Explanations in Autonomous Driving: A Survey

Daniel Omeiza, Helena Webb,
Marina Jirotka, Lars Kunze,

***Abstract*—The automotive industry is seen to have witnessed an increasing level of development in the past decades; from manufacturing manually operated vehicles to manufacturing vehicles with high level of automation. With the recent developments in Artificial Intelligence (AI), automotive companies now employ high performance AI models to enable vehicles to perceive their environment and make driving decisions with little or no influence from a human. With the hope to deploy autonomous vehicles (AV) on a commercial scale, the acceptance of AV by society becomes paramount and may largely depend on their degree of transparency, trustworthiness, and compliance to regulations. The assessment of these acceptance requirements can be facilitated through the provision of explanations for AVs' behaviour. Explainability is therefore seen as an important requirement for AVs. AVs should be able to explain what they have 'seen', done and might do in environments where they operate. In this paper, we provide a comprehensive survey of the existing work in explainable autonomous driving. First, we open by providing a motivation for explanations and examining existing standards related to AVs. Second, we identify and categorise the different stakeholders involved in the development, use, and regulation of AVs and show their perceived need for explanation. Third, we provide a taxonomy of explanations and reviewed previous work on explanation in the different AV operations. Finally, we draw a close by pointing out pertinent challenges and future research directions. This survey serves to provide fundamental knowledge required of researchers who are interested in explanation in autonomous driving.**

***Index Terms*—Explanations, explainable AI, autonomous vehicles, intelligent vehicles, human-machine interaction, regulations, standards, black-box models**

I. Introduction

THE advent of autonomous vehicles (AVs) is a significant miles-stone in the automotive industry. The increasing growth rate in the industry is considered to be precipitated by the accrued research knowledge in vehicle dynamics [1]. Moreover, the enhancements of sensing devices (e.g. LiDAR [2], [3] and Radar [4], [5]) and the emergence of deep learning algorithms are also contributing factors. Despite the technological advancements, the successful deployment of AVs in the real world may greatly depend on users' acceptance and confidence. Due to reports on AV accident cases (e.g. the Uber collision with a pedestrian [6] and Tesla's crash into road attenuator [7]), public skepticism in the acceptance of AVs in society seems to persist. Thus, effective ways to build public confidence on AVs should be devised.

Recently, the National Transportation Safety Board (NTSB) identified *driver distraction* as one of the issues that led to the Tesla crash after investigations. The Tesla driver's cell phone activity logs indicated that the Tesla driver was distracted by a game in his cell phone. The NTSB therefore suggested the need for risk mitigation pertaining to monitoring driver engagement, and the need for event data recording requirements for autonomous driving systems as this might facilitate future accident investigations.

Users (e.g. drivers and passengers) and regulators (e.g. NTSB) will benefit from explanations that explain the behaviour of autonomous systems. For example, in the Tesla crash case, a timely *intelligible and faithful* explanation (i.e. an explanation that is clear, easy to understand and correct/truthful) about the deviation action of the vehicle might have called the driver's attention to intervene. Similarly, the explanations generated from event logs, when present, are very useful in that accident investigators are able to easily identify the causes of an accident through these explanations. Moreover, system auditors can also benefit from an easier auditing process in the presence of explanations.

Further, the general stress on explainable AI (XAI) and the "right to explanation" as stipulated in the General Data Protection Regulation (GDPR) [8] underscores the essence of explanations in complex systems, especially when they are powered by black-box models. The tendency for such complex system as autonomous vehicles to make decisions that are strange and confusing to end-users are bound to occur. This is because human-understandable rules (which includes a subset of road traffic rules) are low dimensional and consider only small numbers of choices and interactions in a given time [9]. This is especially the case for vehicles with high level of automation (e.g vehicle in Level 3 or above [10]). As these vehicles make high stake decisions that can significantly affect end-users, the vehicles should explain or justify their decisions to meet set transparency guidelines.

A number of explainable AI literature focus on explaining single neurons or a single artificial neural network model while only a handful focus on explaining an entire goal-based system like autonomous vehicles which possess unique architecture and different interacting components. Providing explanations for the behaviour of such goal-based systems is therefore essential.

In this paper, we provide a structured and comprehensive overview of recent work on explanations in the autonomous driving. Previous literature surveys such as [11], [12], [13] placed more focused on works aimed at opening black-box machine learning mechanisms (data driven XAI) applied in

This work was funded in part by the UK's Engineering and Physical Sciences Research Council (EPSRC) through project RoboTIPS: Developing Responsible Robots for the Digital Economy, grant reference EP/S005099/1. Partly, it was also funded by the Assuring Autonomy International Programme, a partnership between Lloyd's Register Foundation and the University of York.

D. Omeiza, H. Webb, and M. Jirotka are with the Department of Computer Science, University of Oxford, UK (e-mail: daniel.omeiza@cs.ox.ac.uk).

L. Kunze is with the Oxford Robotics Institute, Department of Engineering Science, University of Oxford, UK (e-mail: lars@robots.ox.ac.uk

deep learning and natural language processing. Sule et al. [14] provided a systematic literature review generally on explainable agency (i.e. explaining the behavior of goal-driven agents and robots) which entailed the use of descriptive statistics to show the number of contributions made in the various aspects of explainable agencies. Zablocki et al. [15] survey was limited to vision-based autonomous driving systems. Survey papers that comprehensively cover explanations for the behaviour of AVs at different levels of operations with the requirements of different stakeholders in mind including interactions does not exist to the best of our knowledge. This paper aims to fill these gaps in the literature through a comprehensive survey. The rest of this paper is organised in nine sections: Section II presents the general need for explanations in autonomous vehicles. Section III presents and discusses the regulations and standards related to explanations in AVs. The different stakeholders who interact with AVs are identified and categorised in Section IV. The categorisation system defined in Section IV is used in the rest of the paper. Section V broadly categorises explanations and provides an explanation taxonomy. Section VI describes the core operations of an AV and reviews existing work on explanations in relation to the different core AV operations. These operations include perception, localisation, planning, and control. Section VII is dedicated to system management, which is the last core AV operation. System management involves event data recorders and human-machine interaction which are crucial to explanations. Section VIII discusses challenges in the explainable AV landscape and provides directions to future work in the field. Section IX concludes the paper with a summary.

II. Need for Explanations

The need for explanations in autonomous vehicles stems from the increasing concerns for transparency and accountability of autonomous vehicles to foster public acceptance and trust. It is believed that explanations is one way of achieving these goals. In this section, we discuss the need for explanation in the light of transparency, accountability and trust.

A. Transparency and Accountability

An autonomous system is considered to be transparent if its decision making process and behaviour are sufficiently expressive to be understood by humans [16]. Transparency is considered an important requirement for ethical systems because of its direct impact on acceptance, use, and trust [17] in the systems. Because people view transparency from different shades, the level of the impact of transparency is dependent on the persona involved. For instance, transparency will be defined to varying levels of detail among regulatory bodies, ethicists or lawyers, and the general society. In whichever way it is defined and understood, its relevance in the assessment of the quality of the behaviour of an autonomous vehicle, and their perceived safety and reliability is high. *Accountability* is seen as a direct effect of transparency. Accountability in this context refers to the ability to determine whether the decision of a system was made in compliance with procedural and substantive standards, and importantly, to hold someone responsible when there is a failure to meet the standards [18].

A transparent AV will facilitate the assignment of responsibility for its actions (especially the wrong ones) that resulted in a discriminatory, inequitable or bad outcome [19]. A way to achieve transparency in AVs is through the provision of explanations and the development and deployment of interpretable AI models to run the AV. Providing justifications in the form of human understandable explanations (or natural language) for the actions that the vehicle makes will be helpful to know who to hold responsible when there is a road traffic offence or a collision. Placing more importance and attention on explainability in AVs through the facilitation of more responsible research and innovation in the landscape would have a great positive effect on transparency and accountable.

B. Trust and Social Acceptance

The information about the functioning of a system at the user's disposal can help the user create a mental model of the system, eventually adding to the user knowledge base [20]. This imply that when adequate information about the operational modes and behaviour of a complex system are not provided, the user makes assumptions which can lead to the development of a wrong mental model about the system in context. This can in turn, affect trust, especially when the system frequently acts outside the expectations of the user. Trust can break down when there are frequent failures adequate without explanations, and regaining trust once lost can be challenging [21], [22]. For AVs, previous reports on AV accidents can have negative impacts society's perception of AVs.

According to Hussainet al. [23], an important challenge evident in the intelligent transport systems is the lack of trust from the consumer perspective [23]. AVs may provide a great deal of benefits which include the potential reduction of traffic accidents that may occur due to drivers' inattention. The public fear that this argument may fail in the appearance of unforeseen traffic conditions with an attentive human driver. This is because human drivers are seen to be able to handle such conditions better than a 'machine' [23]. In addition, failure to provide information and reasons about the decisions and operational modes of the AV when needed can leads to the development of a wrong perception about the AV, which can affect trust [20]. Abraham et al. [24] reported that the consumers' perception of trust is still not as high in spite of the great potentials of AVs. The authors claimed that the public is still hesitant about the technology, and still feel uncomfortable using it. Trust is therefore imperative for achieving commercial manufacturing and deployment of AVs. Researchers (e.g. in [25]) suggest that the provision of meaningful explanations for AVs' to passengers, pedestrians and other road participants is one way to build the necessary trust in AV technology. Explanations in autonomous driving is therefore crucial. In the following section, we will discuss explanations from the regulatory perspective.

III. Regulations and Standards

A. Explanation and AV Regulations

There are increasing concerns on the collection and use of personal data in algorithms that make critical decisions about people in domains like healthcare, finance, insurance, and criminal justice. The European Union GDPR effected in 2018 aims to provide more control rights to individuals over their personal data [26].

The GDPR also sets guidelines related to the explanation of decision made based on users' data. The GDPR guideline demands that controllers (entities handling people's personal data) provide meaningful information about the logic involved in the decisions made based on peoples' data and what the likely consequences are for individuals. It also states the use of appropriate mathematical or statistical procedures on such data. This in summary, is what is popularly referred to as the 'right to explanation'. In addition, the GDPR Articles 12 on transparency demands that the provision of information/explanation to data subjects must be done in an intelligible way (i.e. in a clear and easy to understand form [27]).

These clauses highlight the users right to question the decision of a system and demand for explanation, especially when decisions are made based on the their data. Autonomous vehicles with in-vehicle personalisation technologies are no exceptions.

The acceptance and adoption of AV will introduce new regulations and standards challenges for governments and experts [23]. The current regulations which relate to human drivers will need to be revised when a vehicle drives itself [28]. A few guidelines have been announced recently to govern autonomous driving. Some of them include: The National Association of City Transportation Officials (NATCO) policy statement on automated vehicles [29] released in June 2016, the American Research and Development (RAND) corporation document on autonomous car technology [30]. NHTSA has as well previously made attempts to legalise Level 4 and Level 5 autonomous vehicles [31]. In line explanations, the National Transportation Safety Board (NTSB) is making a call for efficient event data recording in AVs which could facilitate the provision of plausible and faithful explanations to ensure correct accident investigation [7]. See Section VII for more details on event data recording.

Check [32], [33], [34], [35] for more on regulations and ethical guidelines.

B. AV Standards

Intelligent Transport Systems (ITS) apply advanced electronics and information and communications technologies into roads and automobiles to collect, store and provide traffic information in real time all with the aim of providing convenient and safe transport, as well as improving: safety, reliability, efficiency, quality, and reduction in energy consumption [36]. Their deployments and the provision of corresponding services extend road transport to include aviation, railways and maritime. Thus, ITS has gained the attention of policy and legislative initiatives, particularly in Europe [37]. The international standard organisation technical committee 204 (ISO TC204), the IEEE and few other standard organisations have set some standards for AVs and ITS in general. The IEEE Initiatives in particular has a vision for prioritising human well-being with autonomous and intelligent systems, and the assessment of standardisation gaps for safe autonomous driving. These standards directly or indirectly unveils explainablility necessity in AVs. See Table I for a description of some relevant standards. We categorised the standards into two: Human safety related standards, and information or explanation exchange related standard. For more information, refer to the ISO report on intelligent transport systems in [38] and Apex.AI document on automated mobility [32].

