# Advancing Explainable Autonomous Vehicle Systems: A Comprehensive Review and Research Roadmap

SULE TEKKESINOGLU, University of Oxford, Oxford, UK
AZRA HABIBOVIC, Scania CV AB, Sodertalje, Sweden
LARS KUNZE, University of the West of England, Bristol, UK and University of Oxford, Oxford, UK

Given the uncertainty surrounding how existing explainability methods for autonomous vehicles (AVs) meet the diverse needs of stakeholders, a thorough investigation is imperative to determine the contexts requiring explanations and suitable interaction strategies. A comprehensive review becomes crucial to assess the alignment of current approaches with varied interests and expectations within the AV ecosystem. This study presents a review to discuss the complexities associated with explanation generation and presentation to facilitate the development of more effective and inclusive explainable AV systems. Our investigation led to categorising existing literature into three primary topics: explanatory tasks, explanatory information and explanatory information communication. Drawing upon our insights, we have proposed a comprehensive roadmap for future research centred on (i) knowing the interlocutor, (ii) generating timely explanations, (ii) communicating human-friendly explanations and (iv) continuous learning. Our roadmap is underpinned by principles of responsible research and innovation, emphasising the significance of diverse explanation requirements. To effectively tackle the challenges associated with implementing explainable AV systems, we have delineated various research directions, including the development of privacy-preserving data integration, ethical frameworks, real-time analytics, human-centric interaction design and enhanced cross-disciplinary collaborations. By exploring these research directions, the study aims to guide the development and deployment of explainable AVs, informed by a holistic understanding of user needs, technological advancements, regulatory compliance and ethical considerations, thereby ensuring safer and more trustworthy autonomous driving experiences.

CCS Concepts: • **Computer systems organization → Robotics**; • **Human-centered computing → Interaction devices**; • **Computing methodologies → Artificial intelligence**;

Additional Key Words and Phrases: Explainable autonomous vehicles, Human-robot interaction, XAI, Transparency, Feedback communication, Human factors

**ACM Reference format:**
Sule Tekkesinoglu, Azra Habibovic, and Lars Kunze. 2025. Advancing Explainable Autonomous Vehicle Systems: A Comprehensive Review and Research Roadmap. *ACM Trans. Hum.-Robot Interact.* 14, 3, Article 39 (March 2025), 46 pages.
https://doi.org/10.1145/3714478

This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation. L. Kunze was supported by EPSRC Project RAILS (EP/W011344/1).

Authors' Contact Information: Sule Tekkesinoglu (corresponding author), University of Oxford, Oxford, UK; e-mail: sule@robots.ox.ac.uk; Azra Habibovic, Scania CV AB, Sodertalje, Sweden; e-mail: azra.habibovic@scania.com; Lars Kunze, University of the West of England, Bristol, UK and University of Oxford, Oxford, UK; e-mail: lars@robots.ox.ac.uk.

ACM 2573-9522/2025/3-ART39
https://doi.org/10.1145/3714478

## 1 Introduction

The promise of **autonomous vehicles (AVs)** is significant, but their success is contingent on overcoming key challenges, one of which is ensuring that their AI-driven operations are explainable. Explainable AI is crucial to building trust, transparency and accountability in decision-making processes conveyed through user-friendly interfaces [62]. Without effective human–machine communication, the potential benefits of AVs—such as enhanced road safety, improved traffic flow and increased mobility—are unlikely to be fully realised. To meet this need, research in the field strives to make AV actions and decisions easily understood and interpreted by humans. AI experts have endeavoured to design models with explainability features and incorporate them into user-centric interfaces, e.g., within the infotainment systems. Numerous literature surveys have scrutinised current approaches from various angles, encompassing cooperative driving, driving scenarios, trustworthiness and computational infrastructure [102, 113, 138, 154, 209]. These studies have put forth recommendations and frameworks to pave the way for future investigations that aim to overcome the hurdles associated with developing explainable AVs.

