TRUST CASE AND THE LINK TO SAFETY CASE

TOR STÅLHANE[1] & THOR MYKLEBUST[2]
[1]NTNU, Norway
[2]SINTEF, Norway

ABSTRACT

The TrustMe project develops a safety case for autonomous buses. A safety case is mostly based on information from the developers and refers to one or more relevant safety standards. The bases for a safety case are the defined safety standards and proof of compliance, based on the paper trails left by each required activity. A trust case is different and trust and safety assessments are not necessarily correlated. In order to make self-driving buses a success they need to be considered trustworthy. Thus, we need a "Trust case". To ensure that the vehicles are safe and to inform the public, we have developed both a developer safety case and safety case for the public. To take care of the remaining factors we have developed a "Trust case". The trust case has been developed as part of literature studies, surveys and interviews. We have made a survey of 311 passengers and interviewed 18 autonomous bus passengers. Based on literature studies, surveys and interviews, we have proposed a set of issues that should be included into a "Trust case". By providing the public with a "Trust case" together with a "safety case for the public" we will help manufacturers of autonomous vehicles and operators to gain public trust.

Keywords: trust, technology acceptance models, autonomous vehicles, safety.

1 INTRODUCTION

Both safety cases and trust cases are assurance cases. Piovesan and Griffor [1] define an Assurance case as "A structured argument, supported by evidence, intended to justify that a system is acceptably assured relative to a concern". The trust case is one type of justification – the justification of trust in travelling with autonomous buses. ISO/IEC 15026-1:2013 – Systems and software assurance – defines an assurance case as a reasoned, auditable artefact that supports the contention that its top-level claim or set of claims are satisfied. This includes systematic argumentation and its underlying evidence and explicit assumptions that support the claim(s). An assurance case contains the following:

- one or more claims about properties;
- arguments that logically link the evidence and any assumptions to the claim(s);
- a body of evidence and possibly assumptions supporting these arguments for the claim(s);
- justification of the choice of the top-level claim and the method of reasoning.

Trust is used in several ways, depending on the application area. According to Frederiksen [2] trust and risk should be considered different ways to manage uncertainty. According to Perrow, trust may be used in two ways [3]:

- Reliability trust: trust using experiences made in former interactions to assess the degree of uncertainty that is associated with a specific transaction partner, e.g., the subjective probability that a transaction with this partner, e.g., a trip with a bus, will be successful.
- Decision trust: the extent to which an entity is willing to enter into a transaction (interaction) with another.

Trust is in our case mainly reliability trust or lack thereof. The degree of trust will decide whether you will use a particular service or not. In some fora, trust is used in a rather informal

doi:10.2495/SAFE210131

way so that for instance trust, reliability and reliance are all used to identify the same thing. Trust can also be seen as a person-to-person relationship. In this case, trust is a relationship to social actors such as designers, creators and operators of technology.

If trust is a person-to-person relationship, we need to identify the persons that the users of self-driving vehicles need to trust. We also need to extend trust to include organizations such as the service provider or the software development company. For autonomous buses, this is often the vehicle manufacturer. As a starting point, we could choose the service provider, e.g., the bus operator, since he is the legally responsible person. As an alternative, we might consider the personnel at the company that built the software.

In everyday speech, safety and trust are different. As is explained in the Risk Communication Guidelines for Public Officials [4] "Safety is connected to a statistical safety analysis and is hard to grasp. For example, a scientist uses a one-in-a-million comparison to convey a specific risk measurement. Health experts understand this to mean that, given one million persons, there is one person who is at risk. To the non-technical person, however, the one person may be someone they know. The public will often personalize risk with the same conviction that most scientists depersonalize it". Trust is different from reliability – it cannot be estimated but can be based on previous experiences – own and others. In this paper we try to show a way to assess trust. Based on the trust-model use by Man et al. in their TAM model, we will build a trust case analogous to a safety case [5]. In addition to the trust case, the TrustMe project is also developing a safety case – aiming at the bus manufacturer – and a safety case for the public (see Myklebust et al. [6]).

2 THE TAM MODEL OF VENKATESH AND DAVIS

The ESREL 2021 paper on trust and self-driving buses [7] used the extended TAM-model by Venkatesh and Davis [8] (see Fig. 1). The main problem, as the TrustMe team sees it, is that perceived safety is not involved – at least not directly.