IV. Stakeholders

Explanation provision in autonomous driving has many faces due to the different purposes for explanations. The level of detail (in terms of information) anticipated by the explanation recipients, the explanation type and the mode of communication vary with respect to the type of recipient and purpose for the explanation. While lay users who lack domain technical expertise may be happy and satisfied with a sparse and coarse explanation that require less effort to interpret, AI researchers and engineers would prefer a finely detailed explanation that would support a deeper and better conception of the internal functioning of a model [39]. In this light, the consideration of the persona of the explainee is necessary [40] and will help to better scope this literature. Going forward, we refer to explanation recipients or explainee as stakeholders. Having identified the typical personas in the literature, we divided stakeholders into 3 broad categories: Class A (all type of end users and society), Class B (all technical groups e.g. developers), and Class C (all forms of regulatory bodies including insurers). See further description:

1) **Class A: End-Users**
    - Passenger: this is the in-vehicle agent who may interacts with the explanation agency in the AV but not responsible for any driving operation.
    - Pedestrian: this is the agent outside the AV (external agent) who may interact with the AV to convey intentions either through gestures or an external human-machine interface (eHMI).
    - Pedestrian with Reduced Mobility (PRM): this is the agent outside the AV (external agent) who may interact with the AV to convey intentions either through gestures or an external human-machine interface (eHMI) but have reduced mobility capacity e.g. pedestrian on a wheel chair.
    - Other Road Participants: these are every other agent outside the AV (external agent) who may interact with the AV to convey intentions either through gestures or an external human-machine interface (eHMI) e.g. cyclist, other vehicles.
    - Auxiliary Driver: This is a special in-vehicle passenger who may also interact with the explanation agency in the AV and can also participate in the driving operations. This kind of participant exist in SAE level 2-5 vehicles.

TABLE I

SOME STANDARDS FOR AUTONOMOUS VEHICLES. THE STANDARDS UNDERLINES THE IMPORTANCE OF SAFE AND TRANSPARENT. EXPLANATIONS ARE PART OF THE MEASURES THAT COULD HELP ACHIEVE TRANSPARENCY AND PROVIDE ASSURANCE OF SAFETY.

Aim	Standard & Description	Stakeholder	Measures
Human Safety	**ISO 19237:2017** Pedestrian detection and collision mitigation systems	**Class B and C** AV Developers, Regulators, System Auditors, Accident Investigators, Insurer	**Causal filters** Non-contrastive (Why), Contrastive (Why-Not), Counterfactual (What-If). Should show input influence and should be able to explain the system globally
	ISO 22078:2020 Bicyclist detection and collision mitigation systems		
	ISO 26262:2011: Road vehicles – Functional safety. An international standard for functional safety of electrical and/or electronic (E/E) systems in production automobiles (2011). It addresses possible hazards caused by malfunctioning behaviour of E/E safety-related systems, including interaction of these systems.		
	ISO 21448:2019: Safety Of The Intended Functionality (SOTIF). Provides guidance on design, verification and validation measures. Guidelines on data collection (e.g. time of day, vehicle speed, weather conditions (2019). (complementary to ISO 26262).		
	UL 4600: Standard for Safety for the Evaluation of Autonomous Products. a safety case approach to ensuring autonomous product safety in general, and self-driving cars in particular.		
	SaFAD: Safety First for Automated Driving. White paper by eleven companies from the automotive industry and automated driving sector about framework for development, testing and validation of safe automated passenger vehicles (SAE Level 3/4).		
	RSS (Intel) / SFF (NVIDIA): Formal Models & Methods to evaluate safety of AV on top of ISO 26262 and ISO 21448 (proposed by companies).		
	IEEE Initiatives: “Reliable, Safe, Secure, and Time-Deterministic Intelligent Systems (2019)”; “A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems” (2019); “Assessment of standardization gaps for safe autonomous driving (2019)”.		
	The Autonomous: Global safety reference, created by the community leading automotive industry players, which facilitates the adoption of autonomous mobility on a grand scale (2019).		
Information/Data Exchange	**ISO/TR 21707:2008: Integrated transport information, management, and control—Data quality in intelligent transport systems (ITS)**. “specifies a set of standard terminology for defining the quality of data being exchanged between data suppliers and data consumers in the ITS domain” (2018).	**Class A and C** Passengers, Auxiliary Drivers, Pedestrians, Regulators, System Auditors, Accident Investigators Insurers	**Causal filters** Non-contrastive (Why), Contrastive (Why-Not), Counterfactual (What-If). May show input influence
	ISO 13111-1:2017: The use of personal ITS station to support ITS service provision for travellers. “Defines the general information and use cases of the applications based on the personal ITS station to provide and maintain ITS services to travellers including drivers, passengers and pedestrians” (2017).		
	ISO/TR 21707:2008: Integrated transport information, management, and control—Data quality in ITS systems. “Specifies a set of standard terminology for defining the quality of data being exchanged between data suppliers and data consumers in the ITS domain” (2018).		
	ISO 13111-1:2017: The use of personal ITS station to support ITS service provision for travellers. “Defines the general information and use cases of the applications based on the personal ITS station to provide and maintain ITS services to travellers including drivers, passengers and pedestrians” (2017).		
	ISO 15075:2003: In-vehicle navigation systems—Communications message set requirements. “Specifies message content and format utilized by in-vehicle navigation systems” (2003).		
	ISO/TR 20545:2017: Vehicle/roadway warning and control systems. “Provides the results of consideration on potential areas and items of standardization for automated driving systems” (2017).		
	ISO 17361:2017: Lane departure warning.		
	ISO/DIS 23150: Data communication between sensors and data fusion unit for automated driving functions.		

2) **Class B: Technicians and Engineers**
   - AV Developer: the agent who develop the automation software and tools for the AV.
   - Automobile Mechanic: the agent who repairs and maintains the AV
3) **Class C: Regulators and Insurers**
   - System Auditor: the agent who inspect the AVs' design process and operations in order to ascertain compliance to guidelines.
   - Regulator: the agent who set guidelines and regulations for the design, use and maintenance of the AVs.
   - Accident Investigator: the agent who investigates the cause of an accident in which the AV was involved.
   - Insurers: the agents who insure the AV against vandalisations, theft, and accidents.

In the next section, we categorise explanations based on methodologies and provide a taxonomy of explanation and situated the different stakeholders.

V. EXPLANATION CATEGORISATIONS

A. *Methodology-Based Categorisation*

Explanations serve different functions in different contexts [39]. Therefore, the methods of generation and evaluation are context and purpose-dependent [41]. Wang et al. [42] identified three approaches that previous pieces of academic literature have adopted in either developing or evaluating the explanations.

First, Wang et al. [42] posited that **unvalidated guidelines** for the design and evaluations of explanations exist. The

TABLE II
EXPLANATION TYPES AND THEIR INVESTIGATORY QUERIES.

Type	Class	Example Query
Contrastive	Causal	**Why Not:** why did you not do Y?
Non-Contrastive	Causal	**Why:** why did you do X?
Counterfactuals	Causal	**What If:** what would you do if Z?
Informative	Non-Causal	**What:** what are you doing?

authors claim that these kinds of guidelines are provided based on the author's experiences with non-substantial further justification. Therefore, explanation generation algorithms that generate explanations as short rules [43], or those that apply influence scores [44], partial dependence plots [45] without sufficient justification for the explanation choices made are assumed to be grounded on unvalidated guidelines. Hence, an explanation provided by these techniques may not be appropriate for a class A stakeholders due to the low intelligibility attribute of the explanation [39].

Second, researchers suggest in [46] that understanding users' requirements might be helpful in explainable AI research. It is on this premise that some explanation design approaches were classified as **empirically derived methods**. These methods elicit explanation needs from user surveys in order to determine the right explanation for a use-case [42]. For instance, explanation frameworks have been proposed for recommender systems [47], case-based reasoning [48], intelligent decision aids [49], and intelligible context-aware systems [50] upon the elicitation of users' requirements through surveys. Through user studies, Lim and Anind [50] examined explanations based on intelligibility types. The intelligibility types used were: 'why', 'why not' (contrastive), 'what if' and 'how to' (counterfactual) explanations which are considered relevant for filtering causes.

Third, some explanation design methods have been thought to be derived from **psychological constructs and formal theories** in the research literature [42]. Some of these methods (e.g. in [25]) draw on philosophy, cognitive psychology, social science, and AI theories to inform explanation design for explanation frameworks. For instance, Akula et al.'s [51] employed the Theory of Mind (ToM) in the development of an explanation framework (X-ToM). The authors claimed that in their explanation framework, the mental representations in ToM are incorporated to learn an optimal explanation policy that takes into account human's perception and beliefs. Simply put, a policy in this context is an agent's strategy [52].

Based on these design methodologies, different explanation types and frameworks have been proposed. We provide a taxonomy of these explanation types in the next section.

B. *Explanation Taxonomy*

We have identified various dimensions of explanations from the literature and present a representative subset of the reviewed work where these explanation dimensions were primarily discussed or implemented in the context of autonomous driving. The description of the various dimensions is detailed below.

a) Causes Filters: this refers to an explanation that is focused on selected *causes* relevant to interpreting an observation, with respect to existing knowledge [42]. The explanations provided in this category is assumed to be usually generated by investigatory queries like *why, why not, how to,* and *what if* [27]. These investigatory queries are assumed to produce explanation that could be contrastive ('why not' explanation), non-contrastive (sometimes 'why' explanation), or counterfactual ('how to' and 'what if' explanation). See Table II.

b) Content Type: this categorises explanations based refers to the type of information or elements referenced in the explanation. Binns et al. [41] categorised explanations based on the elements referenced in the explanations. They could be in content they contain:

- Input Influence: a list of input variables is presented along with quantitative measures of their influence (either positive or negative) on a decision.
- Sensitivity: shows what magnitude of change is required in an input variable in order to change the output class. Note that this is different from sensitivity or saliency in deep learning models.
- Case-based: picks out a relevant case from the model's training data that is most similar to the decision made, and locally generalises.
- Demographic: explanation provides an aggregate statistics of previous outcomes for people in the same demography.

c) Model Dependence: in this context refers to the possibility of having an explanation method that works for other driving condition examples order than the ones used to define (or perhaps train if a model) the explanation method. If the possibility exists, the explanation method is regarded to as *model agnostic*. Otherwise, it is regarded as *model specific* [53]. For example, an explanation method which uses a model that has been trained on a driving condition dataset with explanations as annotation will be able to provide explanations for new examples of conditions outside of its training examples (e.g. in [54]).

d) Interactivity: this refers to the possibility of a user raising follow-up questions as a way of demanding for further explanations. The conversational style of explaining allows for this [55].

e) System Type: this refers to the nature of the system that the explanation technique is primarily designed for. It could be an explanation technique for *data-driven* systems (e.g. explaining the output of a machine learning model) or *goal-driven* system (e.g. explaining the behaviour of an autonomous agent based on goals) [14]. For example, an explanation method that explains a deep learning model trained on driving scene images or video is data-driven while those that explain plans (or change in plans) in the absence of a trained machine learning model is referred to as goal-driven in this context. See Table III for an overview.

f) Scope: in this context, it refers to the extent of the part of the system being explained. The explanation is global if it explains or can explain the entire bahaviour of the part of the system it is designed for. It is referred to as local if it only explains a subset of the behaviour of the part of the system it was designed to explain. For example, a local explanation

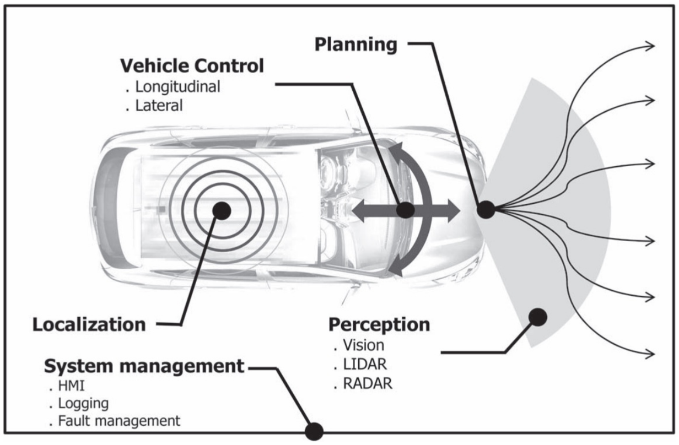

Fig. 1. Basic operations of an autonomous vehicle. Source: [74]. In Section VI and Section VII, we discuss the role of explanations for these basic operations.

method designed to explain a perception system might only be able to explain only stop and deviate actions out of many other actions possible. See Table III.