Atakishiyev et al. offered a thorough analysis of **explainable AI (XAI)** in autonomous driving [6]. Their work described concepts and processes involved in AI explainability and proposed a framework that considers societal and legal requirements for ensuring the explainability of AVs. A survey study by Omeiza et al. reviewed research on explanations for various AV operations [189]. The study identified stakeholders involved in developing, using and regulating AVs. Based on their findings, the authors provided recommendations, including a conceptual framework for AV explainability. In their review, Zablocki et al. discuss the challenges of interpretability and explainability in vision-based autonomous driving [261]. Their study provides a detailed organisation of the *post hoc* explainability methods available for black-box autonomous driving systems. Furthermore, the authors explore several strategies for developing interpretable autonomous driving systems by design. The review by Xing et al. explored human behaviours and cognition in AVs for effective human-autonomy collaboration [256]. The review also delved into the critical factors influencing trust and situational awareness in AVs. Furthermore, the authors provided an analysis of human behaviours and cognition in shared control and takeover situations. Kettle and Lee undertook a comprehensive literature review centred on the **augmented reality (AR)** interfaces implemented in AVs [113]. The review examines the impact of AR interfaces on drivers' situational awareness, performance and trust. While their analysis indicates that integrating AR interfaces in AVs can enhance trust, acceptance and driving safety, the findings underscore the necessity for research on the long-term effects of AR interfaces, impaired visibility, contextual user needs, system reliability and inclusive design.

Despite the growing body of research on explainability for AVs, it remains unclear how to effectively generate and communicate explanations that adapt to diverse user needs and varying levels of transparency across different levels of autonomy. The objective of this study is to assess how task and context, e.g., stakeholders, driving operations and timing, influence the explanation needs and identify situations requiring explanations in various levels of automation. Moreover, we evaluate diverse requirements for communicating explanatory information. After critically evaluating the existing literature and identifying research gaps, we propose **responsible research and innovation (RRI)**-grounded roadmap for future research.[1] With these focus points in mind, our article contributes in the following ways:

[1] RRI is a term used by the European Union's Framework Programmes to describe scientific research and technological development processes that take into account effects and potential impacts on society and the environment.

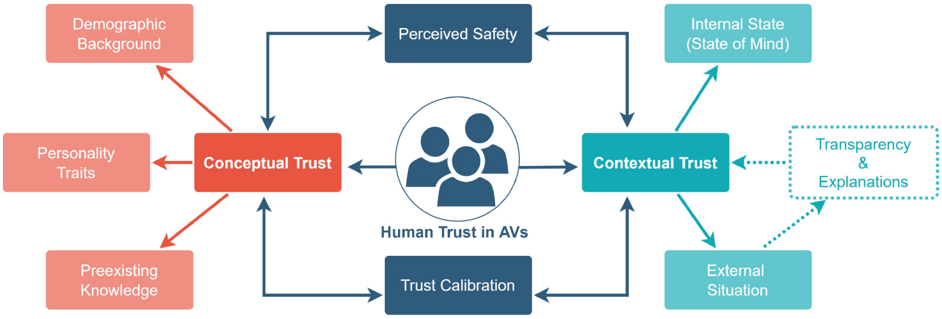


Fig. 1. Our perspective on the factors affecting human trust in AVs and its relation to transparency and explanations.

- Conducting an extensive survey: We provide a comprehensive review of the **state-of-the-art (SotA)** research on explainability in AVs, identifying strengths, challenges and areas where further research is needed.
- Presenting factors that affect explanation generation and communication: We categorise and analyse key factors influencing the explanatory task, information and information communication. This provides a structured framework for researchers to follow when developing explainability solutions.
- Offering RRI grounded roadmap: The roadmap outlines key research directions, ethical considerations and design principles. The application of this roadmap can benefit future research by providing a user-centred and adaptive framework that focuses on the clarity and relevance of explanations.

The rest of the article is organised as follows: Section 2 gives a background on trust factors, transparency and explanations. Section 3 presents the methodology describing the design of the review. Section 4 provides a literature review on the explanatory task, followed by explanatory information (Section 5), and explanatory information communication (Section 6). Section 7 presents the proposed roadmap and discusses a number of future research directions. Section 8 concludes the paper with a summary of key points.

## 2 Background

This section lays the groundwork for the review by discussing factors affecting human trust in AV technology and contextualising its relation to transparency and explanations (see Figure 1). Acceptance and appropriate trust depend on a proper understanding of the technology. One way to attain this is by transparently elucidating its internal processes. Thus, trust and transparency are intrinsically linked. Transparency fosters confidence in the system by providing insights into its reasoning mechanisms, thereby compelling the necessity for explanations.

In the following, we discuss various factors influencing conceptual and contextual trust. While the literature widely explores the impact of explanations on user trust, it is important to note that not all trust factors can be fully addressed through transparency and explanations alone. Understanding these distinct dimensions of trust is critical, as they require different explanatory approaches, e.g., personalisation. When generating explanations, it is essential to consider both conceptual and contextual trust factors to effectively build user confidence.