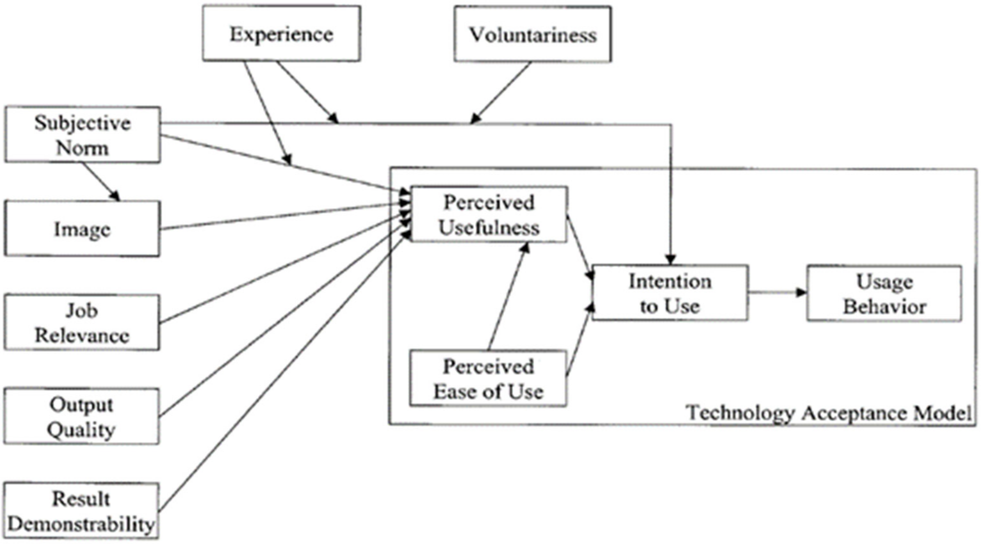


Figure 1: The TAM model of Venkatesh and Davis [8].

The term output quality needs to be clarified. Venkatesh and Davis use the following definition: "Output Quality is related to the tasks a system is capable of performing and the degree to which those tasks match their job goals (job relevance). People will take into consideration how well the system performs its tasks, which we refer to as perceptions of output quality. Empirically, the relationship between perceived output quality and perceived usefulness has been shown before. We expect output quality to be empirically distinct from, and to explain significant unique variance in, perceived usefulness over and above job relevance because a different underlying judgmental process is involved".

In addition to the issues included in output quality by Venkatesh and Davis, the TrustMe project wants to also include relevance for vacations and leisure time and thus replace "Job relevance" with "Transport needs relevance".

3 THE TAM-MODEL OF MAN ET AL.

One weakness with the TAM model of Venkatesh and Davis is that it does not consider trust except as a part of output quality. The TAM-model of Man et al. contains trust as a separate component which makes it easier to discuss this part of technology acceptance.

The dotted connection lines in Fig. 2 are connections that turned out to not be significant based on available data. Note that perceived privacy risk has no significant influence, neither on trust nor on perceived usefulness.

- Compatibility can be defined as the degree to which a technology complies with the needs and lifestyles of users.
- System quality refers to overall consumer perceptions of the excellence and effectiveness of a particular system. Note that this not the same as the quality-in-use as defined by ISO/IEC 25010:2011 – Systems and software engineering – since the standard includes "Freedom from risk" as part of quality in use. According to SQuaRE (Systems and software Quality Requirements and Evaluation), "freedom from risk" has the following components:
    - Economic risk mitigation – a product or system's ability to mitigate the potential risk to financial status, efficient operation, commercial property, reputation or other resources in the intended contexts of use.
    - Health and safety risk mitigation – a product or system's ability to mitigates the potential risk to people in the intended contexts of use.
    - Environmental risk mitigation – a product or system's ability to mitigate the potential risk to property or the environment in the intended contexts of use.

Note the difference between compatibility and system quality. Compatibility is related to customer needs while system quality is related to how well the system meets these needs. For the sake of comparison between the models of Man et al. and the model of Venkatesh and Davis, we will assume that:

- Compatibility is equivalent to subjective norm, image and job relevance. As mentioned in Section 1, TrustMe adds vacations and leisure time so that job relevance is replaced by transport-need relevance.
- System quality is equivalent to output quality (result) and demonstrability (seeing is believing).