VI. Explainable Autonomous Driving Operations

This section provides a high level description of the different operations of an AV and a review of previous work on explanation for each AV operation. The operations include: perception, localisation, planning, control and navigation, system management (which includes event data recorder and human-machine interaction) [74]; see Figure 1.

A. Perception

In this section, we identify various perception datasets that have been used or that can potentially be used for explanation generation purposes. We also review recent data driven explanation methods and previous work that have applied these explanation methods in AV perception task.

1) Driving Datasets For Post-Hoc Explanations: Several driving datasets have been made available for the purpose of training machine learning models for autonomous vehicles (See [75]). Some of these datasets have annotations—e.g. handcrafted explanations [54], [76], vehicle trajectories [64], human driver behaviour [77], [64] or anomaly identification with bounding boxes [58], [76]—that are helpful for post-hoc driving behaviour explanation. We have grouped the dataset into two categories: exteroception and proprioception, based on the type of sensor and level of AV operation the data come from (see Table IV).

Although the datasets are helpful for developing explanation methods, the utilisation of these datasets for explanation purposes comes with varying challenges which include the lack of interpretability, bias due to under represented scenarios, and faithfulness. We discuss this further in Section VIII-B.

2) Vision-Based Explanation Methods useful for Perception: Various methods have been proposed to explain neural networks which are fundamental structures for perception and scene understanding in AVs. The prominent methods include *gradient-based methods*. Gradient-based or backpropagation methods are generally used for explaining convolutional neural network models. The main logic of these methods is dependent on gradients that are backpropagated from the output prediction layer of the CNN back to the input layer [16]. They are often presented in form of heatmaps (see Figure 2). These methods fall under the input influence content style in the explanation taxonomy presented in Table III.

We provide some examples of gradient-based methods that are useful for explanations in AV perception. Refer to Tjoa et al. [78] for a survey on vision based explanation methods.

- Class Activation Map (CAM) [79] and its variants like Gradient Class Activation Map (Grad-CAM) [80], Guided Grad-CAM [81], Grad-CAM++ [82], Smooth Grad-CAM++ [83]
- Other gradient based methods include: VisualBackProp [69], Layer-wise Relevance Propagation (LRP) [84], [85], DeepLift [86], [87], and Guided-Backpropagation [88].

3) Vision-Based Explanations for AVs: Bojarski et al. [69] proposed VisualBackProp for visualising super-pixels of an input image that is most influential to the predictions made by a CNN model. The authors applied VisualBackProp on an end-to-end learning system for autonomous driving (PilotNet [89]) to check whether the explanation method is able to show the parts of a driving scene image that are necessary for the steering operation of the AV model. Kim et al. [54] proposed an approach for explanation generation in autonomous driving. The approach involves training a convolutional neural network end-to-end from images to the vehicle control commands (which are acceleration and change of course). Further, textual explanations of the model actions are produced through an attention-based video-to-text model trained on the BDD-X dataset. Explanations were provided in form of saliency maps and texts (see Figure 2). A related work by Xu et al. [58] focused on scene understanding, highlighting salient objects in input that can potentially lead to a hazard. These objects are described as action inducing since their state can influence the vehicles' decisions. Apart from identifying objects, a sequence of short explanations were generated.

B. Localisation

Precise and robust localisation is critical for AVs in complex environment and scenarios [90]. For effective planning and decision making, the position and orientation information are required to be precise in all weather and traffic conditions. The goal of a precise and robust localisation is to ensure that the AV is aware that it is within its lane [91] for safety purposes. Safety is often considered as the most important design requirement and was highly considered in the derivation of requirements for AVs in [91]. Hence, communicating position per time and with justifications as explanations is crucial to expose increasing error rates right on time before they cause an accident. Although there seem to be no research related to explainable localisation, intelligible explanations remains key. They would allow for easy communication of the position of an AV, its precision, the lateral and longitudinal position error [91] of an autonomous vehicle during the localisation process

TABLE III

SUMMARY OF EXPLANATIONS CATEGORIES. THE TABLE INCLUDES A SUBSET OF THE REVIEWED PAPERS WERE EACH OR A SUBSET OF THE EXPLANATIONS CATEGORIES WAS DISCUSSED.

STAKEHOLDERS: CLASS A—PASSENGER (PA), PEDESTRIAN (PE), PEDESTRIAN WITH REDUCED MOBILITY (PRM), OTHER ROAD PARTICIPANTS (ORP), AUXILIARY DRIVER (AD), CLASS B—DEVELOPER (DV), AUTO-MECHANIC (AM), CLASS C—SYSTEM AUDITOR (SA), REGULATOR (RG), INSURER (IN), ACCIDENT INVESTIGATOR (AI).

METHODS: UNVALIDATED GUIDELINES (UG), EMPIRICALLY DERIVED METHODS (ED), PSYCHOLOGICAL CONSTRUCTS/FORMAL THEORIES (PC).

OPERATIONS: PERCEPTION (P), LOCALISATION (L), PLANNING (PL), CONTROL (C), SYSTEM MANAGEMENT (M)

Literature	**Cause Filters**			**Content Style**				**Interactivity**		**Dependence**		**System**		**Scope**				
	Non-Contrastive	*Contrastive*	*Counterfactual*	*Input Influence*	*Sensitivity*	*Case-based*	*Demographic*	*Conversational*	*Non-conversational*	*Model Agnostic*	*Model Specific*	*Goal-Driven*	*Data Driven*	*Local*	*Global*	*Method*	*Stakeholders (Class)*	*Operation*
Kim et al. [54]	✓					✓			✓	✓			✓	✓		UG + ED	B, C	P, C
Chakraborti et al. [56]	✓			✓					✓	✓		✓			✓	UG	B & C	PL
Raman et al. [57]	✓			✓					✓		✓	✓			✓	UG	B & C	PL
Xu et al. [58]	✓					✓			✓	✓			✓			UG	A & C	P
Kim & Canny [59]	✓					✓			✓		✓		✓	✓		UG	B & A	P
Cultrera et al. [60]	✓			✓					✓		✓		✓	✓		UG	B & A	P
Hayes et al. [61]	✓	✓		✓				✓		✓		✓			✓	UG	B & C	PL
Neerincx et al. [62]	✓	✓		✓				✓		✓		✓			✓	UG	B & C	P, PL
Rahimpour et al. [63]				✓					✓		✓		✓	✓		UG	B & C	P
Shen et al. [64]	✓					✓			✓		✓		✓	✓		ED	B	P
Ben-Younes et al. [65]	✓					✓			✓	✓			✓	✓		UG	A & C	P
Ha et al. [66]	✓			✓					✓			✓				ED + PC	A & B	P
Koo et al. [67]	✓			✓					✓			✓				ED + PC	A & B	P
Cruz et al. [68]	✓	✓		✓				✓		✓		✓			✓	UG	B & C	P
Bojarski et al. [69]	✓					✓			✓		✓		✓	✓		UG	B & C	P
Mori et al. [70]	✓			✓					✓		✓		✓	✓		UG	B & C	P
Liu et al. [71]	✓			✓					✓			✓		✓		ED	A	P
Rizzo et al. [72]	✓			✓					✓		✓		✓	✓		UG	B	P
Liu et al. [73]	✓			✓					✓		✓		✓	✓		UG	B & C	P

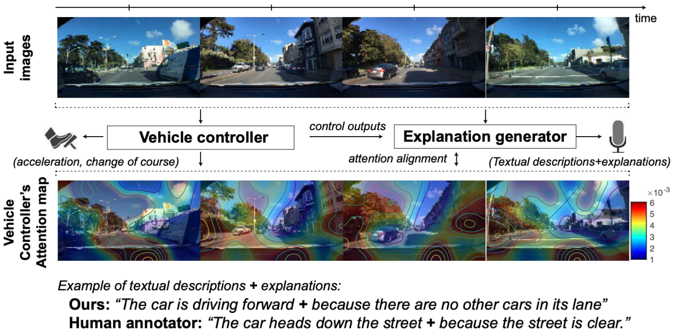
time
Input images
Vehicle controller
control outputs
Explanation generator
(acceleration, change of course)
attention alignment
(Textual descriptions+explanations)
Vehicle Controller's Attention map
×10⁻³
6
5
4
3
2
1
Example of textual descriptions + explanations:
Ours: "The car is driving forward + because there are no other cars in its lane"
Human annotator: "The car heads down the street + because the street is clear."

Fig. 2. The vehicle control model predicts commands such as acceleration and a change of course, the explanation generator model generates natural language explanations and attention maps. Source: [54].

in form of clear and intelligible explanations. Explanations from localisation will be handy for Class B stakeholders (i.e. system developers) for debugging AVs in that it can facilitate positional error correction, give other stakeholders a reliability and safety perspective of AVs. Potentially, it will inform the development process of more efficient localisation procedures.

C. Planning

Through AI planning and scheduling, the sequence of actions required for an agent to complete a task are detailed, and are further utilised in influencing the agent's online decisions or behaviours with respect to the dynamics of the environment it operates [93]. The planning system is an important aspect of autonomous vehicles because of the complex navigation procedures they make in dynamic, complex and sometimes less structured or noisy environments (e.g. urban roads, street roads with lots of pedestrians and other road participants). In fact, road fabric are very dynamic and can change with time, this makes AVs regularly update their plans (and even learn sometimes) as they operate. Often, the decision sets of an AV have higher bounds that humans might be unable to keep track of [9]. Hence, a stakeholder driving in an AV may be left in a confused state when the AV updates its trajectory plan without explanations.

Explainable planning can play a vital role in supporting users and improving their experiences when they interact with autonomous systems in complex decision-making procedures [94]. According to [95], depending on the stakeholder involved, the process may involve the translation of the agent's plans into easily understandable forms that can be easily understood by humans, and the design of the user interfaces that facilitate this understanding. Relevant work include XAI-PLAN [96], WHY-PLAN [97], refinement-based planning (RBP) [98], plan explicability and predictability [99], and plan explanation for model reconciliation [56], [100].

a) XAI-PLAN: is a domain-independent, planning system agnostic, and an explainable plan model that provides initial explanations for the decisions made by an agent planner [96]. The user explores alternative actions in a plan and a comparison is done with the user's resulting plan and the plan

TABLE IV
DRIVING DATASETS THAT CAN BE USED TO DEVELOP EXPLANATION METHODS FOR AVS AND THE STAKEHOLDERS THAT WOULD POTENTIALLY BENEFIT FROM SUCH EXPLANATIONS.

Dataset	**Size**	**Exteroception** Sensors (Camera,...)	**Proprioception** Sensors (CAN)	**Annotation & Explanation**	**Stakeholders** (see Sec. IV)
BDD-X [54]	7K × 40s	✓	✗	Textual *Why explanation* associated with videos segments with heatmaps	Class A and B
BDD-OIA [58]	23K × 5s	✓	✗	Actions and *Why explanation*	Class A, B and C
DoTA [92]	4,677 videos	✓	✗	*What explanation* (Temporal and spatial anomaly identification with bounding boxes)	Class B and C
CTA [76]	1,935 videos	✓	✗	Why explanation for accidents with cause and effects	Class B and C
HDD [77]	104 hours	✓	✓	*What explanations* for driver actions	Class B
BDD-A Extended [64]	$1,103 \times 10s$	✓	✗	Human gaze inciting *why and/or what explanation*, explanation necessity score	Class B
Lyft Level5 [64]	1,000 hours	✓	✗	Trajectory annotation, this incites other explanation types	Class B

that was suggested by the planner. The XAI-PLAN framework then provides an explanation to justify discrepancies. This kind of interaction encourages and enhances mixed-initiative planning which has the potential of improving the final plan. Interestingly, users can pose contrastive form of queries in the form “why does the plan contain action X rather than action Y?”.

b) Refinement-based Planning (RBP): A related transparent and domain independent framework called refinement-based planning (RBP) [98] produces explanations of verbal plans upon a verbal query from a user. It possesses an enhanced representation of the search space, providing a 2-way search (forward and backwards) capability when generating plans. This allows for flaw detection and plan update or optimization. Using states and action primitives, RBP paradigm integrates partial-order causal-link planning and hierarchical planning [101] (hybrid planning framework).

c) Why-Plan: Korpan and Epstein [97] also proposed Why-Plan, an explanation technique in a human-machine collaborative planning. The method juxtaposes a person and an autonomous agent objectives in a path navigation planning process and provides explanations to justify the differences in panning objectives in a meaningful and human friendly fashion. It basically addresses questions like ”why does your plan involve that action?”

