Towards common ethical and safe 'behaviour' standards for automated vehicles

Papadimitriou, Eleonora; Farah, Haneen; van de Kaa, Geerten; Santoni De Sio, Filippo; Hagenzieker, Marjan; van Gelder, Pieter

DOI
10.1016/j.aap.2022.106724

Publication date
2022

Document Version
Final published version

Published in
Accident Analysis and Prevention

Citation (APA)
Papadimitriou, E., Farah, H., van de Kaa, G., Santoni De Sio, F., Hagenzieker, M., & van Gelder, P. (2022). Towards common ethical and safe 'behaviour' standards for automated vehicles. *Accident Analysis and Prevention*, *174*, Article 106724. https://doi.org/10.1016/j.aap.2022.106724

Accident Analysis and Prevention 174 (2022) 106724

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Towards common ethical and safe 'behaviour' standards for automated vehicles

Eleonora Papadimitriou [a,*], Haneen Farah [b], Geerten van de Kaa [a], Filippo Santoni de Sio [a], Marjan Hagenzieker [b], Pieter van Gelder [a]

[a] Technical University Delft, Faculty of Technology, Policy and Management, Jafalaan 5, 2628BX Delft, the Netherlands
[b] Technical University Delft, Faculty of Civil Engineering & Geosciences, Stevinweg 1, 2628 CN Delft, the Netherlands

ARTICLE INFO

Keywords:
Automated vehicles
Safety
Ethics
Standardisation

ABSTRACT

Automated vehicles (AVs) aim to dramatically improve traffic safety by reducing or eliminating human error, which remains the leading cause of road crashes. However, commonly accepted standards for the 'safe driving behaviour of machines' are pending and urgently needed. Unless a common understanding of safety as a design value is achieved, different manufacturers' driving styles may emerge, resulting in inconsistent, unpredictable and potentially unsafe 'behaviour' of AVs in certain situations. This paper aims to explore the main gaps and challenges towards establishing shared safety standards for the 'behaviour' of AVs, and contribute to their responsible traffic integration, by reviewing the state-of-the-art on AV safety in the core relevant disciplines: ethics of technology, safety science (engineering & human factors), and standardisation. The ethical and safety aspects investigated include the users' perception of AV safety, the ethical trade-offs in critical decision-making contexts, the pertinence of data-driven approaches for AVs to mimic human behaviour, and the responsibilities of various actors. Moreover, the paper reviews the current safety patterns, metrics (surrogate measures of safety – SMoS) and their thresholds introduced in existing research for three use cases: mixed traffic of AV and conventional vehicles, AV interaction with pedestrians and cyclists, and transition of control from machine to human driver. The results reveal several knowledge gaps within each discipline and highlights the lack of common understanding of safety across disciplines. On the basis of the results, the paper proposes a framework for further research on AV safety, identifying concrete opportunities for interdisciplinary research, with common goals and methodologies, and explicitly indicating the path for transfer of knowledge between sectors.

1. Introduction

Automated vehicles (AVs) aim to bring dramatic safety improvements by minimizing human driver's role and dramatically reducing human error, which remains the primary contributory factor of road crashes. Superior safety is the principal "banner value" to promote AVs technology, however it has been proved difficult to responsibly claim, as the existing evidence is insufficient (ETSC, 2016; ITF/OECD, 2018). This creates the risk of misinformation, by which societies may be misled into decisions that serve the industry's interest, rather than citizens' interest. Establishing an objective baseline and coherent metrics of road safety that enables a fair assessment of AVs performance relative to non-AVs, and thereby demonstrating AVs societal benefit, is the recommendation no.1 of the recent report of the EU expert group on Ethics of Connected and Automated Vehicles (EC Expert Group E03659, 2020).

A concise definition of safety in the context of AVs that is laid down in safety standards is still lacking and is a significant challenge for ensuring AV safety (Koopman & Wagner, 2017; Dankd & London, 2017). Unless a common understanding of machines' 'ethical and safety behaviour' is achieved, different manufacturers inevitably develop different driving styles of AVs. That may result in inconsistent 'behaviour' of different types of AVs in different situations, discordance of that 'behaviour' with the values and expectations of other road users (e.g. conventional vehicles' drivers, pedestrians, cyclists), and uncertain safety consequences.

In the literature, AVs' safety is investigated from ethical perspectives, safety science perspectives (including, but not limited to, transportation engineering, human factors and systems engineering), as well as standardisation and policy perspectives. However, several questions on AV safety standards touch upon more than one of the above core

* Corresponding author.
E-mail address: e.papadimitriou@tudelft.nl (E. Papadimitriou).

https://doi.org/10.1016/j.aap.2022.106724
Received 13 March 2021; Received in revised form 12 May 2022; Accepted 28 May 2022
Available online 9 June 2022

disciplines, yet they are seldom explored with an interdisciplinary approach (see Fig. 1). Addressing the broad question of how to define safety in the context of AVs ('how safe is safe enough'), involves evident safety engineering and human factors issues (e.g., what metrics should be used, what thresholds, in which scenarios), but also overarching ethical issues (e.g. what moral/legal/societal values should be reflected in machines' decision-making and how should (residual) risk be distributed, what trade-offs exist between safety and other moral and societal values) (Nyholm & Smids, 2020) etc. Furthermore, it involves the question of how to translate values and design principles into standards, and involve all relevant stakeholders from industry and research. This inherent interdisciplinarity has not been taken into account in existing research, and the potential benefits from the integration of efforts in different disciplines, have not been sufficiently investigated.

2. Objectives and methods

This paper aims to identify gaps in existing research towards the definition of typical 'safe behaviour' standards for AVs and to contribute to their consistent design and responsible traffic integration. The research question is: how can acceptable safety standards for the "behaviour" of AVs be developed.

In this paper, automated vehicles are defined and discussed in accordance with the SAE taxonomy as "*vehicles that perform part or all of the dynamic driving task on a sustained basis*" (SAE International, 2018), with 6 levels of automation ranging from: level 0 (no automation), level 1 (automated longitudinal or lateral control, driver assistance), level 2 (automated longitudinal and lateral control, partial automation, driver response needed), level 3 (automated control, conditional automation, response-ready driver), level 4 (fully automated control in certain conditions, manual driving possible), and level 5 (fully automated control in all conditions). Typically, levels 1 & 2 (currently available on the market) are referred to as low levels of automation, while levels 4 & 5 are referred to as high levels of automation.

A dedicated literature search is carried out, with particular emphasis on the principles, methods and indicators used to define safety, the degree of standardisation of approaches and the main challenges involved within the core disciplines: ethics, safety (engineering, human factors), standardisation. It is noted that integration between safety engineering and human factors/behavioural approaches is already in place to a large extent within existing research, hence the term "safety" used in this research includes both aspects. Moreover, both descriptive ethics (empirical study of people's perception) as well as normative ethics (conceptual and normative analysis), are considered. Safety and security issue related to hardware or software development (e.g. mechanical engineering, computer science etc.) were considered out of this search scope.

The literature search was done for the period 2010–2020 by using combinations of terms in the form of <Automated vehicles > AND < safety > AND < discipline >. The following synonyms were included in the search terms:

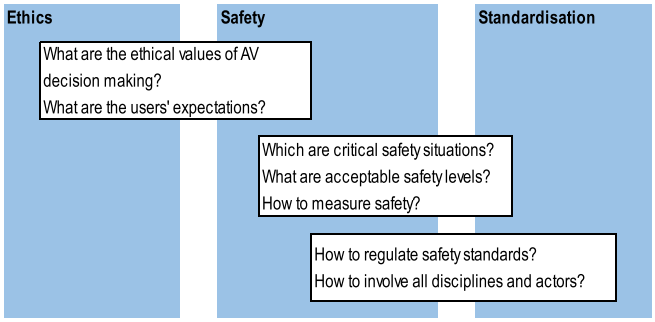

Fig. 1. Inter-disciplinary questions in defining AV safety standards.

- Autonomous vehicles, automated vehicles, driverless
- Safety, safety thresholds, safety patterns, functional safety
- Ethics, moral
- Transport engineering, behaviour, human factors
- Standards, assurance, certification

For the engineering and behavioural disciplines, the search returned a very high number of studies (>1500 'hits'). For this reason, a more focused search approach was taken, i.e., to examine safety in specific safety–critical use cases: mixed traffic (conventional and automated vehicles), interaction of AVs with vulnerable road users (VRUs), and transition of control from automation to human. Therefore, in this case, the search terms were adjusted as < Automated vehicles > AND < safety > AND < use case>, with the following additional terms for use cases:

- Mixed traffic
- Pedestrians, Cyclists, VRUs
- Transition, takeover, handover.

It is noted that a systematic and exhaustive literature review was beyond the scope of this paper; of primary interest was the selection of high-quality studies that are representative of the current state-of-the-art; hence a hybrid search method was used with the following criteria: recent studies (<5 years) were prioritized over older studies, and existing review/*meta*-analysis papers were prioritized over individual studies. For the use cases, quantitative studies, i.e., studies with quantitative estimates of safety parameters were prioritised over qualitative studies, in order to assess the homogeneity of approaches and thresholds used.

The relevance of studies was assessed through title and abstract screening. An initial selection was made once a 'critical mass' of between 10 and 20 recent papers per discipline was reached. The vast majority of these recent studies included detailed and comprehensive literature reviews. Therefore, a targeted selection of key additional studies was made through the references included in the initially selected studies ('backward snowballing') – for instance, the "trolley problem" in AV ethics, and the questions on human-mimic behaviour of AVs, were eventually given further attention. Additional references were added at the suggestion of peer reviewers.

The final selection included 20 studies on ethics (6 studies on user expectations, 9 on ethical trade-offs, 5 on responsibility), 24 studies on engineering and behaviour (5 on human-mimic behaviour of AV, 6 on mixed traffic, 7 on pedestrians or cyclists, and 4 on takeover performance), and 15 studies on standardisation of AVs (6 studies on functional safety and 9 studies on safety assurance/certification).

Based on the results, this paper's remainder is structured as follows: Section 3 explores the safety values, underlying ethical trade-offs, and expectations regarding AVs 'behaviour'. Section 4 reviews existing studies on AV behaviour in different use cases, with focus on user capabilities and acceptable safety thresholds on the basis of relevant surrogate measures of safety (SMoS). Section 5 reviews recent developments in safety assurance of AVs from the system design standards, i.e. the industry perspective. A synthesis of the results in Section 6 leads to identifying gaps within each discipline and recommendations on methodologies and tools that will allow determining and testing the benefits of standardized AV 'behaviour'. Section 7 presents the conclusions of this research.

3. Ethical aspects of automated vehicles safety

3.1. User expectations, perception and acceptance

AVs have attracted the public interest by promising more efficient, inclusive, and safe mobility. Users' acceptance is key to achieve these goals. However, in several recent studies, the user's expectations,

perception, and acceptance of AVs present considerable variations.