In order to make it easier to compare the two previous models and based on the two assumptions above, we can rearrange the model of Venkatesh and Davis shown in Fig. 1 to the one shown in Fig. 3.

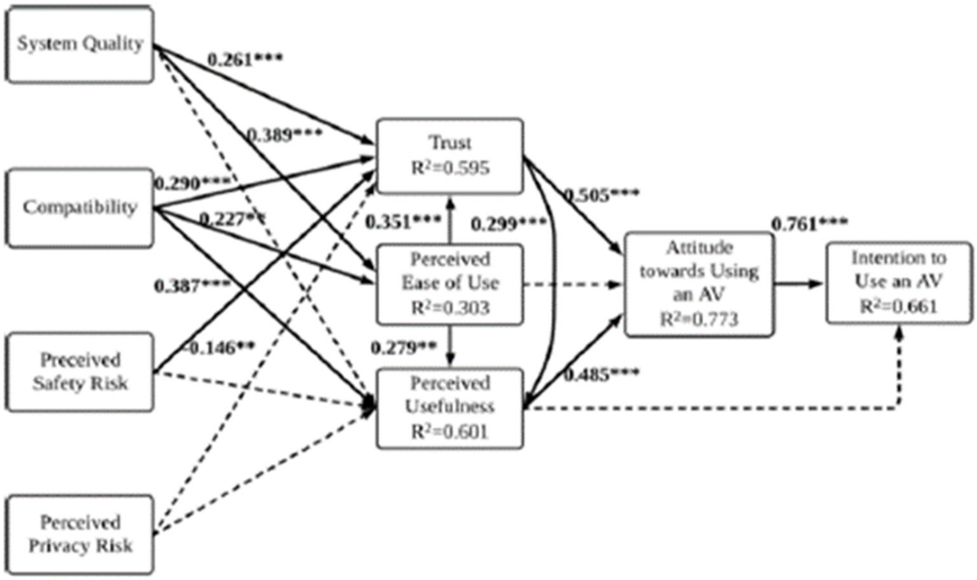


Figure 2: The TAM-model of Man et al [10].

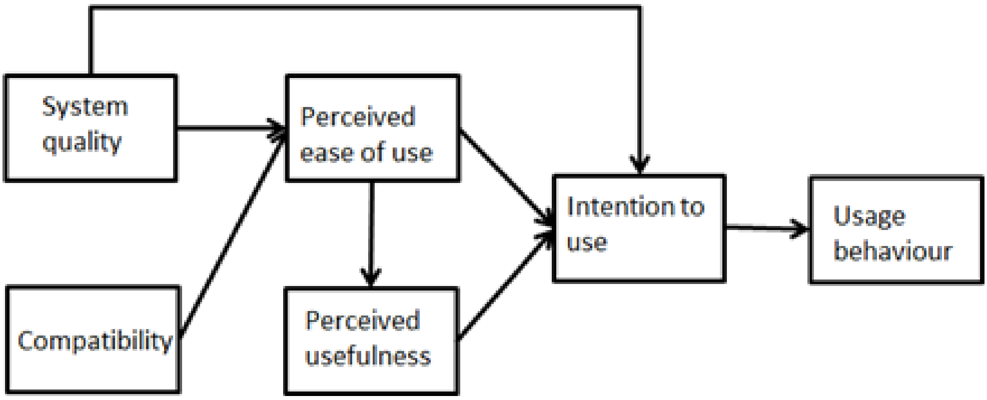


Figure 3: The TAM model of Venkatesh and Davis rearranged [8].

Health and safety risk mitigation is close to what we have called trust. There has, however, been little attention paid to environmental consequences. On the other hand, the potential for autonomous buses to create environmental damage is considered to be low.

4 TRUST ACCORDING TO ISO TR 24028

Earlier, safety and adhering to the relevant standards' requirements was considered to be sufficient for self-driving cars. With the arrival of artificial intelligence (AI) and machine learning (ML) in self-driving vehicles, trust has become an issue also here. According to the ISO TR 24028 "Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence" standard, trust is a combination of:

- Physical trust – often synonymous with the combination of reliability and safety. This will be equivalent to "the risk to safety is low" in Bezai's mode (see Section. 5).
- Cyber trust – concerns often shift to IT infrastructure security requirements. According to the experiments of Zhang et al. [9] and Man et al. [10] neither security nor privacy is considered relevant for peoples' trust. However, Zhang et al. add that "…future studies should explore the role of other factors such as reliability, perceived cyber-security risk, liability concerns, and driving pleasure on acceptance".
- Social trust – based on a person's way of life, belief, character. In our case this is related to the service provider.