The explainable planning frameworks described can serve as a blueprint for plan explanations for AVs.

D. Vehicle Control

Control in an AV generally has to do with the manipulations of vehicle motions such as lane changing, lane keeping, and car following. These manipulations are broadly categorised under longitudinal control (speed regulation with throttle and brake) and lateral control (i.e. automatic steering to follow track reference) [102].

ADAS currently works based on the sensor information generated by sensors observing the vehicle’s environment. Interfaces that come with ADAS now display digital maps [103], vehicle’s position, and track related attributes ahead or around the vehicle. Stakeholders may issue intelligibility queries [27] when the AV makes a decision against their expectations. For instance, the user may want to ask different questions based on current contexts (e.g. near-miss, special vehicle case, or collision). Intelligibility types could be in form of a why question (e.g. ”Why did you turn left?”), why not or contrastive question (e.g. ”why did you switch to the left lane instead of the right lane”), what if or counterfactual questions (e.g. ”what if you turned left instead of right?”), what question (e.g. ”What are you doing?”).

Other than existing in-vehicle visual interfaces such as mixed reality (MR) visualization [104], and other forms of flexible dashboard panels [105], in-vehicle interfaces that support the exchange of messages between the stakeholder and the AV is crucial. The user should be able to query the interface and receive explanations for navigation and control decisions in an appropriate form; either through voice, text, visual, gesture or a combination of any of the options.

In the next section, we address questions regarding explanation as it relates to AV management and interaction with respective stakeholders.

VII. System Management

In this section, we review works relating to event data recording (EDR) in AVs and human-machine interactions involving in-vehicle interfaces and external human-machine interfaces (eHMI) interfaces.

A. Logging and Fault Management: Event Data Recorder

To potentially identify faults and investigate accidents, critical actions of automobile are logged in an event data recorder

(EDR). The EDR serves as a recording device in automobiles to log information related to vehicle crashes or accidents. Upon a post-hoc analysis, a better understanding of how certain faults or accidents come about is achieved [106].

The installation of EDR in passenger vehicles has been a mandatory process in the United State since 2014. Likewise, the European Union made the *automated emergency call system (eCall)* installation compulsory for every passenger vehicle manufactured and released since April 2018 [107]. The eCall is a system installed in vehicles to provide necessary road traffic incidence information and brings quick assistance anywhere in the European Union territory [108]. As autonomous vehicles increase in society and gain more public attention, it is necessary to discriminate human driver errors and negligence from the AV's errors arising from non-adapted or poor product design or a product defect [109], [110] and express these errors in explanations. Martinesco et al. [111] attributed the existing difficulty in ascribing faults to the appropriate traffic participant to the difficulty of identifying and evaluating the correct cause of an accident. The authors provided a categorisation of accidents with respect to their causes. Accidents due to: (i) the negligence of a driver, supervisor or an operator (ii) the poor or inappropriate system design leading to an inappropriate behaviour of the AV (iii) a fault in the system (e.g. sensor damage).

In line with this, the National Highway Traffic Safety Administration (NHTSA) calls for the industry and standard bodies such as SAE and IEEE to develop a uniform approach to address data recording and sharing[1] which may in turn be useful for explanations. Pinter et al. [107] deplored the inability of the existing EDRs and eCall to provide sufficient data needed to reconstruct behaviour of the vehicle before and after the accident, and to a degree that the accident could be analysed in the perspective of liability. As AV functions continue to increase (evidently leading to full autonomy), the storage of a satisfactory number of parameters needed for the reconstruction of the vehicle's behaviour (which includes full movement) and explaining for a reasonable moment before and after the accident becomes crucial.

As an effort towards building more effective EDRs that can support explanation provision, different approaches which include the use of blockchain technologies, and more effective and robust data models have been proposed. Guo et al. [112] proposed a blockchain-inspired EDR system for autonomous vehicles to achieve indisputable accident forensics by providing trustability and verifiability assurance of an event's information. With the tested dynamic federation consensus scheme provided, the verification and confirmation of new block of event data is possible in an efficient way with no central authority involved. In terms of storage mechanisms and reliability, Yao et al. [113] proposed a Smart Black Box (SBB) to supplement traditional low-bandwidth data recording with value-driven high-bandwidth data capture. The SBB uses a deterministic Mealy machine based on data value and similarity to cache short-term histories of data as buffers. By optimising value and storage cost trade-offs, the appropriate compression quality for each frame in the driving history video data is determined. Prioritised data recording prevents the retention of low-value buffers. By discarding them, spaces are made available to store new data.

With the EU upcoming new legislative rules on EDR beginning 2022 [114] and similar in China [115], the question as to whether existing data storage facilities is sufficient for the data needs of accidents investigations involving automated vehicles. For efficient storage space management, a well defined data package which puts the data points (with necessary parameters) and the frequency of measuring and recording that enables full reconstruction of AVs' regular and irregular movements is necessary for events' explanation purposes. The data model from Pinter et al.'s [107] can be used to determine the data content required in an EDR sufficient for accident investigations and it is suitable for vehicles at different autonomy levels. Further, Bohm et al. [116] proposed a broader data base in comparison with the US EDR regulation (NHTSA 49 CFR Part 563.7) after carrying out a study involving the reconstruction of real accidents with ADAS enabled vehicles to investigate requirements. These advancements in EDRs are relevant for the development of explanation techniques for accidents (and other critical events) based on EDRs. This will setout a new landscape for explainable EDR which is currently very much under-explored. Human-machine interaction (HMI) is a key aspect of explanation in AVs. In the next section, we will discuss the relationship between HMI and explanations in the autonomous driving context.

B. Human-Machine Interaction

An AV is made up of an Automated Driving System (ADS) which consists of components for sensing, decision making, and the operation of the vehicles, requiring minimal human driving [117]. An ADS operates in in-deterministic (i.e. complex) environments where there are many decisions to choose from [118]. Thus, posing a challenge to the understandability of their operational modes. AVs are seen to have evolved over the years in terms of automation level, in-vehicle technologies and interfaces (i.e. technologies and interfaces within the vehicle). As the automation level improves, the in-vehicle technologies and interfaces also improve. In-vehicle interface trends encompass the interoperability between mobile devices (e.g. smartphones), integration and interaction with digital information and entertainment systems built into the vehicle, and the interactive interfaces for the intelligent and autonomous features of the vehicle [119]. There are five different levels of driving automation defined by the Society of Automotive Engineers (SAE) in [120]. The SAE seems to be the predominant categorisation model used by automotive engineers. Based on the academic literature and reports on vehicle automation [121], [118] and in-vehicle interfaces [122], [123], [124], we provide a description of each of the automation levels and their respective in-vehicle interfaces. The various in-vehicle interfaces support some forms of explanations (or simply information alerts in some cases) with varying levels of *intelligibility*. See Table V.

[1] See relevant documents here: https://www.nhtsa.gov/fmvss/event-data-recorders-edrs

- Level 0 (No automation): Vehicles without any form of automation are classified as Level 0 by the SAE. The in-vehicle dashboards of the vehicles in this level are made up of either mechanical/analog instrument dial (like speed gauge and revolution gauge) [125] or a combination of an analogue component and a digital component (e.g. the analogue dial and digital readout display [126]). The explanations needed in the vehicles of this level are low due to the low complexity of the vehicles [127]. The attention and involvement of the driver is required in all driving instances.
- Level 1 (Assisted automation): this category of vehicles has a primitive driver assistance system (DAS) which has stability control, anti-lock braking systems and adaptive cruise control [128]. Many vehicles at this level have many of their dashboard components digitised. The digitisation of the dashboard components support phone synchronisation, and also provide satellite navigation information. The need for explanation at this level of automation is low [129].
- Level 2 (Partial automation): vehicles at this level are seen to have partial automation, with some advanced assistance system components such as emergency braking or collision avoidance [130], [131]. Most vehicles in this category are equipped with full digital instrument dial panels that possess minimal physical buttons, large adaptive displays for legibility, and entertainment systems [132]. Examples include the Tesla auto-pilot cars and other recent Mercedes cars with *MBUX system* [133]. Although this category of vehicles possesses advanced functionalities, a driver is still required to sit behind the wheel to ensure that everything works fine. The driver takes over control in complex environments. The need for explanation is considered to be higher than that of level 1 and 0 due to the increased complexity and reduced driver intervention.
- Level 3 (Conditional Automation): according to the SAE, the vehicles in this level have conditional driving automation (i.e. the driver must remain alert to take over control in emergency situations). The vehicles have environment detection capabilities and can make informed decisions for themselves (e.g. accelerating past a slow-moving vehicle). Ekim et al. [134] report that the vehicles' autonomy is restricted in that they can only operate in operational design domains (ODDs)—highways in this case. Audi (Volkswagen) claims to be the first in the world to achieve the production of Level 3 vehicles capable of moving in constrained (i.e. geofenced) highway conditions [134]. However, there are issues with the manual handover from automatic mode to human [135], [136]. Ekim et al. [135] states that the takeover situations increase the risk of collisions. The provision of explanations to in-vehicle participants would be useful in these vehicles due to their increased complexity.
- Level 4 (High automation): the main difference between Level 4 and Level 3 vehicles is that Level 4 vehicles can intervene when something goes wrong or when a system failure occurs [137]. Level 4 vehicles do not need any human attention. Waymo, an Alphabet group already unveiled Level 4 vehicles capable of providing taxi services in Arizona without a safety driver [138]. They have tested the vehicles for over a distance of 10 million miles without an in-vehicle driver. Some Waymo vehicles do not require drivers in them, especially in some geofenced territories. The need for the provision of intelligible explanations is important in vehicles with this level of automation due to the high level of autonomy [129].
- Level 5 (Full automation): According to the SAE, human attention is completely not required and the dynamic driving task (i.e. the control switching between human and the AV) is eliminated. No vehicle currently exists with this level of automation. Future vehicles in this level will not have steering wheels, acceleration and braking pedals. Because of the high degree of sophistication, this category of vehicles would require highly intelligible causal explanations [129].

There are currently no vehicle in Level 4 and 5 automation commercially available. The development and deployment of this levels of vehicles on urban roads faces challenges relating to highly indeterminsitic environmental, weather, and human variables. Moreover, various optimisation goals such as time to reach destination, energy efficiency, comfort, and ride-sharing optimisation increase the complexity of this already difficult to solve problem. As such, carrying all of the dynamic driving tasks safely under strict conditions outside a well defined, geofenced area remains as an open problem. See Table V for a summarised description based on data available in [139], [140].

1) User Studies on In-Vehicle Interaction and Interfaces:

a) Novel Interaction Technologies: As novel interaction technologies come into existence, opportunities seem to abound for the design of useful and attractive in-vehicle user interfaces that abstract and explain vehicle automation operations (e.g. perception, planning, localisation, and control) see Figure 3. The in-vehicle user interface is critical to people's perception of driving experience[141]. There are studies that suggest that the interface design trends impact user driving experience. For example, Jung et al. [142] explore the impact of the displayed precision of instrumentation estimates of range and state-of-charge on drivers' driving experience, and attitude towards varying conditions of resource availability in an all-electric vehicle. Results from the study showed that it can be advantageous to display the uncertainty values associated with a measure rather than concealing it as participants presented with ambiguous display of range measure reported a preserved trust level towards the vehicle. Although presenting users with a single number value increases reading and apprehension time, the implication of disguised uncertainty on user experience and behaviour has to be carefully considered in critical situations. A related work by Mashko et al. [143] involved the assessment of in-vehicle navigation systems with visual display where virtual traffic signs were represented in-vehicle to assist better orient at road sections loaded with information. The use of virtual traffic signs in-vehicle improved the drivers' concentration and reaction to traffic signs on the road.