Hulse et al. (2018) argue that the ethics of AV decisions concerning pedestrians (e.g., programmed to yield, safety choices at the benefit of pedestrians) might affect the passengers' expectations and trust. Furthermore, in an extensive population survey in Finland, the main points of concern over AVs were traffic safety and the uncertainty about AVs' moral decisions (Liljamo et al., 2018). In another survey in the US, it was shown that the vast majority of potential AV users were concerned over issues of AV safety, namely that AVs would not drive as well as human drivers, they would be unpredictable in emergencies, adverse weather, and interactions with pedestrians (Woldeamanuel & Nguyen, 2018). Moody et al. (2020) find that industrialized countries have currently lower perceived safety of AVs, and are more pessimistic over their safety potential. On the other hand, participants from low and middle-income countries, especially those with a high share of vulnerable road users' (VRU) fatalities, appear more optimistic about the number of years at which a safety benefit could be achieved. A relevant study in the EU found a relatively large share of 'hostility' towards AVs, and a negative relationship between the current road safety level in a country and favourable opinions of AVs (Hudson et al., 2019).

Gkartzonikas & Gkritza (2019) made a systematic literature review and found that safety is more frequently encountered as a concern of users than an expected benefit. Jienitz et al. (2019) propose a set of macroscopic risk acceptance thresholds (in terms of accidents per mileage travelled) for highly automated vehicles (HAV), which at the individual level are assumed to be based on the risk acceptance of professional drivers. However, from a societal perspective, it is often suggested that an unacceptable risk would result only from a "noticeable" increase in total traffic victims (PEGASUS Consortium, 2020).

3.2. Safety choices of AVs

There is significant literature devoted to the ethical dilemmas that AVs may encounter in critical situations (Lin, 2013; Bonnefon et al., 2019); these are inspired by the older "trolley problem" and describe situations in which an observer needs to decide who should be hit by a fully autonomous vehicle in a scenario where a crash cannot be avoided, but different potential victims can be selected by, e.g. swerving to the right or to the left. There are different approaches to this dilemma, which is related to vehicle design failure conditions: on the one hand, a utility-based approach would argue that the objective would be to save the higher number of lives, assuming as morally acceptable to sacrifice a smaller number of lives. This approach can be extended to estimate the total utility of each choice in relation to gender, age and other relevant characteristics. Pickering et al. (2019) developed a set of models allowing assessing the utilities of scenarios between collision against a rigid wall or into a group of one to ten pedestrians, and finding that survey respondents accepted utilitarian approaches.

Other studies have extended the dilemmas over sacrificing between own self or others (Millar, 2014), younger or older victims (Lin, 2016), law compliant and uncompliant collision opponents (Goodall, 2014) or combinations of these. 'Moral Machines' is an online application (Massachusetts Institute of Technology, 2016) presenting users with various scenarios of AV ethical dilemmas, aiming to collect data on the values, ethics and eventual choices of humans. For example, Rhim et al. (2020) revealed clusters of moral choice making: altruists, deontologists, and non-determinists (i.e., individuals making context-specific decisions), suggesting that "AV morality is pluralistic" and varies in different cultures. Cultural factors were also identified in Awad et al. (2018) who thoroughly analysed the 'Moral Machine' data.

The EC recommendations also bring forward the dimension of risk inequality: AVs should contribute to the reduction of the disproportional risk exhibited by certain road user groups (EC Expert Group E03659, 2020). Radun et al. (2019) found that humans distinguish between deliberate actions and unforeseen circumstances of the opponent in their decision whether to sacrifice themselves or not.

Santoni de Sio (2017) notices that a simple utilitarian approach would be in contrast with some deep-seated legal principles and practices of many Western jurisdictions. In line with this, and from a more deontological approach, recent German guidelines on the ethics of AVs suggest that damage to property or animals can be acceptable in order to prevent damage to humans. However, it is also suggested that AV development's focus should be placed on the total avoidance of 'genuine dilemmatic situations' involving humans, as it is unfeasible and undesirable to standardize these. The recent EU recommendations on self-driving cars' ethics propose that such dilemma situations are a limiting case of risk management, i.e. they do not require specific principles to be handled, but should be handled in compliance to the general ethical and legal principles proposed for AVs design and behaviour, among others non-maleficence, human dignity and fairness (EC Expert Group E03659, 2020). They also recommend more generally to move away from these rare, fictional scenarios towards more substantive ethical issues such as risk, data ethics, and responsibility. However, lessons learned from retrospective judgments of such incidents could be taken into account in updating regulations (Luetge, 2017). On the other hand, based on Karnouskos (2018) results, it is suggested that users expect AV manufacturers to take into account the commonly accepted ethical values.

Dennis et al. (2016) argue that it is feasible to integrate a ranking of ethical violations into AV behaviour design. Autonomous systems can be designed as (rational) agents with given beliefs, goals and plans, choosing from a set of alternative ethical plans. Mordue et al. (2020) simulated a random decision making of a scenario of a vehicle steering right, going straight or steering left, with different safety consequences in each case (crash outcome, age of victims), within four ethical approaches: egoism, utilitarianism, virtue ethics and 'moral machine' (i.e., the 'democratic' result of a relative online experiment). Although utilitarianism and 'moral machines' resulted in fewer fatalities, the results were comparable to those of the other approaches on the aggregate level; nevertheless, there was significant variation in the composition of fatalities, revealing how different programming of AVs would result in large discrepancies in the related societal impacts.

3.3. Responsibility and liability

As higher levels of automation are achieved, and the control of a vehicle is gradually transferred to the machine, responsibilities will be redistributed across the network of human individuals and organisations involved in their manufacture, deployment, regulation and use (EC Expert Group E03659, 2020, p. 53; Mordue et al., 2020). In this context, vehicle failure conditions will be subject to new responsibility and liability questions. The EC recommendations also stress the importance of distinguishing five senses of responsibility i.e. two forward-looking responsibilities: obligation and capacity to design, regulate, use AVs responsibly; and three backward-looking responsibilities: accountability (duty to explain the behaviour of an AVs), moral culpability (being legitimately blamed or shamed for the behaviour of an AVs) and legal liability (being legally forced to compensate for the wrong behaviour of am AVS).

As for obligations, Banks et al. (2019) use a Risk Management Framework to represent the responsibilities and expectations of different actors at the macro (international organisations, regulators etc.), meso (manufacturers, resource providers etc.) and the micro (end users, equipment and environment) levels. The authors argue that, while manufacturers have been left to manage safety in their own approaches, the responsibility envisaged to be placed on 'meso-level' actors is disproportionate. By a network analysis of relevant stakeholders in the UK, it is recommended that responsibility should lie with macro-level actors who should provide legislation 'leaving little room for interpretation', while manufacturers should be responsible for providing evidence of their compliance with requirements.

In this framework, De Bruyne & Werbrouck (2018) stress the role of

supernational institutions and regulators in clarifying liability for AVs beyond the current – and rather outdated – approach of technical reliability (e.g. EU Directive 85/374/EEC), and highlight several areas leaving room for unclear liability (e.g. does an update of the software of an AV make it a new product?). Noy et al. (2018) distinguish the driver error from driver culpability, pointing out that there may be not only situations where a crash may be due to vehicle or environmental factors, but also situations where driver error is unintentional (e.g. due to misperception) – and it is unlikely that AI algorithms can handle these situations in a straightforward way. The same is the case for intentional violations (e.g. deliberate autopilot disengagements in AVs).

It is expected that the driver is in principle fully responsible at lower levels of automation, at least from insurers' perspective (liability). However, at higher levels of automation, there are at least three elements that need to be known for a fair assessment of the driver's culpability: (i) whether full automation was active, (ii) whether there was a mandatory authority transition request and the response of the driver to that, and (iii) whether there was a violation of traffic rules by

Table 1
Summary of studies on ethical issues of AV safety.

	Ethical issues			Methods					
Source	User acceptance / expectations	moral safety choices	Responsibility / Liability	Survey/experimet	In depth interviews	Review / meta-analysis	Descriptive / Qualitative analysis	Quantitative analysis	Results
Hulse et al. (2018)	•			•			•		AVs are considered safer from pedestrians, but similarly safe for passenger car
Liljamo et al., 2018	•			•			•		Safety and moral choices are the top concerns over AV
Woldeamanuel & Nguyen, 2018	•			•			•		Main safety concerns of users are related to AV behaviour in unexpected situations, poor weather or with VRUs
Hudson et al. 2019	•			•				•	Significant differences in attitudes towards AV in Europe based on country, age, gender, education.
Gkartzonikas & Gkritza, 2019	•					•	•		Safety, moral choices and liability are the key concerns for adoption of AV
Moody et al. (2020)	•			•				•	Higher safety perception and optimism in safety-readiness predictions in LMI countries
Goodall, 2014		•				•	•		In addition to rational ethics (basic rules of behaviour) for Avs, Artificial Intelligence may be used to observe/analyze human decision making, when a clear behavioural/moral rule can not be established.
Lin, 2016		•				•	•		The fact that AV moral choices will be systematic, in contrast to human choices that may be based on different moral principles in different drivers/context, brings a need for absolute transparency from the manufacturer's side.
Dennis et al. (2016)		•							Demonstrates a formal verification for ethical behaviour of autonomous systems, on the basis of ranking of ethical choices
Pickering et al. (2017)		•		•				•	Users accept the collision path of the least overall severity, despite the number of persons at risk
Awad et al. (2018)		•		•			•		A large international sample of responses on numerous AV ethical dilemmas; in addition to individual differences in 'preferences', a strong cultural factor was identified with significant differences between western, eastern and southern countries.
Karnouskos (2018)	•	•		•				•	Deontology, Relativism, Absolutism, and Pluralism
Radun et al. (2019)		•		•				•	When a deliberate action of the other vehicle takes place, drivers are not willing to endanger themselves to save the opponent.
Rhim et al. (2020)		•			•			•	Compared a collectivists' sample with an individualists' sample on moral choices and clustered the responses into 3 groups: Moral Altruist, Moral Non-determinist, and Moral-Deontologist
Mordue et al. (2020)		•		•				•	Utilitarianism and "moral machines" (e.g. surveys of user choices over ethical scenarios) based AV choices result in fewer fatalities, but the results vary.
De Bruyne & Werbrouck (2018)			•			•	•		issues such as continuous software updates, transitions of control between human and software.
Noy et al. (2018)			•			•	•		induced error. AV 'errors' may be due to software errors on which very little is known, or simply the lack of human decision making which is based on reasonable expectations from others.
McCall et al, (2018)			•		•				Insurers' survey on responsibility at handovers: a shift in liability towards the OEMs to be considered at high SAE Levels on the basis of with situational awareness requirements and type of handover alert.
Banks et al. (2019)			•		•		•		A shared responsibility over AV from high-level international regulators to intermediate actors (companies) and the end-users can be achieved through a risk management framework model and network-analysis of all interactions between actors.
Calvert et al. (2020b)			•				•		Meaningful Human Control implies not only operational control (i.e. assignment of tasks) but also understanding of the capacities of all agents involved in the driving task, and making all human agents involved in the design and operation aware of their role in the moral consequences of the system's behaviour.

another road user (Gasser et al., 2013).

