/dam/kpmg/es/pdf/2020/07/2020_KPMG_Autonomous_Vehicles_Readiness_Index.pdf

Krippendorff, K. (2004). *Content analysis: an introduction to its methodology* (2nd ed.

ed.). Thousand Oaks, California: Sage.

Lansing, J., Sunyaev, A. (2016, June). Trust in cloud computing: Conceptual typology and trust-building antecedents. *SIGMIS Database*, *47*(2), 58–96. doi: https://doi.org/10.1145/2963175.2963179

Lee, J. D., See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80. doi: https://doi.org/10.1518/hfes.46.1.50_30392

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 2053951718756684. doi: https://doi.org/10.1177/2053951718756684

Lins, S., Grochol, P., Schneider, S., Sunyaev, A. (2016). Dynamic certification of cloud services: Trust, but verify! *IEEE Security Privacy*, *14*(2), 66-71. doi: https://doi.org/10.1109/MSP.2016.26

Makridakis, S. (2017). The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms. *Futures*, *90*, 46-60. doi: https://doi.org/10.1016/j.futures.2017.03.006

Markus, A., Kors, J., Rijnbeek, P. (2020). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, *113*, 103655. doi: https://doi.org/10.1016/j.jbi.2020.103655

Mayer, R. C., Davis, J. H., Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709-734. doi: https://doi.org/10.5465/amr.1995.9508080335

Mcknight, D., Carter, M., Thatcher, J., Clay, P. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, *2*, 1-25. doi: https://doi.org/10.1145/1985347.1985353

McKnight, D. H., Choudhury, V., Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, *13*(3), 334-359. doi: https://doi.org/10.1287/isre.13.3.334.81

Myers, M. D. (2013). *Qualitative research in business & management* (2nd ed. ed.). London: Sage.

Nastjuk, I., Herrenkind, B., Marrone, M., Brendel, A. B., Kolbe, L. M. (2020). What drives the acceptance of autonomous driving? an investigation of acceptance factors from an end-user's perspective. *Technological Forecasting and Social Change*, *161*, 120319. doi: https://doi.org/10.1016/j.techfore.2020.120319

NBC News. (2018). *Trust in facebook has dropped by 66 percent since the cambridge analytica scandal.* Retrieved 2022-01-07, from `https://www.nbcnews.com/business/consumer/trust-facebook-has-dropped-51-percent-cambridge-analytica-scandal-n867011`

Network of Employers for Traffic Safety. (unknown). *Steer clear of bad driv-*

ing. Retrieved 2021-12-12, from https://www.hopkinsmedicine.org/hse/memos/steerclear.pdf

Omeiza, D., Webb, H., Jirotka, M., Kunze, L. (2021). Explanations in autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 1-21. doi: https://doi.org/10.1109/TITS.2021.3122865

Othman, K. (2021). Public acceptance and perception of autonomous vehicles: a comprehensive review. *AI and Ethics*, 1-33. doi: https://doi.org/10.1007/s43681-021-00041-8

Penmetsa, P., Sheinidashtegol, P., Musaev, A., Adanu, E. K., Hudnall, M. (2021). Effects of the autonomous vehicle crashes on public perception of the technology. *IATSS Research*. doi: https://doi.org/10.1016/j.iatssr.2021.04.003

Poole, D., Mackworth, A., Goebel, R. (1997). *Computational intelligence: A logical approach.* USA: Oxford University Press, Inc.

Public Affairs Council. (2021). *2021 public affairs pulse survey report.* Retrieved 2022-01-07, from https://pac.org/wp-content/uploads/Pulse_2021_Report.pdf

Rai, A. (2020). Explainable ai: from black box to glass box. *Journal of the Academy of Marketing Science*, *48*, 137–141. doi: https://doi.org/10.1007/s11747-019-00710-5

Renner, M., Lins, S., Sunyaev, A. (2021). A taxonomy of is certification's characteristics. In *2021 2nd international conference on internet and e-business* (p. 1–8). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3471988.3471989

Renner, M., Lins, S., Söllner, M., Thiebes, S., Sunyaev, A. (2021, 12). Achieving trustworthy artificial intelligence: Multi-source trust transfer in artificial intelligence-capable technology..

Renner, M., Lins, S., Söllner, M., Thiebes, S., Sunyaev, A. (2022, 01). Understanding the necessary conditions of multi-source trust transfer in artificial intelligence.. doi: 10.24251/HICSS.2022.717

Robert Bosch GmbH. (2020). *Ki-kodex von bosch im Überblick.* Retrieved 2021-12-8, from https://assets.bosch.com/media/de/global/stories/ai_codex/bosch-code-of-ethics-for-ai.pdf

Rousseau, D., Sitkin, S., Burt, R., Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, *23*(3), 393-404. doi: https://doi.org/10.5465/AMR.1998.926617

Russell, S., Norvig, P. (1995). *Artificial intelligence: A modern approach* (1st ed.). Prentice Hall.