TABLE V
CAR DASHBOARD EVOLUTION AND EXPLANATION NEED

#	Explanation Interface	SAE Automation Level	XAI Demand	Vehicle Examples
1	Fully mechanical/Analog instrument dial with speed gauge and rev gauge	Level 0	Low	Old Ford vehicles and similar vehicles back before 1990
2	Analog dial with digital readout display	Level 0	Low	Older Honda Civics, Citroen C4 Picasso and others between 1990 and 2000.
3	Digital dial with Digital readout display	Level 0 and 1	Low	BMW 5 Series, Fiat 500, and Jaguar XF and others between 2010 – 2016
4	Full digital instrument dial panel with minimal buttons and adaptive display	Level 2 and 3	Moderate	Tesla autopilot, Audi A8 2016 to present
5	Specialized dashboard for high automation	Level 4	High	Waymo cars 2016 to present

Fig. 3. As autonomy continues to increase, in-vehicle interface might be designed mainly for explanation provision, entertainment or office support purposes. Source: [145].

Langois [144] proposed an interface (Lighting Peripheral Display—LPD) that creates signals that are able to be handled by peripheral vision (the ability to see objects and movement outside of the direct line of vision) while driving in order to enhance the utility of ADAS. The LPD possessed a box illuminated by light emitting diodes (LEDs), and reflected onto the windscreen. User tests conducted showed that driving performance and comfort are enhanced by LPDs. Sirkin et al. [146] developed Daze, a technique for measuring situation awareness through real-time, in-situ event alerts. The technique is ecologically valid in that it is very similar in look and feel to the applications used by people in actual driving environment, and can be applied in simulators and also in on-road research settings. The authors conducted a study which included a simulated based and on-road test deployments in order to provide assurance that Daze could characterize drivers' awareness of their immediate environment and also understand the practicality of its use.

Having looked at the existing interaction technologies, it would be worthwhile to look at what users actually prefer.

b) User Preferences: Apart of experimenting with novel interaction technologies in the context of AVs, it is would also helpful to learn about the the user experience of drivers and other in-vehicle participants to understand preferences. Mok et al. [147] described a Wizard of Oz study to get insights to how automated vehicles ought to interact with human drivers. Design improvisation sessions were conducted inside a driving simulator with interaction and interface design experts. While the two human operators (wizards) controlled the audio and driving behavior of the car, the participants were driven through a simulated track with different terrain and road conditions. The study noted that:

1) instead of taking over full control, participants wanted to share control with the vehicle;
2) participants like to know exactly when a handover (mode switch) happens and require a clear alert from the vehicle to that effect;
3) to the participants, delayed response and unperformed requests were acceptable as long as the responses provided are correct/proper;
4) AVs have a variety of means to help sustain or improve participants' trust in them.

Fu et al. [148] studied the effect of varying sensitivity and automation levels in vehicle collision avoidance systems. The authors explored this with the automatic emergency braking (AEB) systems in Level 3 autonomous vehicles, where the attention of the driver is needed to monitor the system for failures. Drivers reacted more (in terms of vigilance and awareness) to the system when it is biased to under-report hazards. The result also suggests that higher levels of autonomy in vehicles result in a lower level of driver vigilance and awareness. This was discovered when the roles of the driver and the computer were reversed, where the driver was meant to supervise an imperfect higher-level automated system; the driver performance worsened during a critical event. Related studies are described in [148], [149].

Park et al. [150] conducted a study in an attempt to understand the extent to which semi-AV decision-making should account for individual user preference. Having considered eighteen different scenarios with tactical driving goals, significant differences were discovered in scenario interpretations, AV perceptions, and vehicle decision references. The alignment of individual preference with AV decision yielded more positive changes in impression of the vehicle than unaligned decisions.

These works highlight the importance of the user-centred approach (which could have great influence on the future of human-machine interaction) in the design and development of AVs. In-vehicle user interfaces are a form of visual and voice explanations that can help users' better understand AVs' decisions. However, previous works only focus on providing information about the vehicle to users without communicating reasons and/or causal links for decisions. This remains an open challenge for AV designers and researchers. User experience differ with respect to demography. We elaborate in the next section.

c) User Demography: Previous research has shown that there is a relationship between demography factors and in-vehicle experience. Eby et al. [151]'s study on the relationship between age and length of driving in a vehicle with advance in-vehicle technologies reveals that advanced in-vehicle technologies can help extend the period over which an older adult can drive safely. Some of the in-vehicles technologies used in the study helped older drivers avoid collisions, led to improved comfort of driving, and helped the drivers travel to places and at times that they have termed odd in the past. In a related work, Yangi and Coughlin [152] explain how in-vehicle participants' age related sensory characteristics, motor and cognitive functions influence user experience with the in-vehicle technology in autonomous vehicles. Souder and Charness [153] also investigated trust and willingness of older adults to adopt AVs with comforting in-vehicle technologies and interfaces. Result suggested a correlation between age and in-vehicle comfort level. A large-scale online study that indicated that ADAS is well accepted among older drivers of 65 years and above. However, ADAS with a lower level of autonomy seemed to have slightly higher acceptance level than more autonomous systems [154]. Pettigrew et al. [155] highlighted the potentials of AVs improving the health and well-being of older people. Although, there seems to be no direct relationship between gender and in-vehicle experience, Heimstral [156] suggested that the gender of pedestrians might have a relationship with their behaviour at crosswalks. having considered existing studies on in-vehicle interaction and interfaces, we consider the interaction between AVs and external agents in the next section.

2) AV and External Agents Interaction: There are different categories of traffic participants that an AV has to interact with. This include; pedestrians, cyclists and other vehicles.

a) Pedestrians: The various factors that influence pedestrians' interactions with AVs, the communication of awareness and intent, and external interfaces for interaction between AVs and other road participants are worth discussing in detail.

- Factors Influencing Pedestrians' Interaction: In order to design an effective eHMI or explanation interface for AVs, the basic characteristics of a pedestrian is worth studying as these factors are capable of influencing the behaviour of pedestrian when interacting with autonomous vehicles. Rasouli and Tsotsos [157] identified these characteristics and broadly divided them into two groups: social factors, and environmental factors.

  One of the social factors include group crossing on crosswalks. Heimstra et al. [156] disclosed from a study that children are more likely to cross as a group rather than on individual basis. Sun et al. [158] showed that group size influences the drivers' and pedestrians' behaviour on crosswalks as drivers are more likely to yield to a group of pedestrian than an individual pedestrian. Pedestrians crossing in groups pay less attentions while at crosswalks [159] and even have reduced speed while doing so because they interact in the process [160], [161]. Social norms (or informal rules) as another example of social factors influence traffic participants behaviour and how they anticipate the intentions of each other [162]. Pedestrians tend to also imitate each other [163]. Therefore, imitation is another social factor that influences pedestrians behaviours. Other social factors include demographics (such as age [161], gender [156]), pedestrians' speeds which affects perception of objects and attention [164], cultural characteristics of individuals [165], and the ability (e.g. speed variation) of the participants [160].

  Environmental factors could be seen from three perspectives, physical context, dynamic factors, and traffic characteristics [157]. Physical context includes the traffic signs [166], signals [167], road structures [164], and weather conditions [159]. The dynamic factors include the gap acceptance between different participants involved [167], waiting time [158], and communication modes [162]. Traffic characteristics includes traffic density [163] and vehicle size [168], [169].
- Communicating Awareness and Intent: In communicating awareness and intent in AV-Pedestrian interaction, Mahadevan et al. [170] conducted a study to get insight on designing interfaces that explicitly communicate autonomous vehicle awareness and intent to pedestrians. Mahadevan et al. [170] further developed prototype interfaces and deployed them in studies involving a Segway and a car. Their results suggest that interfaces communicating vehicle awareness and intent: can assist pedestrians attempting to cross at crosswalks; can exist in the environment outside of the vehicle. They suggested a combination of modalities (e.g. visual, auditory, and physical) in the interfaces.
- Pedestrians Reactions to a Ghost Driver: Moore et al. [171] conducted a Wizard-of-Oz driverless vehicle study aimed to test pedestrians' reactions to everyday traffic in the absence of explicit external human- machine interface (eHMI). Although some pedestrians were surprised by the vehicle's autonomy, others neither noticed nor paid attention to its autonomous nature. All the pedestrians crossed in front of the vehicle without explicit signaling. This suggests that the vehicle's implicit eHMI (which is basically its motion) may suffice. Therefore, pedestrians may not need the explicit eHMI in their interaction routine. Similar study by Moore et al. [172] indicated that pedestrians crossed in front of a ghost vehicle with little hesitation even when the vehicle did not give any signal beyond the its motion. However, Li et al. [173] findings contradicts this claim by confirming pedestrians behaviours are different on encountering a vehicle with a hidden-driver based on a study carried out Europe.

3) Interaction with Pedestrians with Reduced Mobility (PRM): Pedestrians with reduced mobility might need their support devices re-engineered to allow for effective interaction with AVs [174]. Pasha et al. [174] carried out a design study to explore interface designs for interaction between AVs and pedestrians with reduced mobility (PRM). The results from the analysis disclosed that visual cues are most important interface elements, and street infrastructures are the most important location for housing cues for this category of pedestrians. They also found that wheel chairs might require interface, and the current wheel chairs would have to be altered to allow for this interface.

4) Other Road-Participants: Vehicle-cyclist interaction could be an important topic to research, especially in environment where cycling is common. Cyclists and drivers currently communicate through implicit cues (vehicle motion) and explicit but imprecise signals such as horns, lights and hand gestures [175]. Hou et al. [175] designed an virtual reality (VR) AV-cyclist immersive simulator and a number of AV-cyclist interfaces to explore interactions between AVs and cyclists. Findings suggest that AV-cyclist interfaces can improve rider confidence in lane merging scenarios. Future AVs could consistently communicate feedback which includes awareness and intent based on their sensor data [175]. More research may needed to explore how the data can be utilised to create effective and efficient interaction interfaces between AVs and other road participants. In the next section, we will present some challenges around explainability in autonomous driving and suggest future research directions.

VIII. Challenges and Future Research Directions

A. User-Centric Explanation Design and Evaluation

1) Fidelity: Most of the existing work examining explanation in the context of autonomous driving do so using the trust objective. These works involved user studies with either synthetic laboratory tasks with little or no direct connection to the real system or with microworlds [176]. A microworld is a simplified version of a real system in which the critical elements are present but with the complexities eliminated to assist easy control of the experimental [177]. Attributes as trust which have been used for explanation assessment in AV [67], [66] can be more accurately measured in a progressive experiment occurring in a real world setting. These recommendations align with the trust node of the UKRI Trustworthy Autonomous Systems programme [2] which stresses the need for an extensive human and autonomous systems interaction studies in relation to trust.

2) Stakeholder Consideration: As seen in Table III, research on explanations in AVs have mainly focused on the theory and implementation of explanations based on perception data without empirical studies. Only a few conducted user studies to elicit the requirements (e.g. when an explanation is needed and the appropriate explanations for each scenario) of the different stakeholders (e.g. some vulnerable group in Class A). Thus, the consideration of stakeholders in terms of explanation utility needs more attention. Further, most of the explanations methods proposed do not fulfil the properties of the standard causal explanations (e.g. being contrasted with relevant foils) [55]. We suggest that relevant user studies and interviews among relevant stakeholder should be carried out to inform the design and development process of explanation methods for AVs. Further, we suggest that relevant theories in the social science (e.g. causal attributions) should be considered in the explanation method designs.

B. AV Operations

1) Faithfulness and Blackboxness: A major limitation in the existing work is the low assurance of faithfulness of the explanations generated. Intermediate data or states data of AV operations are currently missing in nearly all of the existing autonomous driving datasets. Also, the existing work on explainability in AVs are merely explanations provided by humans trying to rationalise the bahaviour of an AV or an ego vehicle. These explanations obtained from rationalisation are used to annotate driving datasets which are subsequently used to train a deep explanation model. The complexity or blackbox nature of this process makes auditing difficult and thereby presents the faithfulness evaluation challenge.

Considering faithfulness on a non-binary scale as suggested in [178], explanations for autonomous driving behaviours could have improved faithfulness when considering data from the different intermediate operations such as localisation and navigation planning. Constructing the AVs' navigation plans and exploring them are possible when the start and end goals of the AV are known before hand. This information can be difficult to assess from AVs, but as a starting point, driving simulators and the Lyft Level5 dataset are helpful for a prototype at the least. To ensure the intelligibility of the explanations, the representation of these navigation plans need to be clear to lay users. Interpretable techniques such as navigation graphs and decision trees could be helpful in this regards.

Also, for intelligibility, high-level commands such as those specified by the National Highway Traffic Safety Administration (NHTSA) report [179] can be used to represent transitions between road and lane segments.