The way responsibility is still shared between human and machine can be better understood when looking into situations of authority transitions between human and machine, for which there is currently no consensus for what would be considered a safe handover request (McCall et al, 2019) – see following section 4.2.3. This makes attribution of fair moral and legal culpability to drivers more difficult. It also arguably gives regulators an extra responsibility to provide drivers with (new) training schemes to learn a proper interaction with the new technology (Heikoop et al., 2020).

Finally, the EC report also recommends that manufacturers and deployers ensure that the logic behind sensitive decisions made by AVs are transparent and explainable to the public (responsibility as accountability) and point to the regulation of explainable automated decision-making under the GDPR as a promising starting point. The report also stresses the importance of creating a "culture of responsibility" where different actors in the chain of design, regulation, control, and use of AVs are not only made aware of their respective responsibility but also actively supported, i.e. given the capacity and a fair opportunity, to do so (EC Expert Group E03659, 2020).

This analysis shows that safety depends on responsibility in different ways. It critically depends on a clear and fair distribution of obligations, culpability and liability across the different actors in the AVs network. Different human actors should be given a sufficient level of competence, (moral) motivation and power to be considered responsible for the AVs' behaviour ("meaningful human control") (Calvert et al., 2020a), to avoid various "responsibility gaps" (Santoni de Sio & Mecacci, 2021). In general, not only a stronger "safety culture" (NTSB – National Transportation Safety Board, 2019), but also a stronger culture of "responsibility" (capacity to comply with one's obligation) and better "accountability" mechanisms (capacity to explain the behaviour of AVs) should be promoted (EC Expert Group E03659, 2020).

Table 1 summarizes the reviewed studies dealing with ethical issues related to AV safety 'behaviour' and choices. The identified gaps and limitations are provided in section 6.1.

4. Safety engineering and behavioural approaches to AV safety

4.1. Human-mimic behaviour of AVs

There is an antithesis in the way AVs safety behaviour is currently conceived: on the one hand, AVs are aimed to eliminate human error, accepting that humans are imperfect drivers; on the other hand, AVs need to be predictable and compatible with user expectations so that trust can be built, and safety in mixed traffic can be ensured. Sparrow & Howard (2017) argue that AVs ideally could outperform even the most experienced and compliant human driver, while at the same time need not be "perfect drivers" in order to be ethically acceptable; performing better than the average human in a given situation would be sufficient.

One approach to defining AV safety standards is the *data-driven approach*, in which field observations are analysed to derive rules of human behaviour in a particular setting. Sadigh et al. (2019) demonstrate this through inversed reinforcement learning and suggest that the 'robot' can also 'learn' the error distribution of humans, minimizing the probability of replicating these. McAree et al. (2017) proposed an algorithm aimed to translate human-like decision making in trajectory finding with situation-awareness building under uncertainty, into the context of AVs; the crossing of a roundabout was used as a case study. Riaz et al. (2018) developed an algorithm regarding the choices of different AV agents (e.g. vehicle types, manufacturers etc.) for collision avoidance, on the basis of a model of human emotions and social norms; decision making is based on the comparison of 'fear' with 'egoism' – leading to comply with norms or not.

An alternative approach for defining AV safety standards would be '*expert judgment*' resulting from knowledge, experience, and data analysis. In this case, however, there is still room for errors, as the experts are humans themselves and thus prone to errors (Sadigh et al., 2019).

A third approach is discussed by Nyholm & Smids (2020): instead of developing a human-like driving behaviour for robots, or assuming that the robot-like behaviour is ideal, an optimal human-robot coordination should be pursued, in which human driving is made more similar to robot-like driving, e.g. through in-vehicle technologies aiming to prevent common human errors or violations.

4.2. Surrogate measures of safety and thresholds

Automated vehicles are designed to perform certain tasks within an Operational Design Domain (ODD). ODD is defined as "Operating conditions under which a given driving automation system or feature thereof is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain traffic or roadway characteristics." (SAE International, 2018). Exiting this ODD corresponds to the system reaching its boundary, and a human operator needs to take over control. Moreover, within this ODD, AVs need to navigate, detect and comprehend other road users' intentions and maintain safety margins or make safety choices.

In this section, existing safety margins and thresholds for risk assessment of AVs are reviewed in three specific use cases: mixed traffic, VRU encounter, and authority transition. These use cases correspond to the most common accident scenarios and causes, as reviewed by Wang et al. (2020); that study reviewed accident investigation reports from on-road tests of automated vehicles in the USA and in China, and concluded that automation disengagements and safety risks caused by other users (conventional vehicles, pedestrians, cyclists) were the vast majority of AV risks. The review presented in this paper for each case is not meant to be exhaustive, but rather to highlight the key SMoS and their thresholds used to define/measure safety in recent studies, and draw some conclusions on the main practices and gaps that can be identified. Table 2 summarizes the methods and findings of the reviewed studies on all use cases.

4.2.1. Safety of mixed traffic

There are several recent studies dealing with simulation of mixed (conventional and automated) traffic. Most of them are dealing with longitudinal control of platoons with Cooperative Adaptive Cruise Control (CACC), whereas in some cases lateral manoeuvres such as lane changing or exit ramps are included. The typical analysis method is (microscopic) traffic simulation.

Ye & Yamamoto (2019) developed a rule-based cellular autonomous model to predict dangerous encounters in mixed traffic via TTC (Time to Collision) assessment. The results suggest that safety is generally improved with the increase in the AV penetration rate, provided that car-following rules of AVs are more conservative (i.e. larger gaps).

Papadoulis et al. (2019) simulated AV traffic on motorways by introducing an external driver model application, setting rules for longitudinal and lateral control on the basis of adjacent traffic and traffic rules. The AV decision making rules have the dual objective of platoon formation while keeping safe distances (gap threshold taken as 0.6 s). The results suggested large safety improvements in terms of conflicts reduction for increased penetration rate; threshold values for TTC and PET were at 1.5 sec and 5 sec respectively.

In a similar study, Virdi et al. (2019) simulated mixed traffic in different junction types, i.e. signal-controlled, priority, roundabout and diamond intersections. Their AV car-following model imposes speed, deceleration (between −8 and 6 m/s^2) and jerk (0.5 m/s^3) thresholds in lane changing, following and braking actions. The results suggest significant reductions in conflicts as the AV penetration rate increases.

Rahman & Abdel-Aty (2018) developed a similar model with a different car-following algorithm and different thresholds (e.g. acceleration/deceleration between −2.8 and 1 m/s^2, minimum gap at 0.6 sec), and evaluated safety through SMoS such as the standard deviation of

Table 2
Review of surrogate measures of safety (SMoS) and their thresholds for selected AV use cases.

Source	Method					Surrogate measures of safety (SMoS)								Results
	Simulation	Survey	Field	Review/meta-analysis	Subective measures	TTC	Time gap	PET	Acceleration / deceleration rate	Speed difference / speed variability	Number of conflicts	Reaction time / takeover time	Others	
Safety case: CACC and mixed traffic														
Asljung et al. (2016)			•			•							•	Metrics that are not dependent on speed such as harsh braking are more robust for estimating thresholds of risky situations
Rahman & Abdel-Aty (2018)	•					•				•				Managed-lane AV platoons on expressways had superior safety performance than all-lane platoons. A base scenario of no AV platoons had the lowest performance.
Ye & Yamamoto (2019)	•					•			•	•				Increased AV penetration rate results in less speed difference, lower acceleration rates and overall smoother traffic, although capacity is compromised.
Papadoulis et al. (2019)	•					•		•			•			Significant safety benefits in terms of number of conflicts even at low penetration rates, without detrimental effect on traffic flow.
Virdi et al. (2019)	•					•	•	•			•			Overall manual driving conflicts/accidents at highways decreased with increased CAV penetration. But there was evidence that conflicts may increase at signalized or diamond-interchanges with low penetration rates.
Sinha et al. (2020)	•					•		•	•	•				Improvement in network performance may come at the expense of safety performance. AV-manual conflicts may in many cases increase before they decrease, and only at very high penetration rates there is stabilization.
Safety case: AV interactions with VRUs (pedestrians and cyclists)														
Rothenbucher et al. 2016			•										•	Qualitative analysis, no sign of different pedestrian behaviour with AV vs conventional vehicle
Rodriguez Palmeiro et al. (2018)			•				•							No singificant difference between AV and conventional vehicles gaps in all the AV configuations
Nuñez Velasco et al. (2019)	•						•						•	Distance gap is the strongest predictor of pedestrian crossing intentions
Razmi Rad et al. (2020)	•					•								Lower TTC reduces the probability to cross, pedestrians are more likely to interrupt AV flow for larger TTCs
Vlakveld et al. (2020)		•											•	In critical conflicts cylcists where cyclists have "right of way", they yield more often for Avs, less often for conventional vehicles, and even less often for AVs that communicate their intentions.
Nuñez Velasco et al. (2021)	•						•							The gap size and the right of way were the primary factors affecting the crossing intentions of cyclists
Safety case: Transition of control from automation to human														
de Winters et al. (2014)				•								•	•	Reaction time to critical events increases with secondary tasks and higher levels of automation.
Zhang et al (2019)				•			•					•		Takeover time ranges from 0.7 to 20 sec, with mean 2.72 sec. In general, longer time budget results in longer TOT.
Mac Donald et al. (2019)				•								•		Takeover time increases per 0.27 sec for each 1 sec of extra time budget available.
Papadimitriou et al. (2020)				•									•	In addition to TOT, takeover quality is an "umbrella" term for many logitudinal/lateral control or driver behaviour indicators that may reflect the safety of takeovers.

speed, time exposed time-to-collision (TET), time integrated time-to-collision (TIT), time-exposed rear-end crash risk index (TERCRI), and sideswipe crash risk (SSCR). A sensitivity analysis of the TTC threshold ranging from 1 to 3 sec showed no significant difference in the results. Rahman et al. (2018) made a similar study with simulation under poor visibility conditions due to fog.