## 2.1 Trust in AVs

Trust in AVs concerns the reliance, confidence and belief individuals place in the technology to perform a task accurately, reliably and safely. *Perceived trust* is crucial in determining how much an individual is willing to accept, intend to use and depend on the automated system [104, 196]. One prominent definition of trust in automation is provided by Mayer et al. as 'the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability' [169]. Several factors influence trust in automation, e.g., perceived safety, design features, brand, road type and weather. Hoff and Bashir [93] proposed a trust in automation model with three layers of variability: dispositional trust, situational trust and learned trust. Numerous studies have assessed trust factors in the context of AVs across cultures and scenarios. Building on Hoff and Bashir's model, we propose a two-category framework—conceptual and contextual trust—that synthesises findings from the current literature on trust in AVs (see Figure 1).

*Conceptual trust*, paralleling dispositional trust, focuses on general perceptions of trust shaped by individual characteristics, such as demographic background, personality traits and pre-existing knowledge. Studies examining inherent factors, e.g., how demographic variables such as age or education affect willingness to adopt AVs, align with conceptual trust [87, 155, 215]. *Contextual trust*, paralleling situational trust, emphasises dynamic factors that influence real-time interactions with the vehicle, defined by the internal state (e.g., emotional or cognitive state) and external situation (e.g., traffic, weather or vehicle behaviour) [11, 105, 226]. Learned trust bridges both categories. It influences conceptual trust by shaping broader perceptions and beliefs through cumulative experience, which directly relates to pre-existing knowledge [111, 272, 273]. Learned trust affects contextual trust by informing situational expectations and responses based on interactions [200]. This categorisation provides a systematic way to interpret existing literature in a manner relevant to understanding trust in AVs.

Before we delve into these concepts, it is important to point out *perceived safety*'s key role in both conceptual and contextual trust-building processes. These concepts have a bidirectional relationship in that changes in one factor can impact the other. Perceived safety could either be objective, based on an objective evaluation of the safety factors, or subjective, that is, safety based on feeling or perception [148]. Perceived safety is a dynamic process in which levels can shift as individuals encounter new information. For instance, the sudden deviation of AV's driving behaviour from the human driver's norm or even accident reports associated with AVs might hamper the feeling of safety and thereby reduce trust [260]. In full automation (**society of automotive engineers (SAE)** Level 5), where user agency is low, perceived safety is heavily influenced by the user's trust in the system.[2] When users feel they have no control, trust becomes the primary factor in perceived safety. AV systems must demonstrate consistent performance, maintain transparency and communicate effectively with users to appease any unnecessary discomfort caused by lack of control. Therefore, understanding and managing *trust calibration*—the measure of the polarity of trust, that is, misplacement of trust, e.g., overtrust, distrust and mistrust, is critical [127]. When users perceive the system to perform effectively and reliably, their trust increases—or remains stable. Conversely, users may recalibrate their trust if the system exhibits errors or inconsistencies, potentially becoming more cautious or sceptical. Thus, justified trust calibration is crucial to foster effective collaboration and safe interactions between humans and machines.

### 2.1.1 *Conceptual Trust.*

Conceptual trust refers to a person's mental conception and propensity to trust and accept technology over time, which is along the lines of dispositional trust [93]. This

[2] SAE defines levels of driving automation in six levels. The levels are: Level 0—No Automation, Level 1—Driver Assistance, Level 2—Partial Automation, Level 3—Conditional Automation, Level 4—High Automation and Level 5—Full Automation.

type of trust is influenced by demographic background, pre-existing knowledge and personality traits. In relation to demographic information, culture, age, gender and education have been widely studied to understand the influence on acceptance and trust in AVs. People from different cultures drive by different implicit and explicit rules. For instance, the work by Edelman et al. work reports that participants from different backgrounds respond differently to overtaking and hindering decisions of the AV [63]. Though some similarities can be drawn among cultures to form a design framework applicable across the globe, an AV is not expected to have the same behaviour worldwide and be understood and accepted by everyone equally well. Cross-cultural differences require AVs in different places to adapt to the culture and transition from one behaviour to another when it travels through cultural borders [201]. Additionally, culture influences key psychological factors such as propensity to trust, which affects how quickly users may place confidence in AV systems [71], and locus of control, which shapes individuals' sense of control over the driving environment and technology [160]. These variations highlight the need for culturally sensitive AV design to accommodate diverse user expectations.