The Sense Access eXplain (SAX) project[3] in the ORI is currently collecting driving data which includes some proprioception information from the CAN bus of their Jaguar Land Rover (JLR) ego vehicle [180]. Some of the relevant CAN bus data include wheel angle, yaw, acceleration, braking among others. Using this data to supplement the exteroception data, navigation plan (if available) with interpretable representations and aglorithms will provide explanations with enhanced intelligibility and faithfulness. More internal state and intermediate AV representation data are also need.

2) Explanation Bias: Another serious limitation in current research is the inherent bias existing in the resulting explanation models built on the perception datasets (presented in Table II). The datasets are finite and are non-representative as they do not cover all possible driving scenarios in fair ratio. For instance, some traffic rules and signs differ with regions or continents. This limitation can lead to incorrect explanations or the *hit and run effect* [181] where the AV performs are serious offence or error and no record or explanation to investigate the

[2] https://www.ukri.org/news/new-trustworthy-autonomous-systems-projects-launched/

[3] https://ori.ox.ac.uk/projects/sense-assess-explain-sax/

offence. This opens up new research questions around data quality.

C. Standards and Regulations

Efforts have been made by NHTSA, SAE, IEEE and ITS on standardising the different operations of an AV to ensure safety and to better communicate information to necessary stakeholders (See Table I). However, from the perspective of explanations, we suggest that these bodies look into updating or creating standards relevant to the availability and access to data from the intermediate operations of an AV to facilitate post-hoc event reconstruction and explainability. The AV industry and research community should define a structure for the inputs and outputs expected in the various components and operation levels of an AV. This will provide a knowledge of the types of input required by explanation methods. The output from explanation methods should also be structured and properly defined to fit the respective interfaces that stakeholders will be interacting with.

IX. Conclusion

We have presented a literature survey on explanations for autonomous driving. We discussed the need for explanations in autonomous vehicles, identified and categorised different autonomous driving stakeholders who utilise explanations, identified and categorised relevant regulations and standards. Having examined the existing academic literature on explanations, we provided a taxonomy for explanations and positioned the existing work in this taxonomy. Related research literature around the basic operations of an AV (e.g. perception, localisation, planning, vehicle control, and system management) where reviewed from the explanation standpoint. Consequently, we identified key challenges around explainability in autonomous driving. Some of the challenges include the limited attention paid to user-centric AV explanation design and evaluation with respect to the different stakeholders identified in this paper. Also, the correctness and faithfulness of AV explanations is challenged by the difficulty in accessing relevant intermediate driving data. Moreover, there is a need for more explanation-specific standards and regulations for autonomous vehicles. Finally, we suggested future research directions for addressing some of the challenges identified.

Acknowledgment

The authors gratefully acknowledge the support by the UK's Engineering and Physical Sciences Research Council (EPSRC) through project RoboTIPS: Developing Responsible Robots for the Digital Economy, grant reference EP/S005099/1. They also thank the Assuring Autonomy International Programme, a partnership between Lloyd's Register Foundation and the University of York.

References

[1] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58 443–58 469, 2020.

[2] G. C. Walsh and A. Aindow, "Lidar system," Nov. 25 2014, uS Patent 8,896,818.

[3] D. S. Hall, "High definition lidar system," June 28 2011, uS Patent 7,969,558.

[4] D. K. Barton, "Modern radar system analysis," *ah*, 1988.

[5] ——, *Radar system analysis and modeling*. Artech House, 2004.

[6] N. A. Stanton, P. M. Salmon, G. H. Walker, and M. Stanton, "Models and methods for collision analysis: a comparison study based on the uber collision with a pedestrian," *Safety Science*, vol. 120, pp. 117–128, 2019.

[7] N. T. S. Board, "Collision between a sport utility vehicle operating with partial driving automation and a crash attenuator mountain view, california," "(accessed October 30, 2020)". [Online]. Available: https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR2001.pdf

[8] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[9] M. Cunneen, M. Mullins, and F. Murphy, "Autonomous vehicles and embedded artificial intelligence: The challenges of framing machine driving decisions," *Applied Artificial Intelligence*, vol. 33, no. 8, pp. 706–731, 2019.

[10] S. O.-R. A. V. S. Committee *et al.*, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," *SAE International: Warrendale, PA, USA*, 2018.

[11] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[12] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[13] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[14] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*. International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1078–1088.

[15] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of vision-based autonomous driving systems: Review and challenges," *arXiv preprint arXiv:2101.05307*, 2021.

[16] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv preprint arXiv:2006.11371*, 2020.

[17] V. Charisi, L. Dennis, M. Fisher, R. Lieck, A. Matthias, M. Slavkovik, J. Sombetzki, A. F. Winfield, and R. Yampolskiy, "Towards moral autonomous systems," *arXiv preprint arXiv:1703.04741*, 2017.

[18] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, S. Schieber, J. Waldo, D. Weinberger, and A. Wood, "Accountability of ai under the law: The role of explanation," *arXiv preprint arXiv:1711.01134*, 2017.

[19] R. Caplan, J. Donovan, L. Hanson, and J. Matthews, "Algorithmic accountability: A primer," *Data & Society*, vol. 18, 2018.

[20] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[21] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *International journal of human-computer studies*, vol. 58, no. 6, pp. 697–718, 2003.

[22] P. Madhavan and D. A. Wiegmann, "Effects of information source, pedigree, and reliability on operator interaction with decision support systems," *Human Factors*, vol. 49, no. 5, pp. 773–785, 2007.

[23] R. Hussain and S. Zeadally, "Autonomous cars: Research results, issues, and future challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1275–1313, 2018.

[24] H. Abraham, C. Lee, S. Brady, C. Fitzgerald, B. Mehler, B. Reimer, and J. F. Coughlin, "Autonomous vehicles, trust, and driving alternatives: A survey of consumer preferences," *Massachusetts Inst. Technol, AgeLab, Cambridge*, vol. 1, p. 16, 2016.

[25] R. R. Hoffman and G. Klein, "Explaining explanation, part 1: theoretical foundations," *IEEE Intelligent Systems*, vol. 32, no. 3, pp. 68–73, 2017.

[26] T. A. T. Institute, "A right to explanation," "(accessed July 24, 2020)". [Online]. Available: https://www.turing.ac.uk/research/impact-stories/a-right-to-explanation

[27] B. Y. Lim, A. K. Dey, and D. Avrahami, “Why and why not explanations improve the intelligibility of context-aware intelligent systems,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2009, pp. 2119–2128.

[28] A. M. Khan, A. Bacchus, and S. Erwin, “Policy challenges of increasing automation in driving,” *IATSS research*, vol. 35, no. 2, pp. 79–89, 2012.

[29] Nacto, “Nacto policy statement on automated vehicles.” ”(accessed October 11, 2020)”. [Online]. Available: https://nacto.org/wp-content/uploads/2016/06/NACTOPolicy-Automated-Vehicles-201606.pdf

[30] J. M. Anderson, K. Nidhi, K. D. Stanley, P. Sorensen, C. Samaras, and O. A. Oluwatola, *Autonomous vehicle technology: A guide for policymakers*. Rand Corporation, 2014.

[31] SAE, “Automated driving levels of driving automation defined in new sae international standard j3016.” ”(accessed October 11, 2020)”. [Online]. Available: https://www.sae.org/standards/content/j3016_201806/

[32] Apex.AI, “An overview of taxonomy, legislation, regulations, and standards for automated mobility,” ” 2020 (accessed February 16, 2020)”. [Online]. Available: https://www.apex.ai/post/legislation-standards-taxonomy-overview

[33] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the gdpr,” *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[34] R. Mahieu, N. J. van Eck, D. van Putten, and J. van den Hoven, “From dignity to security protocols: a scientometric analysis of digital ethics,” *Ethics and Information Technology*, vol. 20, no. 3, pp. 175–187, 2018.

[35] J. A. Kroll, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu, “Accountable algorithms,” *U. Pa. L. Rev.*, vol. 165, p. 633, 2016.

[36] ETSI, “Intelligent transport systems,” ”2018 (accessed July 24, 2020)”. [Online]. Available: https://www.etsi.org/images/files/ETSITechnologyLeaflets/IntelligentTransportSystems.pdf

[37] C. 278, “Intelligent transport systems,” ”(accessed July 24, 2020)”. [Online]. Available: https://www.itsstandards.eu/

[38] A.-P. E. Cooperation, “World report for intelligent transport systems (its) standards - a joint apec-international organization for standardization (iso) study of progress to develop and deploy its standards (iso tr 28682), september 2007,” ”2017 (accessed July 24, 2020)”. [Online]. Available: https://apec.org/Publications

[39] Y. Zhou and D. Danks, “Different” intelligibility” for different folks,” in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 194–199.

[40] P. Langley, “Varieties of explainable agency,” in *ICAPS Workshop on Explainable AI Planning (XAIP)*, 2019.

[41] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, “‘it’s reducing a human being to a percentage’ perceptions of justice in algorithmic decisions,” in *Proceedings of the 2018 Chi conference on human factors in computing systems*, 2018, pp. 1–14.

[42] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, “Designing theory-driven user-centric explainable ai,” in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–15.

[43] H. Lakkaraju, S. H. Bach, and J. Leskovec, “Interpretable decision sets: A joint framework for description and prediction,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1675–1684.

[44] A. Bussone, S. Stumpf, and D. O’Sullivan, “The role of explanations on trust and reliance in clinical decision support systems,” in *2015 International Conference on Healthcare Informatics*. IEEE, 2015, pp. 160–169.

[45] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission,” in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1721–1730.

[46] J. Zhu, A. Liapis, S. Risi, R. Bidarra, and G. M. Youngblood, “Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation,” in *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 2018, pp. 1–8.

[47] J. L. Herlocker, J. A. Konstan, and J. Riedl, “Explaining collaborative filtering recommendations,” in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, 2000, pp. 241–250.

[48] T. R. Roth-Berghofer, “Explanations and case-based reasoning: Foundational issues,” in *European Conference on Case-Based Reasoning*. Springer, 2004, pp. 389–403.

[49] M. S. Silveira, C. S. de Souza, and S. D. Barbosa, “Semiotic engineering contributions for designing online help systems,” in *Proceedings of the 19th annual international conference on Computer documentation*, 2001, pp. 31–38.

[50] B. Y. Lim and A. K. Dey, “Assessing demand for intelligibility in context-aware applications,” in *Proceedings of the 11th international conference on Ubiquitous computing*, 2009, pp. 195–204.

[51] A. R. Akula, C. Liu, S. Saba-Sadiya, H. Lu, S. Todorovic, J. Y. Chai, and S.-C. Zhu, “X-tom: Explaining with theory-of-mind for gaining justified human trust,” *arXiv preprint arXiv:1909.06907*, 2019.

[52] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[53] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, “Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai,” *Information Fusion*, vol. 58, pp. 82–115, 2020.

[54] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, “Textual explanations for self-driving vehicles,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 563–578.

[55] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[56] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, “Plan explanations as model reconciliation: Moving beyond explanation as soliloquy,” *arXiv preprint arXiv:1701.08317*, 2017.

[57] V. Raman, C. Lignos, C. Finucane, K. C. Lee, M. P. Marcus, and H. Kress-Gazit, “Sorry dave, i’m afraid i can’t do that: Explaining unachievable robot tasks using natural language.” in *Robotics: Science and Systems*, vol. 2, no. 1. Citeseer, 2013, pp. 2–1.

[58] Y. Xu, X. Yang, L. Gong, H.-C. Lin, T.-Y. Wu, Y. Li, and N. Vasconcelos, “Explainable object-induced action decision for autonomous vehicles,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9523–9532.

[59] J. Kim and J. Canny, “Interpretable learning for self-driving cars by visualizing causal attention,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2942–2950.

[60] L. Cultrera, L. Seidenari, F. Becattini, P. Pala, and A. Del Bimbo, “Explaining autonomous driving by learning end-to-end visual attention,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 340–341.

[61] B. Hayes and J. A. Shah, “Improving robot controller transparency through autonomous policy explanation,” in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 2017, pp. 303–312.

[62] M. A. Neerincx, J. van der Waa, F. Kaptein, and J. van Diggelen, “Using perceptual and cognitive explanations for enhanced human-agent team performance,” in *International Conference on Engineering Psychology and Cognitive Ergonomics*. Springer, 2018, pp. 204–214.