For a systematic review of recent studies on simulation of AVs at corridor level the reader is referred to Sinha et al. (2020). These authors use the algorithm of Virdi et al. (2019) to simulate AV platoons control and evaluate safety by means of time-to-collision (TTC), post-encroachment time (PET), and relative speed during a TTC. Unlike other studies, the authors report little or no safety benefit for conventional vehicles in all AV penetration rates, compared to the 0% AV penetration rate. In terms of total crash rate, the full benefits of AV can be achieved only for the 100% penetration rate. It is found that reduced TTCs for AVs may result in discomfort for drivers in critical events, defined as those with a PET of <1.5 sec. The study concludes that thresholds for AV 'behaviour', currently taken from previous naturalistic driving studies or standard car-following/simulation tools, may need to be reconsidered for actual AV deployment.

Asljung et al. (2016) suggest that "inevitable collisions states" may be more robust measures for safety assurance of AVs than the traditional conflict-based approaches; for instance, TTC is largely affected by speed and therefore cannot be easily generalized for different situations, while metrics like steering and break response may be more robust.

4.2.2. *Vulnerable road users' encounter*

There have been several studies on the impacts of advanced sensing technology on automated vehicles safety, especially as regards pedestrians. For instance Chen et al. (2021) compared several object detection algorithms – including but not limited to several neural network applications – on the basis of precision, speed and memory consumption. Zhuang et al. (2021) tested an improved and environment-sensitive multispectral network to improve pedestrian detection in challenging illumination and temperature conditions. Pu et al. (2021) proposed an improved algorithm for real-time traffic speed estimation and subsequent traffic mode detection. However, only part of the AV-pedestrian safe interaction can be ensured by means of accurate sensor detection, due to the numerous traffic and human factors involved.

The interactions of non-motorised road users and AVs is based on several components. First, the lack of transparency in the AV 'intentions' may reduce trust and induce stress and confusion to the pedestrian, while over-trust (e.g. AVs always yield for pedestrians) may result in critical conflicts. Moreover, the AV 'behaviour' (expressed by its speed, distance etc.) may affect the pedestrian's situation awareness – e.g. perception, comprehension and projection of the situation (Rodriguez Palmeiro et al., 2018). 'Forward incompatibility', e.g. the potential absence of visual/auditory interaction or other human cues and mannerisms that are common in traffic "negotiations" may reduce trust in AV (Van Loon & Martens, 2015). In this context, the question seems relevant: is mimicking human behaviour safer, or another commonly acceptable AV gap should be established? Or can the "forward incompatibility" be addressed by external HMIs creating room for communication and negotiation between VRUs and AVs?

Rodriguez Palmeiro et al. (2018) conducted a 'Wizard of Oz' experiment in which a pedestrian had to decide to cross the road in front of an approaching AV in different scenarios based on combinations of AV direction, braking behaviour, attention/engagement of the AV 'driver' and recognisability of the AV through signs. The study found no significant effect of any of the combinations on the gap acceptance of pedestrians, although in a post-experiment questionnaire several participants reported having distinguished the differences and been affected by them. The critical gaps measured ranged from 5.66 sec to 7.67 sec, in line with the average gaps accepted by pedestrians when interacting with conventional vehicles.

In 'ghost driver' experiment - where no driver was visible - (Rothenbücher et al., 2016), it was found that pedestrians crossed in front of a (fake) AV the same way that they would cross in front of a conventional vehicle, unless an 'unusual' behaviour of the AV was observed.

In Nuñez Nunez Velasco et al., 2019, the crossing intentions of

pedestrians were studied in a virtual environment, and values of AV speed of 10 km/h and 20 km/h were combined (among other things) with gaps of 2 sec and 4 sec. While the results suggested rather counter-intuitively that lower speeds of the AV resulted in fewer positive crossing intentions, it was found that the distance gap was the most significant predictor of crossing intentions, with the odds of crossing being significantly lower for smaller AV distances.

In another virtual experiment based on agent modelling (Razmi Rad et al., 2020), participants were asked to cross a road while in a hurry to catch a train; in one scenario, AVs were marked as black vehicles, while in another one AVs could use lights to signal their yielding intentions to pedestrians. The study tested three TTC values of 3, 5 and 7 sec. The modelling results showed a reduction of the probability to cross for smaller TTC, while pedestrians were found most likely to interrupt the AV and cross for a TTC of 5 sec.

In Vlakveld et al. (2020), a video experiment displaying cyclist-vehicle conflicts was taken by >1000 participants, comparing early, mid and late decision moments with different types of vehicles (conventional, AV, and AV communicating its intentions,), while controlling for the participants' trust in technology. The results indicated that generally the frequency of yielding increased in early decision moments and with lower trust in AVs. Overall, in situations where participants had priority they yielded less often to conventional vehicles than AVs, but even less often to AVs that communicated its intentions.

Nunez Velasco, (2021) used a 360° video-based virtual reality (VR) method to determine the main factors influencing cyclists' crossing intentions when interacting with an automated vehicle. Four main factors were considered in the study: vehicle type, gap size between cyclist and vehicle, vehicle speed, and right of way. It was found that the gap size and the right of way were the primary factors affecting the crossing intentions of the individuals, while the vehicle type and vehicle speed did not have a significant effect on the crossing intentions.

Overall, the findings suggest that AV 'behaviour' in terms of its visibility/recognition, time/distance gaps maintained and speed will have significant impact on pedestrians' perception, intentions and actual crossing decisions. It is noted that these subjective measures are often used to assess the safety of interactions. It is also shown that diversity of configurations may affect trust and create uncertainty in AV/VRU interactions. A recent *meta*-analysis of external HMIs (eHMIs) that can be used for the AV to "communicate" its level of automation or "behavioural" intentions suggests several knowledge gaps regarding the optimal features of this communication to ensure safe interactions (Dey et al., 2020). Zandi et al. (2020) suggest, on the basis of a survey in six countries, that AV intention messages are more important for pedestrians than AV status signals. Finally, it is suggested that AV systems are not guaranteed to handle safety critical situations when the movements of pedestrians are non-typical (Ondruš et al., 2020).

4.2.3. Authority transitions

Authority transitions between automation and human drivers has received a lot of attention in the literature, and has been identified as a key safety determinant at medium to high levels of automation (levels 3 and 4) (Biondi et al., 2019). In this paper, we focus on handovers that may occur either because of technical failure or because the vehicle has reached the boundaries of its ODD, and may therefore be classified as emergency/non-emergency handovers; human-initiated handovers are out of the scope of this analysis. Of particular interest are the SMoS that are used to assess the takeover performance of human drivers in such automation-initiated takeover alerts.

In this respect, recent reviews and *meta*-analyses have summarized the main determinants and influencing factors of safe transitions (Zhang et al., 2019; McDonald et al., 2019; Papadimitriou et al. 2020). The takeover time budget (TTB) is generally defined as "the time available until the system limit of the automation is reached", while the takeover time (TOT) defined as "the time from the moment of automation disengagement until the first signs of steering corrections or braking behaviour of the driver" (Zhang et al., 2019). The ratio of TOT/TTB is then used as a performance indicator/SMoS for safety of the takeover. McDonald et al. (2019) report other time indicators used to assess takeover performance, e.g. gaze/feet-on/hands-on reaction times, gaze/side-gaze time. A number of other relevant indicators are found in the literature under the general term 'takeover quality' (TQ), including different longitudinal and lateral control indicators, but also driver factors such as hazard perception and situational awareness (McDonald et al., 2019; Papadimitriou et al., 2020).

McDonald et al. (2019) report a range of TTBs used in the literature from 3 to 30 sec. In the *meta*-analysis of Zhang et al. (2019), the mean TOT across 129 studies including various conditions ranged from 0.69 s to 19.79 s, with an average mean of 2.72 s. TOT increases with emergency takeovers, performance of secondary tasks, adverse weather and high traffic density, and decreases with auditory or vibrotactile takeover alert and shorter TTB. McDonald et al. (2019) report a 0.27 s increase in takeover time per a 1 s increase in time budget.

There are no clear conclusions as per the safety-optimal takeover time or TTB and TOT combination. The question becomes critical when taking into account that the time required for a driver to rebuild situational awareness after a disengagement of the automation may range from 4 to 6 sec on average, up to 20 sec in certain contexts (Papadimitriou et al., 2020) and driver response time to critical events significantly worsens at high levels of automation (de Winter et al., 2014).

5. Functional safety and the industry perspective

In the automotive industry, the safety assurance of components and systems is based on the concept of functional safety, i.e. "the specification of functional safety requirements, their allocation to architectural elements and their interaction necessary to achieve safety goals". Safety goals (high-level or task-specific) are associated with an Automotive Safety Integrity Level (ASIL), defined as "one of four levels to specify the item's or element's necessary requirements of ISO 26262 and safety measures to apply for avoiding an unreasonable residual risk" (ISO, 2018). The four levels are denoted as QM (lowest level), and ASIL A, B, C, D (highest level), and are estimated on the basis of severity, exposure and controllability ratings. The approach is based on the HARA (Hazard Assessment and Risk Analysis) method, in which hazards are assessed from a systems perspective and risks are calculated from decomposition methods. A review of these methods is beyond the scope of the present paper, the reader is referred to Khastgir et al. (2017). In the following paragraphs, the functional architectures relevant to AVs safety are reviewed, and a number of key studies are described with regard to the existing approaches for AV safety assurance. Gaps and limitations of the state of the art are analysed in section 6.1.

5.1. Safety in AV functional architectures

A functional architecture refers to the system's logical decomposition into components and sub-components, and the data flows between them. In the automotive industry, it is the basis for the introduction of "intelligence" in the system, reflecting the tasks handled by automation (Behere & Törngren, 2016).

Behere and Törngren (2016) define three types of AV functional components: (i) perception (including sensing, localization, world model, and semantic understanding); (ii) decision & control (including trajectory generation, energy management, diagnosis & fault management, reactive control, and vehicle platform abstraction); and (iii) vehicle platform manipulation (including passive safety, trajectory execution in terms of propulsion, steering, and braking) – in an analogous approach as the one in Calvert et al. (2020a) where different components of the whole traffic system are relevant for safety and responsibility. Wang et al. (2020) indicate a typical AV system architecture as (i) perception layer, (ii) decision layer and (iii) action layer).

Typically, the architecture(s) are tested using scenarios (use cases)

upon which they are validated, and gaps breaking the architecture are identified (Behere et al., 2013). Several studies stress the difficulty in identifying all the possible degradation processes that may be relevant to AVs within the architecture and the need for more functional safety-specific applications (Behere & Törngren, 2016). Behere et al. (2013) note that safety is often seen as a non-functional property of the system, and in most cases, it relies upon the approach taken by the system architect.

Mauborgne et al (2017) discuss how the functional safety concept can be integrated into the system's logical architecture. The (sub) system's logical architecture is drafted, including its dysfunctional failure modes and critical propagation paths. As a second step, safety functions are created to describe the behavioural view of the architecture. As a third step, alternative functions are provided. All steps and activities can be linked to safety goals. The example of unintended acceleration of a vehicle is used for that purpose.