Regarding the age factor, research revealed no significant differences between the average trust ratings of younger and older adults [87, 155, 215]. Both younger and older participants reported autonomous driving as trustworthy and acceptable. Liu et al. reported that the younger participants had a higher positive attitude and acceptance than the older participants [155]. At the same time, Hartwich's study revealed that older people also had a strong positive attitude toward AVs [87]. This suggests that age may not be a significant factor on its own but could be a proxy for other underlying variables. Further research is needed to assess if consistent age differences exist in individual trust in AVs. In terms of gender differences, previous studies suggest that men are more inclined towards accepting AVs [206, 274]. Regarding the education level, Liu et al. showed that education was positively correlated with respondents' willingness and ability to pay for AV technology [155].

On another note, research has explored how personality traits, particularly the big five, influence the acceptance of AVs. People with high openness, conscientiousness, extraversion and agreeableness had a more positive perspective, while high neuroticism had a negative effect [99, 205]. Personal innovativeness, considered a relevant personality trait influencing one's willingness to accept and implement new ideas, products and systems [3], is also perceived to positively influence the adoption of AVs [90]. Risk preference (risk-seeking or risk-averse personality) is also considered an important factor in AV adoption. The research shows that risk preference is highly influenced by age, income and education [246].

Pre-existing knowledge is another factor influencing trust formation even before an actual interaction with AVs occurs. Prior experience with similar technology—or other transport modes and information conveyed by external sources (e.g., media)—could lead to overtrust, distrust, or mistrust toward AVs [197, 240]. Previous studies reported that participants with pre-existing knowledge of AVs held more optimistic views towards using AVs in the future than those without such knowledge [111]. Research by Shammut et al. also highlighted that respondents who tended to trust AVs commented on how safe and reliable AVs are based on their pre-existing knowledge [225]. Another work revealed that responders who reported greater knowledge and experience of new technologies were more accepting of AVs [140]. On the other hand, the study by Zhou et al. revealed that misinformation could be the most significant factor causing distrust in AVs due to risks associated with software malfunctions or cyberattacks [272]. According to Zhou et al., negative news coverage about accidents involving AVs could cause people to distrust them [273]. Conversely, overstated assurances or overpromising marketing claims can lead to developing unrealistic expectations about the safety and capabilities of AVs [77]. While initial trust is formed based on pre-existing knowledge and previous experiences with technology, interacting with AVs and understanding their capabilities and limits generates proper trust calibration.

### 2.1.2 *Contextual Trust.*

Contextual trust is determined by the particular situation or location in which an interaction occurs, which aligns with situational trust [93]. Contextual trust is consequential and temporal, influenced by both the *external* (e.g., driving behaviour) and *internal* environment (e.g., state of mind) [112]. People might feel positive trust toward a system concerning certain tasks and goals yet mistrust or distrust when other tasks, goals, or situations are involved. For instance, existing literature highlights contextual elements, such as traffic signals and driving behaviour (e.g., defensive, normal and aggressive), are important external factors related to contextual trust in pedestrian-AV interaction [40, 105, 241]. Research showed that pedestrians generally expressed more trust in AVs at signalised crosswalks than in unsignalised crossing. This finding is based on the assumption that AVs are expected to be law-abiding (e.g., stopping at red lights) under all circumstances [105].

In future, human-driven vehicles and AVs will inevitably share the road. A recent work studied how human drivers' trust toward AVs varies with contextual attributes such as traffic density and road congestion. Sun et al. discovered a positive correlation between the vehicle gap and trust level, while factors such as relative speed and traffic density exhibited a negative correlation [233]. Osikana et al. found that the vehicle's speed and distance have the most influence on the cyclists' trust levels [190]. Another study found that motorcyclists trust AVs more than human drivers due to human unpredictable driving behaviours, not checking their blind spots and speeding [194].

Researchers have investigated the influence of perceived risk on drivers' trust in automated driving. Ayoub et al. examined drivers' trust in AVs' takeover scenarios with different system performances [10]. The results show that drivers quickly calibrate their trust level when AV clearly demonstrates its limitations. Azevedo-Sa et al. considered two risk types (i.e., low system reliability and low visibility from foggy weather) as contextual factors where low visibility did not significantly impact drivers' trust in automated driving [11].

Trust develops inconsistently in different contexts due to *internal factors* such as emotional state, attentional capacity and self-confidence [163, 258]. The attentional capacity of a driver often depends on the task—monotonous automated driving promotes passive task-related fatigue [126]. A high level of positive trust in automation might reduce attention allocation and situational awareness. Inactivity and engagement in **non-driving-related tasks (NDRTs)** have been suggested to impair the driver's ability to handle the takeover process safely, which might cause accidents resulting from over-trusting automation [180, 198]. In another work, contextual trust was associated with emotions, where a high level of trust in AVs significantly improved participants' positive emotions [7]. Furthermore, factors such as motivation, stress, sleep deprivation and boredom influence trust dynamics in driving scenarios [93].