[63] A. Rahimpour, S. Martin, A. Tawari, and H. Qi, “Context aware road-user importance estimation (icare),” in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 2337–2343.

[64] Y. Shen, S. Jiang, Y. Chen, E. Yang, X. Jin, Y. Fan, and K. D. Campbell, “To explain or not to explain: A study on the necessity of explanations for autonomous vehicles,” *arXiv preprint arXiv:2006.11684*, 2020.

[65] H. Ben-Younes, É. Zablocki, P. Pérez, and M. Cord, “Driving behavior explanation with multi-level fusion,” *arXiv preprint arXiv:2012.04983*, 2020.

[66] T. Ha, S. Kim, D. Seo, and S. Lee, “Effects of explanation types and perceived risk on trust in autonomous vehicles,” *Transportation research part F: traffic psychology and behaviour*, vol. 73, pp. 271–280, 2020.

[67] J. Koo, J. Kwac, W. Ju, M. Steinert, L. Leifer, and C. Nass, “Why did my car just do that? explaining semi-autonomous driving actions to improve driver understanding, trust, and performance,” *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 9, no. 4, pp. 269–275, 2015.

[68] F. Cruz, R. Dazeley, and P. Vamplew, “Memory-based explainable reinforcement learning,” in *Australasian Joint Conference on Artificial Intelligence*. Springer, 2019, pp. 66–77.

[69] M. Bojarski, A. Choromanska, K. Choromanski, B. Firner, L. J. Ackel, U. Muller, P. Yeres, and K. Zieba, “Visualbackprop: Efficient visualization of cnns for autonomous driving,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 4701–4708.

[70] K. Mori, H. Fukui, T. Murase, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, “Visual explanation by attention branch network for end-to-end learning-based self-driving,” in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1577–1582.

[71] T. Liu, H. Zhou, M. Itoh, and S. Kitazaki, "The impact of explanation on possibility of hazard detection failure on driver intervention under partial driving automation," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 150–155.

[72] S. G. Rizzo, G. Vantini, and S. Chawla, "Reinforcement learning with explainability for traffic signal control," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 3567–3572.

[73] Y.-C. Liu, Y.-A. Hsieh, M.-H. Chen, C.-H. H. Yang, J. Tegner, and Y.-C. J. Tsai, "Interpretable self-attention temporal reasoning for driving behavior understanding," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2338–2342.

[74] K. Jo, J. Kim, D. Kim, C. Jang, and M. Sunwoo, "Development of autonomous car—part i: Distributed system architecture and development process," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 12, pp. 7131–7140, 2014.

[75] J. Janai, F. Güney, A. Behl, A. Geiger, *et al.*, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," *Foundations and Trends® in Computer Graphics and Vision*, vol. 12, no. 1–3, pp. 1–308, 2020.

[76] T. You and B. Han, "Traffic accident benchmark for causality recognition," in *European Conference on Computer Vision*. Springer, 2020, pp. 540–556.

[77] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7699–7707.

[78] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): towards medical xai," *arXiv preprint arXiv:1907.07374*, 2019.

[79] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[80] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[81] Z. Tang, K. V. Chuang, C. DeCarli, L.-W. Jin, L. Beckett, M. J. Keiser, and B. N. Dugger, "Interpretable classification of alzheimer's disease pathologies with a convolutional neural network pipeline," *Nature communications*, vol. 10, no. 1, pp. 1–14, 2019.

[82] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 839–847.

[83] D. Omeiza, S. Speakman, C. Cintas, and K. Weldermariam, "Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models," *arXiv preprint arXiv:1908.01224*, 2019.

[84] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature communications*, vol. 10, no. 1, pp. 1–8, 2019.

[85] W. Samek, G. Montavon, A. Binder, S. Lapuschkin, and K.-R. Müller, "Interpreting the predictions of complex ml models by layer-wise relevance propagation," *arXiv preprint arXiv:1611.08191*, 2016.

[86] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *arXiv preprint arXiv:1704.02685*, 2017.

[87] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[88] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[89] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[90] L. Wang, Y. Zhang, and J. Wang, "Map-based localization method for autonomous vehicles using 3d-lidar," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 276–281, 2017.

[91] T. G. Reid, S. E. Houts, R. Cammarata, G. Mills, S. Agarwal, A. Vora, and G. Pandey, "Localization requirements for autonomous vehicles," *arXiv preprint arXiv:1906.01061*, 2019.

[92] Y. Yao, X. Wang, M. Xu, Z. Pu, E. Atkins, and D. Crandall, "When, where, and what? a new dataset for anomaly detection in driving videos," *arXiv preprint arXiv:2004.03044*, 2020.

[93] F. Ingrand and M. Ghallab, "Deliberation for autonomous robots: A survey," *Artificial Intelligence*, vol. 247, pp. 10–44, 2017.

[94] T. Chakraborti, S. Sreedharan, and S. Kambhampati, "The emerging landscape of explainable automated planning & decision making," 2020.

[95] F. Sado, C. K. Loo, M. Kerzel, and S. Wermter, "Explainable goal-driven agents and robots–a comprehensive review and new framework," *arXiv preprint arXiv:2004.09705*, 2020.

[96] R. Borgo, M. Cashmore, and D. Magazzeni, "Towards providing explanations for ai planner decisions," *arXiv preprint arXiv:1810.06338*, 2018.

[97] R. Korpan and S. L. Epstein, "Toward natural explanations for a robot's navigation plans," *Notes from the Explainable Robotic Systems Worshop, Human-Robot Interaction*, 2018.

[98] J. Bidot, S. Biundo, T. Heinroth, W. Minker, F. Nothdurft, and B. Schattenberg, "Verbal plan explanations for hybrid planning." in *MKWI*. Citeseer, 2010, pp. 2309–2320.

[99] Y. Zhang, S. Sreedharan, A. Kulkarni, T. Chakraborti, H. H. Zhuo, and S. Kambhampati, "Plan explicability and predictability for robot task planning," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 1313–1320.

[100] T. Chakraborti, S. Sreedharan, S. Grover, and S. Kambhampati, "Plan explanations as model reconciliation–an empirical study," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 258–266.

[101] S. Biundo and B. Schattenberg, "From abstract crisis to concrete relief—a preliminary report on combining state abstraction and htn planning," in *Sixth European Conference on Planning*, 2014.

[102] A. Khodayari, A. Ghaffari, S. Ameli, and J. Flahatgar, "A historical review on lateral and longitudinal control of autonomous vehicle motions," in *2010 International Conference on Mechanical and Electrical Technology*. IEEE, 2010, pp. 421–429.

[103] V. Blervaque, K. Mezger, L. Beuk, and J. Loewenau, "Adas horizon—how digital maps can contribute to road safety," in *Advanced Microsystems for Automotive Applications 2006*. Springer, 2006, pp. 427–436.

[104] S. Sasai, I. Kitahara, Y. Kameda, Y. Ohta, M. Kanbara, Y. Morales, N. Ukita, N. Hagita, T. Ikeda, and K. Shinozawa, "Mr visualization of wheel trajectories of driving vehicle by seeing-through dashboard," in *2015 IEEE International Symposium on Mixed and Augmented Reality Workshops*. IEEE, 2015, pp. 40–46.

[105] L. Marques, V. Vasconcelos, P. Pedreiras, and L. Almeida, "A flexible dashboard panel for a small electric vehicle," in *6th Iberian Conference on Information Systems and Technologies (CISTI 2011)*. IEEE, 2011, pp. 1–4.

[106] B.-F. Wu, Y.-H. Chen, and C.-H. Yeh, "Driving behaviour-based event data recorder," *IET Intelligent Transport Systems*, vol. 8, no. 4, pp. 361–367, 2013.

[107] K. Pinter, Z. Szalay, and G. Vida, "Road accident reconstruction using on-board data, especially focusing on the applicability in case of autonomous vehicles," *Periodica Polytechnica Transportation Engineering*, 2020.

[108] N. Virtanen, A. Schirokoff, and J. Luom, "Impacts of an automatic emergency call system on accident consequences," in *Proceedings of the 18th ICTCT, Workshop Transport telemetric and safety. Finland*, 2005, pp. 1–6.

[109] U. Bose, "The black box solution to autonomous liability," *Wash. UL Rev.*, vol. 92, p. 1325, 2014.

[110] W. J. Kohler and A. Colbert-Taylor, "Current law and potential legal issues pertaining to automated, autonomous and connected vehicles," *Santa Clara Computer & High Tech. LJ*, vol. 31, p. 99, 2014.

[111] A. Martinesco, M. Netto, A. M. Neto, and V. H. Etgens, "A note on accidents involving autonomous vehicles: Interdependence of event data recorder, human-vehicle cooperation and legal aspects," *IFAC-PapersOnLine*, vol. 51, no. 34, pp. 407–410, 2019.

[112] H. Guo, E. Meamari, and C.-C. Shen, "Blockchain-inspired event recording system for autonomous vehicles," in *2018 1st IEEE international conference on hot information-centric networking (HotICN)*. IEEE, 2018, pp. 218–222.

[113] Y. Yao and E. Atkins, "The smart black box: A value-driven high-bandwidth automotive event data recorder," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[114] E. Commission, "Teuropean commission - press release: Road safety: Commission welcomes agreement on new eu rules to help save

lives." "2019 (accessed February 8, 2021)". [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/IP_19_1793

[115] UNECE, "Working party on automated/autonomous and connected vehicles (grva): Edr/dssad 1st session. edr-dssad-01-06 overview of edr," "2019 (accessed February 8, 2021)". [Online]. Available: https://wiki.unece.org/pages/viewpage.action?pageId=87621710

[116] K. Böhm, T. Kubjatko, D. Paula, and H.-G. Schweiger, "New developments on edr (event data recorder) for automated vehicles," *Open Engineering*, vol. 10, no. 1, pp. 140–146, 2020.

[117] B. W. Smith and J. Svensson, "Automated and autonomous driving: regulation under uncertainty," 2015.

[118] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58 443–58 469, 2020.

[119] S. Damiani, E. Deregibus, and L. Andreone, "Driver-vehicle interfaces and interaction: where are they going?" *European transport research review*, vol. 1, no. 2, pp. 87–96, 2009.

[120] S. international, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," *SAE International,(J3016)*, 2016.

[121] Y. F. Payalan and M. A. Guvensan, "Towards next-generation vehicles featuring the vehicle intelligence," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 1, pp. 30–47, 2019.

[122] A. Ulahannan, R. Cain, S. Thompson, L. Skrypchuk, A. Mouzakitis, P. Jennings, and S. Birrell, "User expectations of partial driving automation capabilities and their effect on information design preferences in the vehicle," *Applied ergonomics*, vol. 82, p. 102969, 2020.

[123] R. Li, Q.-X. Qu, and Z. Lu, "Interactive design of digital car dashboard interfaces," in *International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*. Springer, 2017, pp. 343–353.

[124] CarBuyTom, "The evolution of the car dashboard," "2014 (accessed October 11, 2020)". [Online]. Available: https://www.carbuyertom.com/blog/evolution-car-dashboard/

[125] H. Yokota, S. Saotome, and T. Matsumura, "Display unit for vehicle," July 13 2010, uS Patent 7,755,601.

[126] J. A. Castellano, *Handbook of display technology*. Elsevier, 2012.

[127] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, and A. Holzinger, "Explainable ai: the new 42?" in *International cross-domain conference for machine learning and knowledge extraction*. Springer, 2018, pp. 295–303.

[128] R. Rajamani, *Vehicle dynamics and control*. Springer Science & Business Media, 2011.

[129] M. Kyriakidis, J. C. de Winter, N. Stanton, T. Bellet, B. van Arem, K. Brookhuis, M. H. Martens, K. Bengler, J. Andersson, N. Merat, *et al.*, "A human factors perspective on automated driving," *Theoretical Issues in Ergonomics Science*, vol. 20, no. 3, pp. 223–249, 2019.

[130] M. R. Hafner, D. Cunningham, L. Caminiti, and D. Del Vecchio, "Cooperative collision avoidance at intersections: Algorithms and experiments," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1162–1175, 2013.

[131] A. Colombo and D. Del Vecchio, "Efficient algorithms for collision avoidance at intersections," in *Proceedings of the 15th ACM international conference on Hybrid Systems: Computation and Control*, 2012, pp. 145–154.