5.2. Safety assurance

Khastgir et al. (2017) argue that the ISO 26262 lacks concrete guidelines for implementation, inducing uncertainty and concerns on reliability and validity. They report on expert ratings based on a dedicated HARA over two collision scenarios of a low-speed autonomous pod: pedestrian collision and unintended braking. The study drafted parametrized rules for severity assessment of these scenarios based on vehicle/oncoming obstacle velocity and type of obstacle, and controllability assessment based on deceleration value, distance to obstacle, TTC, and vehicle velocity. While initial results showed variation in ASIL ratings, eventual convergence after a certain number of rounds was achieved.

Sini & Violante (2020) suggest that using data from driving simulator experiments may reduce the subjectiveness of expert ratings or the experts' limited previous knowledge, which are considered significant drawbacks of ASIL determination. The authors tested AEB (Autonomous Emergency Braking) based on scenarios obtained from NHTSA, EuroNCAP, and European Commission reports. They used thresholds from the literature on the relative speed between the test vehicle and the target vehicle for the collision severity, and TTC for the controllability, while the exposure was assigned manually.

Saraoglu et al. (2019) developed a simulation model with 'fault injection' at the vehicle front sensor and speed sensor components in order to test the safety of AV driving systems, using the functional architecture of Behere and Törngren (2016). The model allows for fault-failure-error analysis along the entire chain from the component level to the vehicle level and the traffic level. Based on data from real-life collisions, the authors propose severity threshold metrics of 1.4 sec time headway (2.5 m distance) in AV platoons and deceleration of $-7 m/s^2$ for emergency braking. Their simulations show that a single AV's sensor failures may affect traffic safety (with or without the occurrence of crash) in different ways, depending on the failure and traffic conditions' timing.

A multi-agent simulation tool based on a belief-desire-intention model was presented by Kamali et al. (2017) to verify the safety of vehicle platooning. The study goes in-depth into the AI algorithm of vehicle control but is based on a general safety principle according to which "a vehicle can join and leave a platoon if, and only if, the whole platooning remains safe". The authors conclude that the actual AI code needs verification, and the system perspective might not be adequate for all critical situations.

Elgharbawy et al. (2019) discuss AV safety assurance for heavy vehicles and note that functional safety assurance is mostly targeted at higher levels of automation, in which the driver cannot over-ride the system in case of failure. The authors suggest that a combination of an (iterative) simulator, laboratory, and field testing is needed in order to result in only a negligible residual risk due to minor imperfection of sensors.

An earlier project FUSE Consortium (2016) proposed methods to gradually reduce the risk of missing hazardous events within the HARA analysis of ISO26262. Interestingly, the project suggests that AVs safety assurance may, in fact, eliminate "trolley-type" problems, as long as the functional requirements allow the AV AI systems to "anticipate" well in advance these situations.

Burton et al. (2017) note that "the safety standard is limited to avoiding potentially safety–critical situations caused by systematic software and random hardware failures. Safety violations due to technological and system-technical deficiencies remain outside the scope of ISO 26262:2018 (e.g., insufficient robustness, uncertainty issues with perception sensors, etc.)".

In a recent review of research on AV safety assurance, Batsch et al. (2020) review relevant testing strategies with regard to a number of specific tasks (modelling capabilities, automation subsystem, driving task level and the metrics for safety evaluation) and conclude that testing strategies are usually targeted as specific tasks, and the lack of appropriate metrics limit the testing capabilities for complex scenarios.

5.3. Standardisation

Technology standardisation is the scientific discipline studying the dynamics of standardisation processes and their consequences, both from a theoretical and empirical perspective (David & Greenstein, 1990). Standardisation may follow two processes (Wiegmann et al., 2017). *De jure* standardisation refers to a process whereby actors (mostly firms) come together in standardisation organizations. The coordination mechanism here is cooperation (Van de Kaa & De Bruijn, 2015). *De facto* standardisation refers to a situation whereby standards or products are developed by multiple companies that are fighting for market dominance. The coordination mechanism here is competition (van de Kaa et al. 2011). It can be understood that both processes are active at present towards the standardisation of AV safety: on the one hand, national and international governments have initiated stakeholder consultations, and standardisation bodies have issued some technical standards. On the other hand, AV manufacturers define their own safety features and engage in the competition, thereby creating de facto standards that may be neither homogeneous nor safety-proof.

Li et al. (2019) reviewed policy developments regarding AVs and concluded that while research has been conducted in the fields of governmental, legal, ethical, licensing/testing and certification, the lack of a common framework for AV deployment may have significant impacts on eventual AV safety and security.

As shown in the previous section, the concept of verification of safety includes both the quantification of safety and the demonstration that all hazardous situations can be handled by the AV, in a challenging context where all systems are continuously active. However, field testing of all possible AVs scenarios is economically unfeasible, while the simulation of these situations requires their full coverage via adequate metrics and prior knowledge (Asljung et al., 2016). Results from the New Car Assessment Programme (Euro NCAP, 2018), where current vehicle systems are evaluated and assigned star ratings, revealed that none of the level 2 systems could offer adequate assistance in all scenarios and tests, and the lack of knowledge about the systems design and limitations, as well as the lack of common safety goals, does not allow for meaningful star ratings of automated vehicles.

Eventually, formal safety verification could be done via online performance assessment against commonly agreed standards or against a 'library' of verified scenarios and test cases (Schwarting et al., 2018). Junietz et al. (2019) further suggest that "unexpected critical situations and near misses must be monitored similar to air traffic, in order to find flaws in the system (including infrastructure and human interaction) with the chance to improve them".

Cummings & Britton (2020) review the regulation approach to autonomous systems in different sectors (aviation, medical, automotive) in the USA and note that the approach taken for AVs is the least conservative, with NHTSA's approach assuming a priori the safety of new

technological features on a car, unless shown to induce unreasonable risks *after* introduction into the marketplace.

Junietz et al. (2019) also underline that, despite ISO26262 and its subsequent specification for safety of the intended function (SOTIF), there is still uncertainty in the safety verification of AV components and the system as a whole, and note that the uncertainty will remain during their introduction on the roads.

There are several recent or ongoing projects, mostly within automotive industry partnerships, dealing with AV deployment standardisation, including safety features. For instance, the earlier project CityMobil (Van Dijke et al., 2012) defined a procedure for the certification of automated transport systems based on comparison with the current level of risk of conventional vehicles. The authors provide results from early trials in Italy, France, and the UAE. The ARCADE project brings together national and European stakeholders involved in AV deployment to ensure synergies' cooperation and exploitation (Arrúe et al., 2018).

The Pegasus project used a fusion of data from accident databases, simulator trials and naturalistic driving studies, to generate scenarios of human performance in critical situations, in order to define acceptable AV 'behaviours' and safety margins for validation and verification of AVs. These logical scenarios are described based on a six-layer model (road, infrastructure, temporary influences, movable objects, environment conditions, digital information). An algorithm is used to extract minimal and maximal parameter values to describe parameter ranges within the logical scenarios. These criteria are represented by metrics, such as TTC, THW, etc. Subsequently, test cases are performed in simulation or field tests, in which performance is compared to the specified criteria. The final step is creating a "Safety Argumentation" framework for the securing and approval of high-level AVs, within an overall architecture. The method aims to serve as the first proof of concept that will be continuously evolving with additional use cases and ODDs.

Shladover & Nowakowski (2019) summarize experience from the California PATH program and underline that "in the absence of clearly defined standards and testing procedures, it is not yet clear how safety can be certified by the developer, a regulatory agency, or a third party". The authors suggest four potential targets for certification:

(i) The manufacturer's functional safety system development through ISO26262 or similar.

(ii) The manufacturer's specific system design's functional safety through an intensive hazard analysis and mitigation for all components.

(iii) Testing the performance of the automated driving system relative to required "behavioural competencies": there are enormous challenges for achieving adequate testing time, scenario coverage and a representative population of drivers (including "extreme" AV encounters).

(iv) Simulating the behavioural competency and system performance under a more comprehensive range of scenarios and conditions than would be practical for on-road or test track testing - in this case, however, validity concerns would be raised, and the simulation would need a certification process itself.

6. An interdisciplinary approach for AV safe behaviour

6.1. Critical assessment of existing research

The above review aimed to span a set of disciplines relevant to the research question of how to develop acceptable ethical and safety standards for the "behaviour" of AVs. Ethical issues are involved in AVs' design, both from the individual, societal, and normative perspective. There are certain expectations at the societal level as per the safety impact of AVs, together with a degree of uncertainty and mistrust on the "promises" made by policy and industry stakeholders. These are reflected in the controversies of the proposed approaches to AVs' actual ethical choices in certain critical situations (descriptive ethics). These ethical choices seem to be pluralistic: they can be affected by demographic, socioeconomic, cultural or contextual factors. Most importantly, some studies have demonstrated how AI algorithms' different safety configurations can affect the final outcomes in terms of traffic victims (e.g. Mordue et al., 2020). On the other hand, different general objections to the "trolley problem" approach are: that they should be avoided 'by design' so that AVs will never be found in such situations and, in any case, that they should play a less central role in ethical discussions, as other issues are more urgent and relevant for AVs safety.

The literature on moral dilemmas with AVs show that engineering approaches to safety choices cannot evade the issue of choosing the (ethical) principles to be followed. Different approaches lead to different results. While laypersons tend to disagree on the principles to be followed, ethics and legal experts notice that some of these choices are more or less compatible with deeply seated shared ethical and even legal principles of our institutions, and in any case recommend to promote a public, critical and inclusive deliberation on these norms, values and principles as opposed to leave them to the decision of (individual) manufacturers or users.

Nevertheless, AVs are still widely expected to make similar choices as an "ideal" human driver, although in practice, all drivers may exhibit different driving choices in different circumstances. Several studies show that an AV behaviour in line with user expectations will enhance trust in automation and ensure safety encounters in mixed traffic; therefore, a data-driven approach to identifying distributions and margins of safe human behaviour is often recommended. On the other hand, this behaviour may not necessarily comply with scientific standards based on acceptable safety margins in given situations, which may raise ethical concerns as to the legitimacy of such an approach.

Currently, simulation and field studies on AVs mixed traffic adopt a set of AV navigation rules and define or identify relevant thresholds or reference values of SMoS in particular situations. However, there are no commonly accepted values of these thresholds. It is interesting to note that in most studies AVs reaction time is assumed to be zero (Ye & Yamamoto, 2019). The typical TTC taken for safety purposes is ~1sec, while smaller values are assumed in CACC car following situations. The majority of studies suggest impressive safety improvements with increased AV deployment. In addition, most of these simulation studies assume that human drivers will interact with AVs in the same way as they interact with other human driven vehicles, i.e. lack of behavioural adaptation. Therefore, future studies need to test this assumption.