Social trust is vaguely related to subjective norms and image in the original TAM model by Venkatesh and Davis (see Fig. 1). However, this perspective is left out in the TAM of Man et al. and is replaced by system quality, which refers to overall consumer perceptions of the excellence and effectiveness of a particular system.

5 UNIFYING THE MODELS OF VENKATESH AND MAN

By adding trust and perceived safety risk to Venkatesh's model, we get the unified model shown in Fig. 4. The two models referred in Figs 1 and 2 allow us to include perceived safety and trust as model factors. The dotted lines show how trust and perceived risk are added to Venkatesh's model. From Fig. 4 we see that trust, in our opinion, depends on four factors:

- system quality;
- relevance – compatibility;
- perceived ease of use;
- perceived safety risk.

Even though some psychologists and sociologists separate trust, reliability and reliance, we will follow those who claim that (1) trust is confidence in or reliance on some person, organization or quality and (2) risk assessment and rational calculations reduce uncertainty and thus increase trust. However, this does not reduce the risks. For an engineer's point of view, risk mitigation will also play an important role but this will not affect the general public.

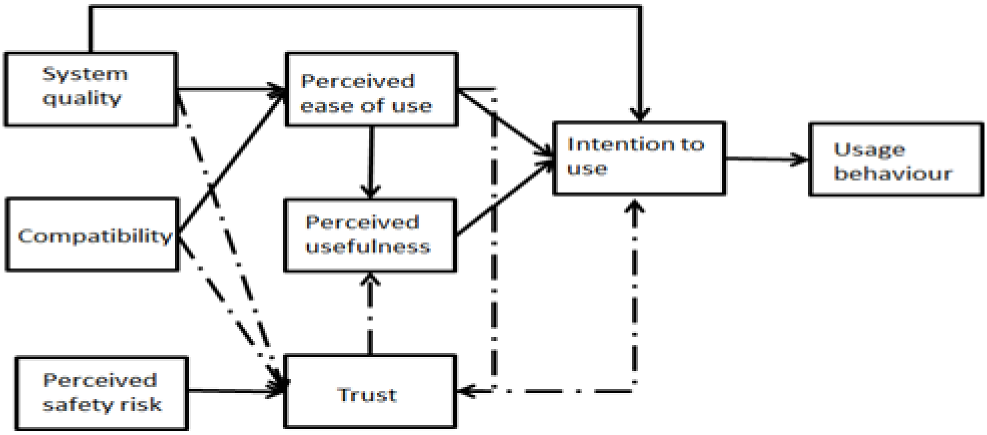


Figure 4: An attempt to unify the models of Venkatesh and Davis [8] and Man et al. [10].

The engineering view is simple and uses trust synonymous with reliability – an engineer trusts a component or system with high reliability. Thus, to build a trust case – analogous to a safety case – we need to convince the users that the four factors summed up above as influencing trust have been taken care of.

Bezai et al. [11] have developed a model that can be used to describe user's acceptance and behaviour. As we see, their model, shown in Fig. 5 has three main components – perception, vehicle performance and usage and cost. Our suggested model for a trust case has several ideas in common with this model, mostly related to vehicle performance and usage.

- Vehicle performance and usage: System quality, that the system satisfies our needs and that the system is easy to use
- Perception – the risk to safety is low and safety feeling condition