A non-task-related condition, **attention deficit hyperactivity disorder (ADHD)**, could influence trust in AVs due to the cognitive challenges it poses in monitoring and maintaining awareness [126]. Since people with ADHD are more prone to distraction, their ability to monitor automation and respond appropriately might be compromised, further impacting their willingness to use automation. On a different note, self-confidence is another context-dependent variable that can alter trust in automation and control allocation [93]. A recent study compared participants who underwent training for engaging automated driving systems with those who did not receive such training. The findings unveiled notable disparities in response to emergency events requiring drivers to assume control. Trained drivers demonstrated reduced reaction times and exhibited more calibrated trust levels in automation compared to their non-trained counterparts [200]. Overall, it is essential to consider internal state factors along with locus of control and propensity to trust when assessing trust in AVs, as they collectively shape individuals' perceptions and acceptance of the technology.

As discussed above, some factors influencing trust in AVs may not be addressed by training or technological advancements. Regardless, certain elements are central to establishing trustworthiness objectively, including technical competence, adept situation management and transparency [33, 179]. Technical competence refers to the user's belief that the AV meets performance and reliability expectations (e.g., GPS connection, security and ability to choose the best course of action). Situation management relates to the user's belief that they can gain control over the vehicle or contact a human operator whenever required. Transparency provides a clear view of an AV's abilities and operations to promote appropriate trust and discourage automation misuse and disuse. In the following, we delve into the transparency requirement further.

### 2.2 Transparency

System transparency refers to the degree of observability and predictability of the operations to give a sense of control and acceptance in autonomous driving [179]. Transparency is usually a precondition for *accountability* in which AV should be able to explain its plans and actions, especially in unexpected events. It concerns the extent to which the responsibility for the outcome can be ascribed to an agent (e.g., governments, companies, system developers, drivers) legally or ethically [242]. This is because, for an action to be evaluated properly, relevant stakeholders should have access to all necessary information.

Transparency requirements vary depending on the levels of autonomy (e.g., SAE Level 2 vs. Level 5) [217]. In semi-autonomous systems, transparency is required when the vehicle needs to hand the control to the driver at certain stages of the journey. In the highest level of vehicular automation, where vehicles can handle all traffic situations, transparency requirements may not necessarily be safety-critical, but in the attempt to regulate trust in the automation, thereby improving user experience and comfort [55, 66]. Regardless of the level of autonomy, AVs in action are expected to communicate information about the driving decisions, reasons and plans transparently in a timely manner to calibrate an appropriate level of trust [256].

### 2.3 Explanations

Explanations are a mechanism to promote greater transparency and help address problems caused by the black-box nature of automated decision-making. The literature has extensively studied explanations for AV's behaviour to improve transparency of system decisions and trust. The research explored transparency and explanations for situations including system uncertainty [91], perceived risk (e.g., weather and driving speed) [82], intent communication for vulnerable road users [54], motion planning [81], driver takeover request [125] and lane positioning [119].

We categorised the literature on AV explanations into two primary approaches based on their intended focus and purpose: 'user-centric' and 'technical-centric'. User-centric explanations prioritise conveying information to users (e.g., the general public or non-experts) in a comprehensible and meaningful manner. Research highlights that users benefit from simplified, legible explanations that enhance comprehension and trust in the system [181]. These explanations involve intuitive visualisations and simple language to communicate complex technical decisions and are typically assessed through user studies conducted in virtual or simulated environments. User-centric explanations centre around the driving context and the kind of information or interaction that is relevant. Contexts (e.g., near-crash) and individual attributes (e.g., aggressive driving style) significantly influence the desire for explanation. A study showed that people tend to agree on the need for an explanation for near-crash or emergency driving scenarios and less for the ordinary driving