Several recent studies have been conducted on AVs' safety behaviour when encountering non-motorized traffic, i.e., VRUs. The results suggest a more conservative attitude of pedestrians and cyclists towards AVs compared to conventional vehicles; however, in several cases the findings did not reach statistical significance. There is a variation in the time or distance gaps accepted by pedestrians (as is the case with conventional vehicles), and further research is needed to conclude a preferred gap maintained by the AV.

As also Kolekar et al. (2020) underline, a fragmentary approach is mostly used to analysing AV safety situations: selected use cases are modelled, focusing either on longitudinal or lateral movement, but seldom both. The authors indicate that this bottom-up approach will require identifying all possible scenarios and integrating human decision-making principles in all of them – resulting in an entirely unfeasible task.

Most importantly, the ethical, behavioural and traffic engineering studies dealing with the above questions are of entirely different methodological and conceptual frameworks than those used by the AV manufacturers and AI developers. A very recent study (Martinho et al., 2021) reviewed ethical issues presence in industry reports and found that, while the industry is aware of and concerned about moral dilemmas in 'extreme' situations, more focus is placed on safety and cybersecurity issues. The industry is primarily driven by the concept of functional safety, as defined in ISO26262 or relevant standards. Functional safety may concern individual failures of sensors, connectivity or

security issues – it is therefore mostly relevant to connected vehicles and their communication with other vehicles, pedestrians, and environment (V2I, V2P, V2X) - but the impacts are seldom assessed at the traffic system level (i.e. including all road users) or at socio-technical system level (i.e. including all actors: regulators, designers, managers).

Therefore, while this approach is proved useful for ADAS and lower automation levels, where individual systems control specific functions, the safety verification of higher automation levels will be much more challenging because system architectures and the numbers of embedded interacting components become immense. At the highest levels of automation, safety is often perceived as merely a software verification task. However, sensor or algorithm failures are only a part of potential causes of AV crashes, and numerous other crash causation factors e.g. traffic and human factors, may be involved. Again, there are no commonly agreed ethical and safe "behaviours" to be reflected in the sensor capabilities and AI algorithms, and the safety goals are often very aggregate. At this stage of AV development, safety assurance is mostly placed on the manufacturers, however, this responsibility is disproportional once the size of potential safety consequences is considered.

6.2. *Framework for further research*

In light of the gaps presented in the previous section, we propose a new framework towards common standards to be acceptable by all stakeholders: (i) for AV moral choices with shared or democratically accepted moral and legal standards, (ii) for AV "behaviour" and relevant safety thresholds in critical situations, and (iii) for safety assurance/verification procedures targeted to the agreed safety "behaviour" rather than the reliability of technical implementation.

The proposed framework is depicted in Fig. 2, together with the concrete steps and research goals that need to be addressed for its implementation. The scientific disciplines involved are displayed as three vertical pillars (ethics, safety & standardisation), while the interdisciplinary research goals are displayed as horizontal boxes spanning different disciplines. For each of these goals, we recommend a set of specific methodologies and tools with high potential for achieving the goals.

First, an interdisciplinary state of the art review is proposed, in order to learn from safety standardisation experiences in other sectors; examples may include (but are not limited to) aviation, traffic signs and signals, telecommunications, cybersecurity protocols.

Subsequently, research should tackle AVs safety standards at different layers/disciplines:

At the ethical layer, research should further explore the safety values, underlying ethical trade-offs, and expectations regarding AV 'behaviour', aiming to a set of ethical guidelines and safety design principles (including vehicle failure conditions), both at a macroscopic level (i.e., values and norms) and at microscopic level (i.e. choices, dilemmas, etc.). These questions are usually tackled using questionnaire survey data or 'moral machines' experiments, however we suggest that broader ethical and philosophical research is required, both theoretical and empirical. Moreover, the interaction between actors is seldom taken into account in previous studies; therefore, a promising method to consolidate the acceptable trade-offs may include serious gaming applications. Also, norms and values should be discussed and shaped by a broad critical, inclusive deliberation process, and not only be observed and extracted from empirical studies about people perceptions and attitudes.

At the safety layer, as a next step, research should analyse more systematically the variability in human driving behaviours with respect to the agreed ethical principles, user capabilities and existing safety standards, in order to define acceptable safety thresholds of relevant SMoS in various scenarios (EC Expert Group E03659, 2020). In this paper we suggest a combination of approaches, in contrast to the 'binary' current practice. On the one hand, simulated or naturalistic driving data will be useful for that purpose in order to identify human driving patterns that correspond to optimal safety (data-driven

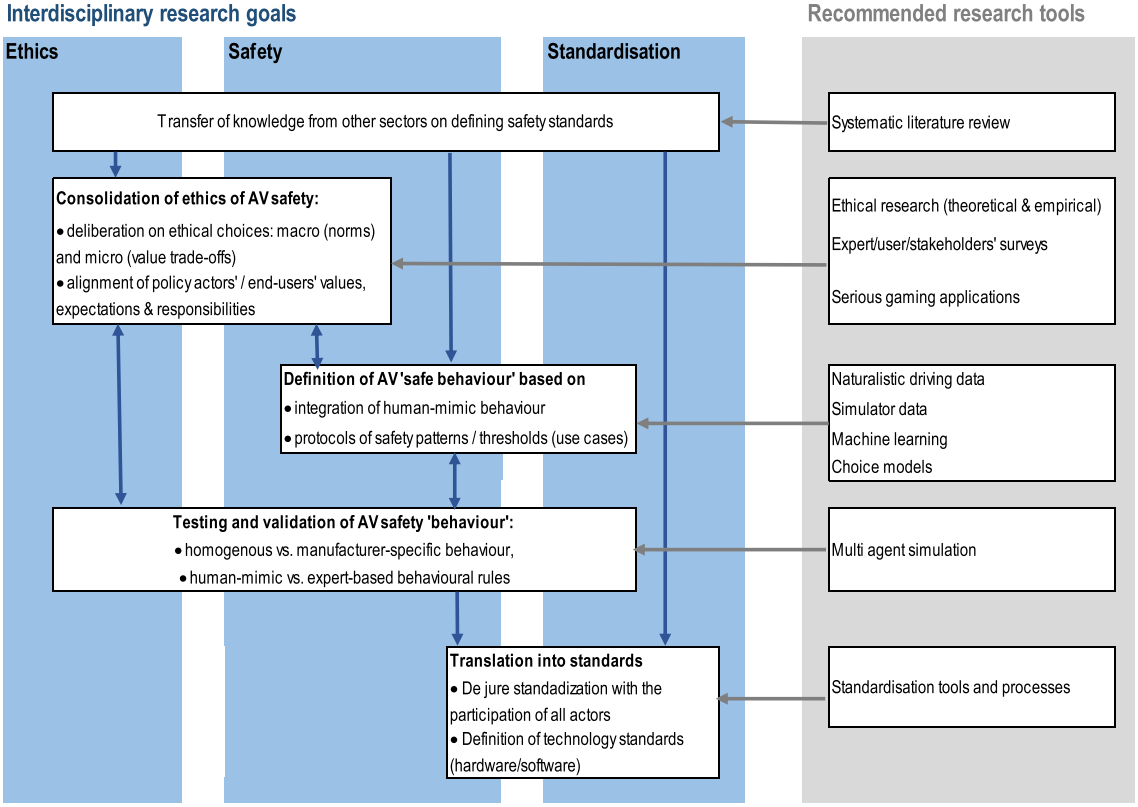

Fig. 2. Proposed interdisciplinary framework of research goals and tools for commonly accepted AV safe behaviour standards.

approach). At the same time, expert knowledge needs to be explicitly integrated in the process to ensure scientific robustness and acceptability. The EU underlines the importance of open data policies with respect to high-value datasets related to AVs (e.g. spatial data, satellite data, weather data, data on crash or near-crash situations including and not including AVs, and data on mobility & traffic patterns), as a means to ensure their ethical and safety conscious deployment (EC Expert Group E03659, 2020). However, data from field tests of industry actors remain inaccessible, and concrete steps should be taken by regulators and all parties involved to increase their availability.

The next research step is to further study how both ethical and safety design principles can be implemented in standards at the standardisation layer. Existing standards should be adapted to aim at a more macroscopic level of standardisation than that of technical components, including the specific and commonly agreed safety goals and thresholds identified per function. This approach can form the basis for actual standardisation of AV safety, with all relevant stakeholders' participation. In this framework, the role of international stakeholders involved in vehicle safety assessment, e.g. the New Car Assessment Programme can play a crucial role in defining safety standards, also because their vehicle testing and star-rating programmes may have a substantial impact on consumer preferences, and this may significantly influence the design and deployment of new technologies from the manufacturers.

Summarizing, the main research methods and tools recommended within this framework are:

(i) methodologies for defining safety in the context of AVs, that can be useful to researchers and industry partners studying the AV safety in various ODDs, and also for international regulators and standardisation institutes who need to know what ethical and safety aspects to take into account in the standards they develop.

(ii) protocols for safety thresholds in the 'behaviour' of AVs in selected use cases/ODDs, which can be readily implemented in simulation and field testing for AV safety assessment.

(iii) simulation tools for testing the safety benefits of various scenarios of standardised AV design (against heterogeneous and "unpredictable" behaviour scenarios), that can be used by researchers and practitioners involved in AV safety assessment projects, as well as authorities involved in the development of AV safety standards.

7. Conclusions

This paper reviewed existing approaches to AVs safety from an ethical, safety (engineering, behavioural) and standardisation perspective. The paper takes an interdisciplinary approach – in contrast to previous review studies that focus either on ethical or on safety issues – and establishes for the first time a robust link with the scientific field of standardisation.

From this review, it was concluded that there is no common understanding of AV safety; on the contrary, there is little recognition of safety as a distinct design value in AVs research and development projects. There are useful methods and concepts within each discipline, however a 'silos' effect is mostly in place, as safety is examined from entirely different perspectives, ranging from ethical to purely technical. The results reveal the need for a common and interdisciplinary approach to AV safety, with two main objectives: (i) a complete definition of safety in the context of AVs, and (ii) a structured transfer of knowledge between sectors and relevant stakeholders.

The framework proposed in this research uses a "comprehensive engineering" approach, which combines three different perspectives for analysing complex societal challenges: values, systems and governance (Weijnen & Herder, 2018). This approach although not exhaustive, may allow a convergence of AV safety goals, testing objectives, expectations and assessment frameworks, resulting in new opportunities for researchers, industry partners, road authorities, cities, standardisation authorities etc. This approach's added value is that it contributes to a smoother and more responsible path to the desired safety level of AVs, so that the ambitious AV safety vision can be achieved, in terms of smooth coexistence of AVs with conventional drivers and non-motorized road users, increased trust in AVs, and fewer road casualties.