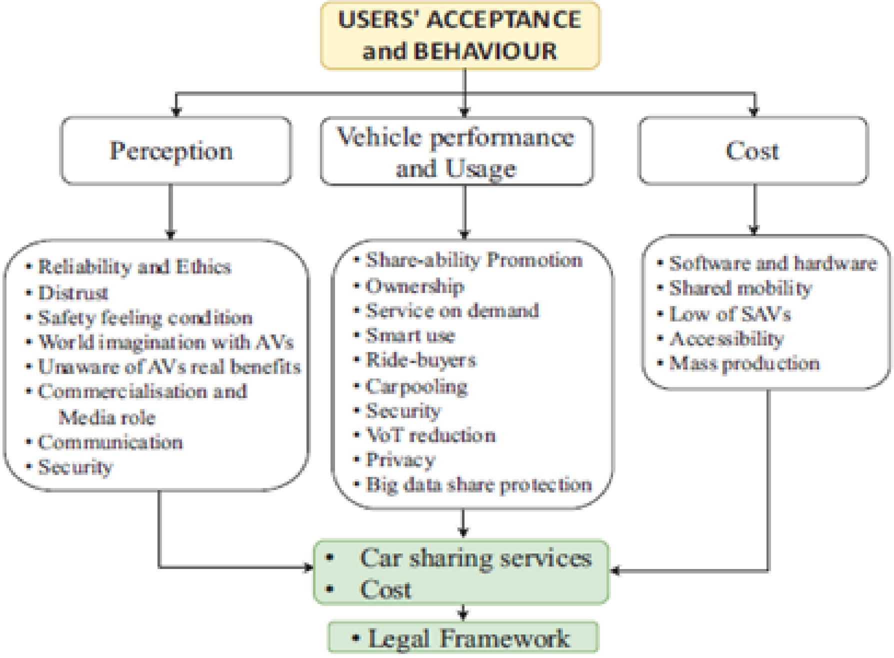
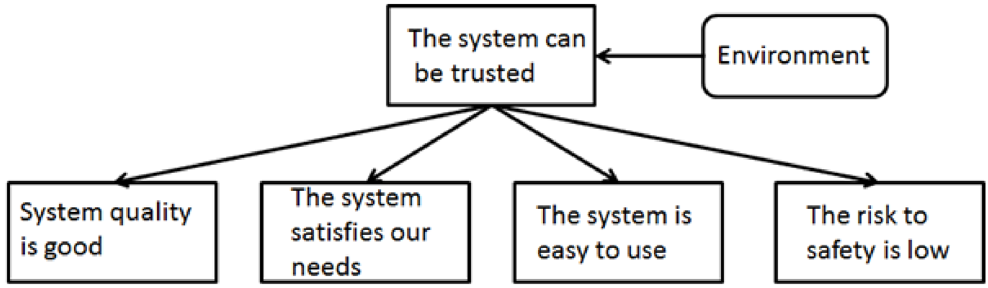


Figure 5: Bezai et al.'s model for user acceptance and behaviour [11].

The model of Bezai is useful, since it points out issues that we have not yet considered, e.g., VoT (Value of Time), ethics, commercialisation and media's role. Ethics is not relevant here for two reasons: (1) the autonomous buses considered here are level 3 (conditional driving automation) but will later move to level 4 (High driving automation). Most discussions on ethics are related to when the system should keep control and when the control should be handed over to the driver – a concept that is not relevant for self-driving buses. Ethics might, however, be relevant later when more public focus might be on the producers of autonomous buses (see also [12]). The model shown below also includes costs, which is not all that relevant for autonomous buses. In addition, the cost component in Bezai's model is mostly related to vehicle production costs.

Fig. 6 shows the four main components of trust according to the model in Fig. 3. The model may be refined by adding concepts related to each main component in the table below each main component. None of the TAM models mention environment. We need it here to cater to such things as weather, traffic density and road quality. Environment in Fig. 4 is equivalent to the ODD (operational design domain). When we use the term "The system is easy to use" we are referring to the system controlled by the computer, not the computer system itself.


The system can be trusted
Environment
System quality is good
The system satisfies our needs
The system is easy to use
The risk to safety is low

The bus works as expected	Punctuality – departure and arrival	Easy to enter and exit the bus, e.g., for wheelchair users	The bus will follow the defined route
The bus is not stopped due to SW, HW or data errors	Frequency	Easy to get ticket if necessary	If the bus has an accident, we will get help quickly
The bus can handle power loss, e.g., to the computer	Relevant areas are covered	Easy to find the right bus – to where you want to go	No harm to other vehicles, bicycles and other micro transport systems
	Comfort, e.g., smooth rides and available seating		Sensors can discover and handle all kinds of obstacles
			The computer's ability to handle complex and unknown situations
			Keep the passengers safe from e.g., robbery
			Manual operation of door-opener

Figure 6: The trust case and the relevant topic.