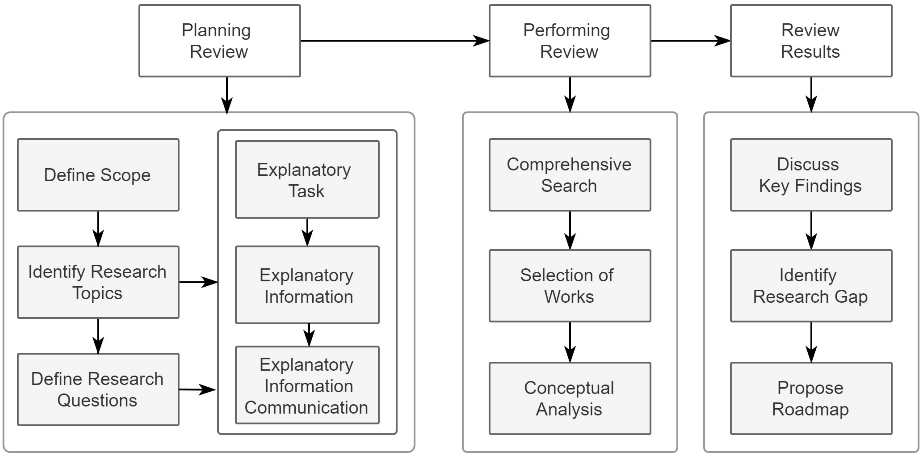


Fig. 2. The SotA review approach.

situation [226]. The context also has a strong influence on the types of questions the AV may receive. Omeiza et al. explored explanation types (i.e., causal and non-causal) and their investigatory queries (e.g., why, why not, what if and how), demonstrating 'why-not' (i.e., contrastive) explanations have the most positive impact on participants' understandability [188].

Technical-centric explanations, on the other hand, focus on the intricate details of how the AV operates, emphasising interpretability and exploring technical aspects such as the sensor data and algorithms used in decision-making [47, 115, 144, 193]. These explanations are more geared towards experts or individuals with technical backgrounds who seek in-depth insights into the underlying mechanisms of the AV's actions and are evaluated through quantitative metrics. Often, a combination of both user and technical-centric approaches might be necessary to address various user needs and levels of expertise.

## 3 Review Methodology

This study conducts a review of the design considerations for explainable AV systems by adopting a 'SotA review' approach [79]. This approach focuses on analysing the most recent advancements in the field, identifying emerging trends and contrasting recent advancements with established studies. The SotA review is particularly well-suited to this work as it offers more flexibility than formal methodologies such as systematic reviews. It allows the inclusion of the latest findings from diverse sources (e.g., preprints, technical reports) without being constrained by rigid protocols. It also enables to focus on the criticality of studies without the formal quality assessments typically required in systematic reviews. Figure 2 illustrates the key steps involved in the review process.

The scope of this study is limited to the explainability of AVs, focusing specifically on the unique challenges in this domain rather than the broader field of explainable AI. To address the challenges of explainability in AVs, we conceptualised the explainability framework into three key dimensions—*explanatory tasks*, *explanatory information* and *explanatory information communication*—representing *why to explain*, *what to explain* and *how to explain*, respectively. These dimensions guide the formulation of research questions that aim to explore the unique requirements of AV explainability. Below, we outline how each dimension identifies with specific research questions:

*Explanatory Tasks.* This dimension focuses on the purpose and factors motivating the need for explanations in AV systems. Research on explainability highlights the importance of identifying the contexts and triggers where explanations are essential, particularly in dynamic and safety-critical situations [24, 250]. The following research questions examine the core elements that define why explanations are required, including stakeholders, driving operations and the function of explanations. It also addresses the need to adapt explanations to the urgency and stakes of specific scenarios:

RQ1. What factors are integral to the explanatory task?
RQ2. How do timing and situation criticality influence the explanatory task?

*Explanatory Information.* This dimension explores the content of the explanation, focusing on the types of information users need to understand AV behaviour. Prior studies highlight the importance of structuring information across layers of transparency, e.g., abstract information vs. causal reasoning [28, 253]. The following research questions examine how to organise explanatory information to meet diverse user needs and how this structural approach is considered within explanatory tasks and context:

RQ3. How is the explanatory information structured across different layers of transparency?
RQ4. How do task and context influence the explanatory information across transparency layers?

*Explanatory Information Communication.* This dimension addresses the modalities and methods for delivering explanations to different stakeholders. Research highlights the importance of tailoring communication methods (e.g., visual, verbal and multimodal) to meet varying stakeholder needs [86, 254]. Explanation requirements for internal stakeholders differ from those for external stakeholders. These research questions examine how explanations can be optimised for various users and explore methods to ensure clarity and usability in these communications:

RQ5. How is the explanatory information communicated to internal stakeholders?
RQ6. How is the explanatory information communicated to external stakeholders?

To obtain relevant articles, we applied the keywords ‘explainable + autonomous + vehicles + driving’ and ‘transparency + feedback + communication + autonomous + vehicle + car + driving’ on Google Scholar for publications between 2018 and 2023. Our initial search yielded 169 papers; through the review process, we eliminated papers that lacked a component of explainability/transparency, only presented a conceptual framework, or were duplicates. This left us with 135 relevant papers. After critically evaluating these papers and identifying research gaps, we formulated a conceptual roadmap for future research tackling advancing explainable AV systems.