The present study has some limitations. Due to the large number of papers available in each of the disciplines examined, we carried out a targeted, rather than a systematic literature review. The proposed framework is based on the gaps identified in the literature, and the expertise of our interdisciplinary team. A much broader consultation would certainly allow to fine-tune the proposed research goals and tools. Moreover, the full potential of the proposed framework needs to be pilot tested with actual datasets relevant to each research goal (e.g. from AV use cases, stakeholder surveys).

It should be noted that in the literature, it is often suggested that standards should remain constant. However, ethical values and safety aspects that society finds essential are not always known at the time that the standard is defined (see e.g. van de Poel, 2018). Therefore, it is also necessary to study whether standards' flexibility in terms of the incorporation of ethical values throughout their life cycle will affect the standards' acceptance and AV systems' adoption, and how broad stakeholder support will affect flexibility and adoption.

There are several complicating factors in this respect; on the one hand, the current lack of a common approach requires a change of culture and perspective by researchers, practitioners, and policy/industry stakeholders. On the other hand, it will be challenging to converge the various safety aspects involved in the numerous initiatives currently in place in Europe and beyond. International coordination of critical Institutes and authorities involved in AV research and dedicated consultations will facilitate the fusion of needs, strategies, and safety perceptions.

CRediT authorship contribution statement

Eleonora Papadimitriou: Conceptualization, Investigation, Methodology, Writing – original draft. **Haneen Farah:** Conceptualization, Methodology, Writing – review & editing. **Geerten van de Kaa:** Conceptualization, Methodology, Writing – review & editing. **Filippo Santoni de Sio:** Conceptualization, Methodology, Writing – review & editing. **Marjan Hagenzieker:** Conceptualization, Methodology, Writing – review & editing. **Pieter van Gelder:** Conceptualization, Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Arrúe A., van Montfort S., Kotte J., Maes J., Dugernier G., Rutten B., Dirnwoeber M. (2018). Position Paper on Safety validation and roadworthiness testing. CARTRE – Coordination of Automated Road Transport Deployment for Europe. Brussels, October 2018. Available online at: https://connectedautomateddriving.eu/wp-content/uploads/2018/10/CARTRE-Roadworthiness-Testing-Safety-Validation-position-Paper_3_After_review.pdf [Accessed June 2020].

Asljung, D., Nilsson, J., Fredriksson, J., 2016. Comparing collision threat measures for verification of autonomous vehicles using extreme value theory. IFAC-PapersOnLine 49–15 (2016), 057–062.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., Rahwan, I., 2018. The moral machine experiment. Nature 563 (7729), 59–64.

Banks, V.A., Stanton, N.A., Plant, K.L., 2019. Who is responsible for automated driving? A macro-level insight into automated driving in the United Kingdom using the Risk Management Framework and Social Network Analysis. Appl. Ergon. 81 (2019), 102904.

Batsch, F., Kanarachos, S., Cheah, M., Ponticelli, R., Blundell, M., 2020. A taxonomy of validation strategies to ensure the safe operation of highly automated vehicles. Journal of Intelligent. Transp. Syst. 6, 1–20.

Behere, S., Törngren, M., Chen, D.-J., 2013. A reference architecture for cooperative driving. J. Syst. Archit. 59 (2013), 1095–1112.

Behere, S., Törngren, M., 2016. A functional reference architecture for autonomous driving. Inf. Softw. Technol. 73, 136–150.

Biondi, F., Alvarez, I., Jeong, K.A., 2019. Human-vehicle cooperation in automated driving: a multidisciplinary review and appraisal. Int. J. Hum.-Comput. Interact. 35 (11–15), 1–15.
Bonnefon, J.F., Shariff, A., Rahwan, I., 2019. The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars. Proc. IEEE 107 (3), 502–504.
Burton, S., Gauerhof, L., Heinzemann, C., 2017. In: Making the case for safety of machine learning in highly automated driving. and Security. Springer, pp. 5–16.
Calvert, S.C., Heikoop, D.D., Mecacci, G., van Arem, B., 2020a. A human centric framework for the analysis of automated driving systems based on meaningful human control. Theor. Issues Ergon. Sci. 21 (4), 478–506.
Chen, L., Lin, S., Lu, X., Cao, D., Wu, H., Guo, C., Liu, C., Wang, F.-Y., 2021. Deep neural network based vehicle and pedestrian detection for autonomous driving: a survey. IEEE Trans. Intell. Transp. Syst. 22 (6), 3234–3246.
Cummings, M.L., Britton, D., 2020. Regulating safety-critical autonomous systems: past, present, and future perspectives. In: Pak, R., de Visser, E.J., Rovira, E. (Eds.), Living with Robots: Emerging Issues on the Psychological and Social Implications of Robotics. Academic Press, pp. 119–140.
Dankd, D., London, A.J., 2017. Regulating autonomous systems: beyond standards. IEEE Intell. Syst. 32 (1), 88–91.
David, P.A., Greenstein, S., 1990. The Economics Of Compatibility Standards: An Introduction To Recent Research 1, 3–41.
Dey, D., Habibovic, A., Löcken, A., Wintersberger, P., Pfleging, B., Riener, A., Martens, M., Terken, J., 2020. Taming the eHMI jungle: a classification taxonomy to guide, compare, and assess the design principles of automated vehicles' external human-machine interfaces. Transp. Res. Interdiscip. Perspect. 7, 100174.
De Bruyne, J., Werbrouck, J., 2018. Merging self-driving cars with the law. Comput. Law Sec. Rev. 34 (2018), 1150–1153.
Dennis, L., Fisher, M., Slavkovik, M., Webstera, M., 2016. Formal verification of ethical choices in autonomous systems. Rob. Auton. Syst. 77 (2016), 1–14.
Elgharbawy, M., Scherhaufer, I., Oberhollenzer, K., Frey, M., Gauterin, F., 2019. Adaptive functional testing for autonomous trucks. Int. J. Transp. Sci. Technol. 8 (2), 202–218.
Euro NCAP (2018). 2018 Automated driving tests. Retrieved from: https://www.euroncap.com/en/vehicle-safety/safety-campaigns/2018-automated-driving-tests/.
ETSC, 2016. Prioritising the Safety Potential of Automated Driving in Europe. European Transport Safety Council, Brussels.
FUSE Consortium (2016). FUnctional Safety and Evolvable architectures for autonomy. Project Summary. Available online at: https://esplanade-project.se/fuse/FUSE_Leaflet_v2.pdf [Accessed June 2020].
Gasser, T.M., Arzt, C., Ayoubi, M., Bartels, A., Bürkle, L., Eier, J., Flemisch, F., Häcker, D., Hesse, T., Huber, W., Lotz, C. (2013). Legal consequences of an increase in vehicle automation. Bundesanstalt für Straßenwesen [Online]. Available:<http://bast.opus.hbz-nrw.de/volltexte/2013/723/pdf/Legal_consequences_of_an_increase_in_vehicle_automation.pdf>.
Gkartzonikas, C., Gkritza, K., 2019. What have we learned? A review of stated preference and choice studies on autonomous vehicles. Transp. Res. Part C 98 (2019), 323–337.
Goodall, N.J., 2014. Ethical decision making during automated vehicle crashes. Transp. Res. Rec. 2424 (1), 58–65.
Heikoop D.D., Calvert S.C., Mecacci G., Hagenzieker (2020). A practitioner's View of Driver Training for automated driving From Driving Examiners: A Focus group Discussion. 2020 Forum on Integrated and Sustainable Transportation Systems (FISTS), Delft, Netherlands, 2020, pp. 14-19.
Hudson, J., Orviska, M., Hunady, J., 2019. People's attitudes to autonomous vehicles. Transp. Res. Part A 121 (2019), 164–176.
Hulse, L.M., Xie, H., Galea, E.R., 2018. Perceptions of autonomous vehicles: relationships with road users, risk, gender and age. Saf. Sci. 102, 1–13.
ISO (2018). ISO 26262–1:2018, Road vehicles - Functional safety. 2nd Edition. Geneva, 2018.
EC Expert Group to advise on specific ethical issues raised by driverless mobility (E03659). Ethics of Connected and Automated Vehicles: recommendations on road safety, privacy, fairness, explainability and responsibility. 2020. Publication Office of the European Union: Luxembourg.
ITF, OECD, 2018. Safer Roads with Automated Vehicles? ITF, Paris.
Junietz, J., Steininger, U., Winner, H., 2019. Macroscopic safety requirements for highly automated driving. Transp. Res. Rec. 2673 (3), 1–10.
Kamali, M., Dennis, L.A., McAree, O., Fisher, M., Veres, S.M., 2017. Formal verification of autonomous vehicle platooning. Sci. Comput. Program. 148 (2017), 88–106.
Karnouskos, S., 2018. Self-driving car acceptance and the role of ethics. IEEE Trans. Eng. Manage.
Khastgir, S., Birrell, S., Dhadyalla, G., Sivencrona, H., Jennings, P., 2017. Towards increased reliability by objectification of Hazard Analysis and Risk Assessment (HARA) of automated automotive systems. Saf. Sci. 99 (2017), 166–177.
Koopman, P., Wagner, M., 2017. Autonomous vehicle safety: an interdisciplinary challenge. IEEE Intell. Transp. Syst. Mag. 9 (1), 90–96.
Kolekar, S., de Winter, J., d., 2020. Human-like driving behaviour emerges from a risk-based driver model. Nat. Commun. 11, 4850.
Li, S., Sui, P.-C., Xiaoa, J., Chahine, R., 2019. Policy formulation for highly automated vehicles: emerging importance, research frontiers and insights. Transp. Res. Part A 124 (2019), 573–586.
Liljamo, T., Liimatainen, H., Pöllänen, M., 2018. Attitudes and concerns on automated vehicles. Transportation Research Part F 59 (2018), 24–44.
Lin, P., 2016. Why ethics matters for autonomous cars. Autonomous driving. In: Maurer, M., Gerdes, J.C., Lenz, B., Winner, H. (Eds.), Autonomous Driving. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 69–85.
Lin, P., (2013). The Ethics of Autonomous Cars [online]. Available at: [Accessed 2nd July 2017].
Luetge, C., 2017. The German ethics code for automated and connected driving. Philos. Technol. 30 (4), 547–558.
Martinho, A., Herber, N., Kroesen, M., Chorus, C., 2021. Ethical issues in focus by the autonomous vehicles industry. Transp. Rev. 41 (5), 556–577.
McDonald, A.D., Alambeigi, H., Engström, J., Markkula, G., Vogelpohl, T., Dunne, J., Yuma, N., 2019. Toward computational simulations of behaviour during automated driving takeovers: a review of the empirical and modelling literatures. Hum. Factors 61 (4), 642–688.
Massachusetts Institute of Technology, 2016. Moral machine. Retrieved June 25, 2019 from. http://moralmachine.mit.edu/.
Mauborgne, P., Deniaud, S., Levrat, E., Bonjour, E., Micaëlli, J.-O., Loise, D., 2017. The determination of functional safety concept coupled with the definition of logical architecture: a framework of analysis from the automotive industry. IFAC PapersOnLine 50–1 (2017), 7278–7283.
McAree, O., Aitken, J.M., Veres, S.M., 2017. Towards artificial situation awareness by autonomous vehicles. IFAC PapersOnLine 50–1 (2017), 7038–7043.
McCall, R., McGee, F., Mirnig, A., Meschtscherjakov, A., Louveton, N., Engel, T., Tscheligi, M., 2019. A taxonomy of autonomous vehicle handover situations. Transp. Res. Part A 124 (2019), 507–522.
Millar, J.. An ethical dilemma: When robot cars must kill, who should pick the victim? Communication platform that brings together experts in robotics research. (2014). Retrieved March 17, 2019 from https://robohub.org/an-ethical-dilemma-wh enrobot- cars-must-kill-who-should-pick-the-victim/11.
Moody, J., Bailey, N., Zhao, J., 2020. Public perceptions of autonomous vehicle safety: an international comparison. Saf. Sci. 121 (2020), 634–650.
Mordue, G., Yeung, A., Wu, F., 2020. The looming challenges of regulating high level autonomous vehicles. Transp. Res. Part A 132 (2020), 174–187.
NTSB – National Transportation Safety Board (2019). Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018. NTSB/HAR-19/03 PB2019-101402. (Available on-line at: https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf, Accessed February 2021).
Noy, I.Y., Shinar, D., Horrey, W.J., 2018. Automated driving: Safety blind spots. Saf. Sci. 102 (2018), 68–78.
Nunez Velasco, J.P., Farah, H., van Arem, B., Hagenzieker, M.P., 2019. Studying pedestrians' crossing behaviour when interacting with automated vehicles using virtual reality. Transp. Res. Part F 66 (2019), 1–14.
Nunez Velasco, J.P., 2021. Should I Stop or Should I Cross: Interactions between vulnerable road users and automated vehicles. PhD dissertation. Delft University of Technology.
Nyholm, S., Smids, J., 2020. Automated cars meet human drivers: responsible human-robot coordination and the ethics of mixed traffic. Ethics Inf. Technol. 22, 335–344. https://doi.org/10.1007/s10676-018-9445-9.
Rodriguez Palmeiro, A.R., van der Kint, S., Vissers, L., Farah, H., de Winter, J.C., Hagenzieker, M., 2018. Interaction between pedestrians and automated vehicles: a Wizard of Oz experiment. Transp. Res. Part F: Traffic Psychol. Behav. 58, 1005–1020.
Ondruš, J., Kolla, E., Vertaľ, P., Šarić, Z., 2020. How do autonomous cars work? LOGI 2019 – Horizons of autonomous mobility in Europe. Transp. Res. Procedia 44 (2020), 226–233.
Papadimitriou, E., Schneider, C., Aguinaga Tello, J., Damen, W., Lomba Vrouenraets, M., ten Broeke, A., 2020. Transport safety and human factors in the era of automation: what can transport modes learn from each other? Article in Press. Accid. Anal. Prev. 144, 105656.
Papadoulis, A., Quddus, M., Imprialou, M., 2019. Evaluating the safety impact of connected and autonomous vehicles on motorways. Accid. Anal. Prev. 124 (2019), 12–22.
PEGASUS Consortium (2020). Pegasus method: An Overview. Available online at: https://www.pegasusprojekt.de/en/pegasus-method [Accessed June 2020].
Pickering, J.E., Podsiadly, M., Burnham, K.J., 2019. A model-to-decision approach for the autonomous vehicle (AV) ethical dilemma: AV collision with a Barrier/Pedestrian(s). IFAC-PapersOnLine 52 (8), 257–264.
Pu, Z., et al. (2021). Multi-modal traffic speed monitoring: a real-time system based on passive Wi-Fi and bluetooth sensing technology. IEEE Internet Things J. (2021).
Radun, I., Radun, J., Kaistinen, J., Olivier, J., Kecklund, G., Theorell, T., 2019. Endangering yourself to save another: a real life ethical dilemma. Transp. Res. Part F 64 (2019), 318–322.
Rahman, M.S., Abdel-Aty, M., Wang, L., Lee, J., 2018. Understanding the highway safety benefits of different approaches of connected vehicles in reduced visibility conditions. Transp. Res. Record 2672 (19), 91–101.
Rahman, M.S., Abdel-Aty, M., 2018. Longitudinal safety evaluation of connected vehicles'platooning on expressways. Accid. Anal. Prev. 117, 381–391.
Razmi, S., de Almeida Correia, G.H., Hagenzieker, M., 2020. Pedestrians' road crossing behaviour in front of automated vehicles: results from a pedestrian simulation experiment using agent-based modelling. Transp. Res. Part F 69 (2020), 101–119.
Rhim, J., Lee, G., Lee, J.-H., 2020. Human moral reasoning types in autonomous vehicle moral dilemma: a cross-cultural comparison of Korea and Canada. Comput. Hum. Behav. 102 (2020), 39–56.
Riaz, F., Jabbar, S., Sajid, M., Ahmad, M., Naseer, K., Ali, N., 2018. A collision avoidance scheme for autonomous vehicles inspired by human social norms. Comput. Electr. Eng. 69, 690–704.
Rothenbücher, D., Li, J., Sirkin, D., Mok, B., Ju, W. (2016). Ghost driver: A field study investigating the interaction between pedestrians and driverless vehicles. Robot Hum. Interact. Commun. (RO-MAN), 2016 25th IEEE Int. Symp. 795–802.