[132] J. L. Campbell, J. L. Brown, J. S. Graving, C. M. Richard, M. G. Lichty, L. P. Bacon, J. F. Morgan, H. Li, D. N. Williams, and T. Sanquist, "Human factors design guidance for level 2 and level 3 automated driving concepts," Tech. Rep., 2018.

[133] R. J. C. Eleotério, B. S. Paterlini, E. N. Baldissera, and A. S. Prullage, "A brief discussion on driver feedback systems and their results," SAE Technical Paper, Tech. Rep., 2018.

[134] P. E. Ross, "The audi a8: the world's first production car to achieve level 3 autonomy," *IEEE Spectrum*, vol. 1, 2017.

[135] C. Gold, M. Körber, D. Lechner, and K. Bengler, "Taking over control from highly automated vehicles in complex traffic situations: the role of traffic density," *Human factors*, vol. 58, no. 4, pp. 642–652, 2016.

[136] N. Merat, A. H. Jamson, F. C. Lai, M. Daly, and O. M. Carsten, "Transition to manual: Driver behaviour when resuming control from a highly automated vehicle," *Transportation research part F: traffic psychology and behaviour*, vol. 27, pp. 274–282, 2014.

[137] J. Davis, "Dreaming of driverless: What's the difference between level 2 and level 5 autonomy?" "2018 (accessed October 10, 2020)". [Online]. Available: https://blogs.nvidia.com/blog/2018/01/25/whats-difference-level-2-level-5-autonomy/

[138] A. Sage, "Waymo unveils self-driving taxi service in arizona for paying customers," *Rueters, December*, 2018.

[139] D. Lavrinc, "Cars with a digital dashboard," "2018 (accessed July 24, 2020)". [Online]. Available: hhttps://www.buyacar.co.uk/cars/902/cars-with-a-digital-dashboard

[140] C. Edwards, "Car safety with a digital dashboard," *Engineering & Technology*, vol. 10, no. 9, pp. 60–64, 2014.

[141] A. Schmidt, A. K. Dey, A. L. Kun, and W. Spiessl, "Automotive user interfaces: human computer interaction in the car," in *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, 2010, pp. 3177–3180.

[142] M. F. Jung, D. Sirkin, T. M. Gür, and M. Steinert, "Displayed uncertainty improves driving experience and behavior: The case of range anxiety in an electric car," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 2201–2210.

[143] A. Mashko, P. Bouchner, D. Rozhdestvenskiy, and S. Novotnỳ, "Virtual traffic signs-assessment of an alternative adas user interface with use of driving simulator." *Advances in Transportation Studies*, no. 1, 2016.

[144] S. Langlois, "Adas hmi using peripheral vision," in *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2013, pp. 74–81.

[145] M. Hussey, "Driverless cars designed for use as mobile offices," "2014 (accessed February 17, 2021)". [Online]. Available: https://www.dezeen.com/2014/02/21/driverless-car-concept-vehicle-xchange-by-rinspeed/

[146] D. Sirkin, N. Martelaro, M. Johns, and W. Ju, "Toward measurement of situation awareness in autonomous vehicles," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 405–415.

[147] B. K.-J. Mok, D. Sirkin, S. Sibi, D. B. Miller, and W. Ju, "Understanding driver-automated vehicle interactions through wizard of oz design improvisation," 2015.

[148] E. Fu, M. Johns, D. A. Hyde, S. Sibi, M. Fischer, and D. Sirkin, "Is too much system caution counterproductive? effects of varying sensitivity and automation levels in vehicle collision avoidance systems," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.

[149] B. Mok, M. Johns, K. J. Lee, D. Miller, D. Sirkin, P. Ive, and W. Ju, "Emergency, automation off: Unstructured transition timing for distracted drivers of automated vehicles," in *2015 IEEE 18th international conference on intelligent transportation systems*. IEEE, 2015, pp. 2458–2464.

[150] S. Y. Park, D. J. Moore, and D. Sirkin, "What a driver wants: User preferences in semi-autonomous vehicle decision-making," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.

[151] D. W. Eby, L. J. Molnar, L. Zhang, R. M. S. Louis, N. Zanier, L. P. Kostyniuk, and S. Stanciu, "Use, perceptions, and benefits of automotive technologies among aging drivers," *Injury epidemiology*, vol. 3, no. 1, p. 28, 2016.

[152] J. Yang and J. F. Coughlin, "In-vehicle technology for self-driving cars: Advantages and challenges for aging drivers," *International Journal of Automotive Technology*, vol. 15, no. 2, pp. 333–340, 2014.

[153] D. Souders and N. Charness, "Challenges of older drivers' adoption of advanced driver assistance systems and autonomous vehicles," in *International Conference on Human Aspects of IT for the Aged Population*. Springer, 2016, pp. 428–440.

[154] H. Braun, M. Gärtner, S. Trösterer, L. E. Akkermans, M. Seinen, A. Meschtscherjakov, and M. Tscheligi, "Advanced driver assistance systems for aging drivers: Insights on 65+ drivers' acceptance of and intention to use adas," in *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2019, pp. 123–133.

[155] S. Pettigrew, S. L. Cronin, and R. Norman, "Brief report: the unrealized potential of autonomous vehicles for an aging population," *Journal of aging & social policy*, vol. 31, no. 5, pp. 486–496, 2019.

[156] N. W. Heimstra, J. Nichols, and G. Martin, "An experimental methodology for analysis of child pedestrian behavior," *Pediatrics*, vol. 44, no. 5, pp. 832–838, 1969.

[157] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 900–918, 2019.

[158] D. Sun, S. Ukkusuri, R. F. Benekohal, and S. T. Waller, "Modeling of motorist-pedestrian interaction at uncontrolled mid-block crosswalks," in *Transportation Research Record, TRB Annual Meeting CD-ROM, Washington, DC*, 2003.

[159] W. A. Harrell, “Factors influencing pedestrian cautiousness in crossing streets,” *The Journal of Social Psychology*, vol. 131, no. 3, pp. 367–372, 1991.

[160] R. Sun, X. Zhuang, C. Wu, G. Zhao, and K. Zhang, “The estimation of vehicle speed and stopping distance by pedestrians crossing streets in a naturalistic traffic environment,” *Transportation research part F: traffic psychology and behaviour*, vol. 30, pp. 97–106, 2015.

[161] M. M. Ishaque and R. B. Noland, “Behavioural issues in pedestrian speed choice and street crossing behaviour: a review,” *Transport Reviews*, vol. 28, no. 1, pp. 61–85, 2008.

[162] G. S. Wilde, “Immediate and delayed social interaction in road user behaviour,” *Applied Psychology*, vol. 29, no. 4, pp. 439–460, 1980.

[163] M. Šucha, “Road users’ strategies and communication: driver-pedestrian interaction,” *Transport Research Arena (TRA)*, 2014.

[164] R. R. Oudejans, C. F. Michaels, B. Van Dort, and E. J. Frissen, “To cross or not to cross: The effect of locomotion on street-crossing behavior,” *Ecological psychology*, vol. 8, no. 3, pp. 259–267, 1996.

[165] G. M. Björklund and L. Åberg, “Driver behaviour in intersections: Formal and informal traffic rules,” *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 8, no. 3, pp. 239–253, 2005.

[166] R. L. Moore, “Pedestrian choice and judgment,” *Journal of the Operational Research Society*, vol. 4, no. 1, pp. 3–10, 1953.

[167] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, “Agreeing to cross: How drivers and pedestrians communicate,” in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 264–269.

[168] S. Das, C. F. Manski, and M. D. Manuszak, “Walk or wait? an empirical analysis of street crossing decisions,” *Journal of applied econometrics*, vol. 20, no. 4, pp. 529–548, 2005.

[169] J. Caird and P. Hancock, “The perception of arrival time for different oncoming vehicles at an intersection,” *Ecological Psychology*, vol. 6, no. 2, pp. 83–109, 1994.

[170] K. Mahadevan, S. Somanath, and E. Sharlin, “Communicating awareness and intent in autonomous vehicle-pedestrian interaction,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–12.

[171] D. Moore, R. Currano, G. E. Strack, and D. Sirkin, “The case for implicit external human-machine interfaces for autonomous vehicles,” in *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2019, pp. 295–307.

[172] D. Moore, G. E. Strack, R. Currano, and D. Sirkin, “Visualizing implicit ehmi for autonomous vehicles,” in *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings*, 2019, pp. 475–477.

[173] J. Li, R. Currano, D. Sirkin, D. Goedicke, H. Tennent, A. Levine, V. Evers, and W. Ju, “On-road and online studies to investigate beliefs and behaviors of netherlands, us and mexico pedestrians encountering hidden-driver vehicles,” in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 141–149.

[174] A. Z. Asha, C. Smith, L. Oehlberg, S. Somanath, and E. Sharlin, “Views from the wheelchair: Understanding interaction between autonomous vehicle and pedestrians with reduced mobility,” in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–8.

[175] M. Hou, K. Mahadevan, S. Somanath, E. Sharlin, and L. Oehlberg, “Autonomous vehicle-cyclist interaction: Peril and promise,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.

[176] J. D. Lee and K. A. See, “Trust in automation: Designing for appropriate reliance,” *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[177] B. Brehmer and D. Dörner, “Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study,” *Computers in human behavior*, vol. 9, no. 2-3, pp. 171–184, 1993.

[178] A. Jacovi and Y. Goldberg, “Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?” *arXiv preprint arXiv:2004.03685*, 2020.

[179] W. G. Najm, J. D. Smith, M. Yanagisawa, *et al.*, “Pre-crash scenario typology for crash avoidance research,” United States. National Highway Traffic Safety Administration, Tech. Rep., 2007.

[180] O. R. Institute, “Sense-assess-explain (sax):building trust in autonomous vehicles in challenging real-world driving scenarios,” ”2020 (accessed February 9, 2021)”. [Online]. Available: https://www.york.ac.uk/assuring-autonomy/projects/sax/

[181] M. McQuillen and D. A. Makled, “Hit-and-run detection,” July 10 2018, uS Patent 10,019,857.

Daniel Omeiza received a bachelors degree in Computer Science from the University of Ilorin in 2015. He obtained a masters degree in Information Technology from Carnegie Mellon University in 2019. He is currently working towards a DPhil degree at the University of Oxford. His research interests primarily lie on explainability in autonomous driving. He is a student member of the IEEE.

Dr. Helena Webb is a Senior Researcher in the Department of Computer Science at Oxford. She is interested in the ways that users interact with technologies in different kinds of setting and how social action both shapes and is shaped by innovation. She works on projects that seek to identify mechanisms for the improved design, responsible development and effective regulation of technology. Whilst at Oxford she has worked on projects relating to, amongst others, harmful content on social media, algorithm bias, resources in STEM education, and responsible robotics. She is an interdisciplinary researcher and specialises in the application of qualitative research methods. She is also very interested in the methodological innovation to combine detailed, granular analysis with larger scale computational work.

Marina Jirotka is Professor of Human Centred Computing at the University of Oxford and Governing Body Fellow at St Cross College. She leads an interdisciplinary research group developing methods for building computing systems responsibly to support human, societal and environmental values. The team focusses on responsible innovation, in a range of ICT fields including robotics, AI, machine learning, quantum computing, social media and the digital economy. She obtained her BSc in Psychology and Social Anthropology (Cons) from The University of London Goldsmiths College in 1985 and her Masters in Computing and Artificial Intelligence from the University of South Bank in 1987. Her doctorate in Computer Science, ”An Investigation into Contextual Approaches to Requirements Capture”, was undertaken at the University of Oxford in 2000. She leads an interdisciplinary research group developing methods for building computing systems responsibly to support human, societal and environmental values.

Lars Kunze is a Departmental Lecturer in Robotics in the Oxford Robotics Institute (ORI) and the Department of Engineering Science at Oxford University. At ORI, Dr Kunze leads the Cognitive Robotics Group. He is also a Programme Fellow of the Assuring Autonomy International Programme (AAIP) and a Co-Editor of the German Journal of Artificial Intelligence (KI Journal, Springer). Before joining Oxford, he was a Research Fellow in the Intelligent Robotics Lab at Birmingham University. He studied Cognitive Science (BSc, 2006) and Computer Science (MSc, 2008) at the University of Osnabrück, Germany, and partly at the University of Edinburgh, UK. He received a PhD (Dr. rer. nat.) from the Technical University of Munich in 2014.