SAE International. (2016). *Cybersecurity guidebook for cyber-physical vehicle systems* (Standard No. J3061_201601). Warrendale, PA, USA: SAE International. Retrieved 2021-12-15, from https://www.sae.org/standards/content/j3061_201601/

SAE International. (2020). *Automated driving system data logger* (Standard No.

J3197_202004). Warrendale, PA, USA: SAE International. Retrieved 2021-01-10, from https://www.sae.org/standards/content/j3197_202004/

SAE International. (2021a). *Sae levels of driving automation™ refined for clarity and international audience.* Retrieved 2022-01-05, from https://www.sae.org/blog/sae-j3016-update

SAE International. (2021b). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles* (Standard No. J3016_202104). Warrendale, PA, USA: SAE International. Retrieved 2021-12-15, from https://www.sae.org/standards/content/j3016_202104/

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.*, *3*, 210-229.

Shin, D. (2020). User perceptions of algorithmic decisions in the personalized ai system: perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media*, *64*(4), 541-565. doi: https://doi.org/10.1080/08838151.2020.1843357

Shin, D., Park, Y. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, *98*, 277-284. doi: https://doi.org/10.1016/j.chb.2019.04.019

Statista. (2016). *Vertrauen in vw sinkt.* Retrieved 2021-12-8, from https://de.statista.com/infografik/6868/vertrauen-in-vw-nach-dieselgate/

Sunyaev, A., Schneider, S. (2013, feb). Cloud services certification. *Commun. ACM*, *56*(2), 33–36. doi: https://doi.org/10.1145/2408776.2408789

Thatcher, J. B., McKnight, D. H., Baker, E. W., Arsal, R. E., Roberts, N. H. (2011). The role of trust in postadoption it exploration: An empirical examination of knowledge management systems. *IEEE Transactions on Engineering Management*, *58*(1), 56-70. doi: https://doi.org/10.1109/TEM.2009.2028320

Thiebes, S., Lins, S., Sunyaev, A. (2020). Trustworthy artificial intelligence. *Electronic Markets*, *31*, 447–464. doi: https://doi.org/10.1007/s12525-020-00441-4

U.S. Department of Labor Occupational Safety and Health Administration. (2006). *Guidelines for employers to reduce motor vehicle crashes.* Retrieved 2021-12-12, from https://www.osha.gov/sites/default/files/publications/motor_vehicle_guide.pdf

Valeo S.A. (2019). *Ai for humanity: French industry engages on artificial intelligence.* Retrieved 2021-12-8, from https://www.valeo.com/en/ai-for-humanity-french-industry-engages-on-artificial-intelligence/

Villani Mission on artificial intelligence. (2019a). *Ai for humanity: French industry engages on artificial intelligence.* Retrieved 2021-12-8, from https://www.edf.fr/sites/default/files/contrib/groupe-edf/espaces-dedies/espace-medias/cp/2019/20190703-cp-iahumanity-en.pdf

Villani Mission on artificial intelligence. (2019b). *Executive summary.* Retrieved 2021-12-8, from `https://www.aiforhumanity.fr/pdfs/MissionVillani_Summary_ENG.pdf`

Volkswagen AG. (2019). *Volkswagen electric car id.3 communicates using light.* Retrieved 2021-12-18, from `https://www.volkswagen-newsroom.com/en/press-releases/volkswagen-electric-car-id3-communicates-using-light-5424`

Volkswagen Group Machine Learning Research Lab. (2020a). *machine learning: next ethical steps.* Retrieved 2021-12-8, from `https://www.inf.elte.hu/dstore/document/1786/Patrick%20van%20der%20Smagt%202020-05-ELTE-Neumann.pdf`

Volkswagen Group Machine Learning Research Lab. (2020b). *Patrick van der smagt.* Retrieved 2021-12-8, from `https://argmax.ai/team/patrick-van-der-smagt/`

Wachenfeld, W., Winner, H. (2016). The release of autonomous vehicles. In M. Maurer, J. C. Gerdes, B. Lenz, H. Winner (Eds.), *Autonomous driving: Technical, legal and social aspects* (p. 425–449). Berlin, Heidelberg: Springer. doi: https://doi.org/10.1007/978-3-662-48847-8

Yang, K., Stoyanovich, J. (2017). Measuring fairness in ranked outputs. In *Proceedings of the 29th international conference on scientific and statistical database management* (p. 1–6). New York, NY, USA: Association for Computing Machinery. doi: https://doi.org/10.1145/3085504.3085526

Özpolat, K., Gao, G. G., Jank, W., Viswanathan, S. (2013). Research note — the value of third-party assurance seals in online retailing: An empirical investigation. *Information Systems Research*, *24*(4), 1100-1111. doi: https://doi.org/10.1287/isre.2013.0489

A Appendix

Information available upon request

A.1 Interview Flyer

Information available upon request

A.2 Interview Guideline

Information available upon request

A.3 Transcript Interview Number 1

Information available upon request

A.4 Transcript Interview Number 2

Information available upon request

A.5 Transcript Interview Number 3

Information available upon request

A.6 Transcript Interview Number 4

Information available upon request

A.7 Transcript Interview Number 5

Information available upon request