Sadigh, D., Sastry, S.S., Seshia, S.A., 2019. Verifying robustness of human-aware autonomous cars. Int. Federation Automatic Control, IFAC PapersOnLine 51–34 (2019), 131–138.

SAE International,, 2018. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. Retrieved from. https://www.sae.org/standards/content/j3016_202104/.

Santoni de Sio, F., 2017. Killing by autonomous vehicles and the legal doctrine of necessity. Ethic Theory Moral Prac 20 (2), 411–429.

Santoni de Sio, F., Mecacci, G., 2021. Four responsibility gaps with Artificial Intelligence: why they matter and how to address them. Manuscript under review. Philos. Technol.

Saraoglu, M., Morozov, A., Janschek, K., 2019. MOBATSim: model-based autonomous traffic simulation framework for fault-error-failure chain analysis. IFAC PapersOnLine 52–8 (2019), 239–244.

Schwarting, W., Alonso-Mora, J., Rus, D., 2018. Planning and decision-making for autonomous vehicles. Annu. Rev. Control Robot. Autonomous Syst. 1, 187–210.

Shladover, S.E., Nowakowski, C., 2019. Regulatory challenges for road vehicle automation: lessons from the California experience. Transp. Res. Part A 122 (2019), 125–133.

Sinha, A., Chand, S., Wijayaratna, K.P., Virdi, N., Dixit, V., 2020. Comprehensive safety assessment in mixed fleets with connected and automated vehicles: a crash severity and rate evaluation of conventional vehicles. Accid. Anal. Prev. 142, 105567.

Sini, J., Violante, M., 2020. A simulation-based methodology for aiding advanced driver assistance systems hazard analysis and risk assessment. Microelectron. Reliab. 109 (2020), 113661.

Sparrow, R., Howard, M., 2017. When human beings are like drunk robots: driverless vehicles, ethics, and the future of transport. Transp. Res. Part C 80 (2017), 206–215.

van de Kaa, G., van den Ende, J., de Vries, H.J., van Heck, E., 2011. Factors for winning interface format battles: a review and synthesis of the literature. Technol. Forecast. Soc. Chang. 78 (8), 1397–1411.

van de Kaa, G., de Bruijn, H., 2015. Platforms and incentives for consensus building on complex ICT systems: the development of WiFi, Telecommunication. Policy 39 (7), 580–589.

van de Poel, I., 2018. Design for value change. Ethics Inf. Technol. 23 (1), 27–31.

van Dijke, J., van Schijndel, M., Nashashibi, F., de la Fortelle, A., 2012. Certification of automated transport systems. Procedia – Soc. Behav. Sci. 48 (2012), 3461–3470.

Van Loon, R.J., Martens, M.H., 2015. Automated driving and its effect on the safety ecosystem: how do compatibility issues affect the transition period? Procedia Manuf. 3, 3280–3285.

Virdi, N., Grzybowska, H., Waller, S.T., Dixit, V., 2019. A safety assessment of mixed fleets with connected and autonomous vehicles using the Surrogate Safety Assessment Module. Accid. Anal. Prev. 131, 95–111.

Vlakveld, W., Van der Kint, S., Hagenzieker, M. (2020). Cyclists' intentions to yield for automated cars at intersections when they have right of way: results of an experiment using high-quality video animations. Transp. Res. Part F: Traffic Psychol. Behav. 71, 288-307. doi:https://www.sciencedirect.com/science/article/pii/S1369847820304083?dgcid=coauthor.

Wang, J., Zhang, L., Huang, Y., Zhao, J., Bella, F., 2020. Safety of autonomous vehicles. J. Adv. Transp. 2020, 1–13.

Weijnen, M., Herder, P., 2018. Technology, Policy and Management: Co-evolving or Converging? In: Subrahmanian, E., Odumosu, T., Tsao, J. (Eds.), Engineering a Better Future. Springer, Cham. https://doi.org/10.1007/978-3-319-91134-2_2.

Wiegmann, P.M., de Vries, H.J., Blind, K., 2017. Multi-mode standardisation: a critical review and a research agenda. Res. Policy 46, 1370–1386.

Woldeamanuel, M., Nguyen, D., 2018. Perceived benefits and concerns of autonomous vehicles: an exploratory study of millennials' sentiments of an emerging market. Res. Transp. Econ. 71, 44–53.

Ye, L., Yamamoto, T., 2019. Evaluating the impact of connected and autonomous vehicles on traffic safety. Phys. A 526.

Zandi, B., Singer, T., Kobbert, J., Khanh, T.C., 2020. International study on the importance of communication between automated vehicles and pedestrians. Transp. Res. Part F 74 (2020), 52–66.

Zhang, B., de Winter, J., Varotto, S., Happee, R., Martens, M., 2019. Determinants of takeover time from automated driving: a meta-analysis of 129 studies. Transp. Res. Part F Traffic Psychol. Behav. 64, 285–307.

Zhuang, Y., et al. (2021) Illumination and Temperature-Aware Multispectral Networks for Edge-Computing-Enabled Pedestrian Detection. IEEE Transactions on Network Science and Engineering (2021).