The roadmap is grounded in a RRI framework. RRI frameworks provide guidelines and best practices to promote inclusive and sustainable research and innovation design, proactively addressing challenges before they arise when deploying technologies in the broader society [21]. The roadmap presented here follows the AREA framework, which stands for *A*nticipate, *R*eflect on**,** *E*ngage with and *A*ct upon [238]. Each holds a specific role in the research and innovation process, as follows:

*Anticipate*: Conducting an in-depth analysis of the intended or unintended consequences that may arise. This aims to facilitate the exploration of potential implications that may otherwise remain overlooked and under-discussed.

*Reflect*: Reflecting on the objectives, motivations and potential implications of the research, along with any unknowns, gaps in knowledge, presuppositions, dilemmas and societal shifts that may arise.

*Engage*: Encouraging diverse perspectives and inquiries to be shared and discussed in a collaborative and inclusive manner.

*Act*: Utilizing these conversations to shape the course of the research and innovation process itself.

Taken together, we propose potential areas of improvement related to each research topic and then discuss these further within the AREA framework in Section 7.

## 4 Explanatory Task

Explanations serve different purposes depending on the setting. Various factors influence the context and purpose, shaping the explanatory task. We reviewed the relevant literature to discuss these factors and circumstances that affect the explanatory task, specifically addressing RQ1 and RQ2.

### 4.1 RQ1: What Factors Are Integral to the Explanatory Task?

The literature discusses factors influencing the explanatory task based on intended stakeholder, driving operation and function of explanation depending on the level of autonomy. Table 1 provides a comprehensive overview of the literature on these factors and their respective variants. The following section offers an in-depth analysis of the literature through the lens of these factors.

*4.1.1 Stakeholders.* Within the autonomous driving domain, explanations are utilised for multiple purposes, resulting in various types of stakeholders. The level of detail, form and mode of communication necessary for an explanation vary depending on the target audience and the explanation's purpose. As such, it is vital to tailor the explanation to meet the specific needs of the intended stakeholders. Omeiza et al. identified three primary stakeholder categories: Class A (including all end-users and society), Class B (consisting of technical groups such as system developers) and Class C (encompassing regulatory bodies, such as guarantors) [189].

We further divided Class A into two main classes of users: vehicle internal and external stakeholders. Internal stakeholders include drivers who participate in driving operations and passenger(s) who may interact with the AV but are not responsible for its operation. External stakeholders can be categorised into different groups based on their relation to the AV and the risks they face. Remote operators assist or control the AV from a distance, where they are not exposed to physical risks but still carry legal responsibilities and emotional stress [85]. In contrast, vulnerable road users, such as pedestrians, cyclists, motorcyclists and mobility scooter users and protected road users, including drivers of cars, trucks and buses, are present in the AV's environment and face varying levels of physical risk [95]. The results revealed that out of 135 papers, only 67 explicitly specified the intended stakeholders for the explanations. These were categorised as follows: driver (32 papers), passenger (8 papers), both driver and passenger (7 papers), pedestrian (16 papers) and other vehicles (4 papers). The remaining papers did not specify a target stakeholder; however, it was apparent that these studies were primarily aimed at technical users to explain model behaviour. Technical-centric explanations are relevant to Class A users as well; however, further research is necessary to study how these approaches can be adapted to meet the end-user requirements and integrated into appropriate **human–machine interfaces (HMIs)**. Moreover, previous studies have not yet touched upon integrating stakeholder observations/feedback into explanations.

*4.1.2 Driving Operations.* Another important factor that influences the explanatory task is the driving operation. AVs, equipped with sensory capabilities to perceive their surroundings, rely on a series of interconnected operational stages to make informed driving decisions in real-time. These stages encompass perception, localisation, planning and control.

Table 1. Classification of Reviewed Literature Based on Factors Integral to the Explanatory Task