It is tempting to equate "The risk to safety is low" with the safety case or the safety case for the public, but it is important to bear in mind that risk here is really reliability trust, i.e., the subjective probability that a transaction with this partner, e.g., a bus, will be successful while reliability in a safety case is related to proof of compliance with a standard. However, the model shown in Fig. 5 might be included in the safety case for the public as a spate component.

Note that security is not included in Fig. 6. This is mainly based on the results from Zhang et al. [9] and Man et al. [10] who included security and privacy in their experiment but its influence on neither trust nor perceived usefulness was statistically significant. In addition, security was mentioned only once during the two focus group meetings. In Fig. 5, we have included the issues identified during two focus group meetings in 2020 [13]. The meetings were arranged by AtB – a local bus service provider in Trondheim. As you can see, the majority of the issues are related to the risk to safety. If we want to combine the trust case and the safety case the trust case component "the risk to safety is low" will be the component that bridge trust case and safety case. The environment component will be part of both cases. Trust Case will be further developed, for instance in a 1–2 pages information sheet similar to the safety case for the public [6].

The experiments described in Zhang et al. [9] and Man et al. [10] both shows that security and privacy are not considered important for trust or perceived usefulness. The results are surprising, considering the strong focus on privacy and security we have lately seen in the media. Thus, we will perform a new survey on autonomous buses and be more explicit when it comes to security challenges.

6 CONCLUSIONS

One of the goals of the TrustMe project is to develop a complete Trust case for the public for autonomous buses. The model presented in Fig. 5 is the first step for this work. Both the technological and psychological factors are important. However, for the general public, the psychological factors are the most important ones, see, for instance, the quote from Risk Communication Guidelines for Public Officials [4]: "A scientist uses a one-in-a-million comparison to convey a specific risk measurement. Health experts understand this to mean that, given one million persons, there is one person who is at risk. To the non-technical person, however, the one person may be someone they know. The public will often personalize risk with the same conviction that most scientists depersonalize it".

Thus we need to focus more on this side of the acceptance of autonomous buses. The psychological factors are included in a trust case, as shown in Fig. 6.

ACKNOWLEDGEMENT

The "TrustMe" project is supported by The Norwegian Research Council, project number 309207 – IPOFFENTLIG19.

REFERENCES

[1] Piovesan, A. & Griffor, E., Reasoning about safety and security. *Handbook of System Safety and Security*, ed. E. Griffor, Elsevier, Chapter 7, 2017.

[2] Frederiksen, M., Trust in the face of uncertainty: A qualitative study of inter-subjective trust and risk. *International Review of Sociology*, **24**(1), 130–144, 2014.

[3] Perrow, C., *Normal Accidents: Living with High Risk Technologies*, Basic Books, p. 326, 1984.

[4] U.S. Department of Health and Human Services, Communicating in a crises: Risk Communication Guidelines for Public Officials. Public Health Service: Rockville.

[5] Myklebust, T. & Stålhane, T., *The Agile Safety Case*, Springer, 2018.

[6] Myklebust, T., Stålhane, T. & Jensen, G., *Safety Case for the Public*, ESREL, 2021.

[7] Stålhane, T., Myklebust, T. & Haug, I.S., Trust and acceptance of self-driving buses. Submitted to ESREL, 2021.

[8] Venkatesh, V. & Davis, F.D., A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, **46**(2), pp. 186–204, 2000.

[9] Zhang, T., Tao, D., Qu, X., Zhang, X., Lin, R. & Zhang, W., The roles of initial trust and perceived risk in public's acceptance of automated vehicles. *Transportation Research Part C*, 2019.

[10] Man, S., Xiong, W., Chang, F. & Chan, A.H.S., Critical factors influencing acceptance of automated vehicles by Hong Kong, *IEEE Access*, 2020.

[11] Bezai, N.E., Medjdouba, B., Al-Habaibeha, A., Chalal, M.L. & Fadli, F., Future cities and autonomous vehicles: Analysis of the barriers to full adoption. *Energy and Built Environment*, **2**, 2021.

[12] Narayanan, A., When is it right and good for an intelligent autonomous vehicle to take over control (and hand it back)? Department of Computer Science School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, New Zealand.

[13] Haug, I.S., The customers' experiences with self-driving buses, November 2020. (In Norwegian.)