| Task and Context | Variants | SAE Level 1-2-3 | SAE Level 4-5 |
|---|---|---|---|
| Stakeholders | Internal | User-centric [8, 29, 30, 38, 40, 43, 52, 76, 78, 91, 129, 131–135, 174, 186, 216, 221, 227, 254, 262, 268, 269] Technical-centric [14, 17, 24, 26, 47–51, 56, 92, 98, 106, 107, 120, 130, 141, 144, 149, 150, 158, 166, 177, 195, 230, 249, 255, 264, 266] | User-centric [36, 41, 58, 64, 70, 81, 86, 110, 119, 136, 187, 202, 222, 223, 232, 235, 244, 245, 251, 267] Technical-centric [1, 2, 15, 18–20, 22, 23, 25, 31, 32, 69, 97, 108, 115–118, 121, 137, 146, 152, 153, 164, 165, 184, 185, 188, 193, 204, 208, 211, 218, 220, 224, 231, 239, 257, 259, 265, 270] |
| | External | User-centric [9, 84, 96] | User-centric [34, 35, 37, 39, 42, 68, 83, 109, 139, 161, 182, 191, 228, 241, 271] Technical centric [210, 213] |
| | Observations | ø | ø |
| Driving operations with explainability | Perception | User-centric [38, 43, 52, 254] Technical-centric [24, 49, 177, 195] | User-centric [35, 119] Technical-centric [1, 31, 110, 117, 118, 121, 165, 184, 208, 218, 232, 239, 257, 259] |
| | Planning | Technical-centric [17, 50, 51, 141, 166, 266] | Technical-centric [2, 18, 19, 23, 146, 152, 210, 211, 213, 265] |
| | Localisation | Technical-centric [255] | Technical-centric [25] |
| | Control | User-centric [8, 9, 29, 30, 40, 76, 78, 84, 91, 96, 129, 131–135, 174, 186, 216, 221, 227, 262, 268, 269] Technical-centric [14, 26, 47, 48, 56, 92, 98, 107, 120, 130, 144, 149, 150, 158, 230, 249, 264] | User-centric [34, 36, 37, 39, 41, 42, 58, 64, 68, 70, 81, 83, 86, 109, 136, 139, 161, 182, 187, 191, 202, 222, 223, 228, 241, 245, 267, 271] Technical-centric [15, 20, 22, 31, 32, 69, 108, 116, 137, 153, 164, 185, 188, 193, 204, 220, 224, 231, 235, 270] |
| | Integrated | Technical-centric [106] | User-centric [244, 251] Technical-centric [97, 115] |
| Function of explanations | Situation awareness | User-Centric [8, 76, 78, 129, 132–135, 221, 268] Technical-centric [14, 17, 24, 26, 47–51, 56, 92, 98, 106, 107, 120, 130, 141, 144, 149, 150, 158, 166, 177, 195, 230, 249, 255, 264, 266] | User-centric [36, 70, 187, 251] Technical-centric [1, 2, 15, 18–20, 22, 23, 25, 31, 32, 69, 97, 108, 110, 115–118, 121, 137, 146, 153, 164, 165, 184, 188, 193, 204, 208, 211, 218, 220, 224, 231, 232, 235, 239, 257, 259, 265, 270] |
| | Safe mode transition | User-centric [29, 30, 40, 52, 131, 174, 227, 262, 269] | User-centric [136, 202] |
| | Trust calibration | User-centric [38, 43, 91, 186, 216, 254] | User-centric [41, 58, 64, 81, 86, 119, 222, 223, 244, 245, 267] Technical-centric [152] |
| | External interaction | User-centric [9, 84, 96] | User-centric [34, 35, 37, 39, 42, 68, 83, 109, 136, 139, 161, 182, 191, 202, 228, 241, 271] Technical-centric [210, 213] |
| | Commentary driving | ø | Technical-centric [185] |

Perception is the sensing of an operational environment defined as a combination of two tasks: road surface extraction and on-road object detection through sensors such as vision, LiDAR, RADAR and ultrasonic technology. Gradient-based methods are the most commonly used approaches for scene understanding tasks [1, 121, 165]. Various explainability approaches have been developed for early anticipation of traffic accidents [173], risk assessment in complex road conditions [259] and segmentation under hazy weather [218] based on the vision data. Explanations are explored for LiDAR and RADAR data to detect 3D objects [195] and sensor input uncertainty for possible deceptive attacks [208]. Other works have proposed explainable localisation and mapping by extracting non-semantic features from LiDAR scans to accurately determine AV's position in the environment [25, 255].

As the AV perceives its surroundings and gets its precise localisation, it plans the trajectory from the initial point to the final destination. Planning is a complex operation in that the amount of data the AV processes per time unit is hard to keep track of continuously and accurately. Research in this field has focused on either route planning (i.e., selection of a route) [2, 152, 211] or behaviour planning, which involves anticipating and interacting with other road users who share the same trajectory [17, 141, 266, 269]. Much of this research has focused on *explicable planning* models in which the outcomes can be explained through action implicitly. The process requires further translating the AV's updated plans into comprehensible and user-friendly formats.

Finally, control of an AV involves the execution of planned motions. Feedback controllers mainly manage this function by interacting with the sensors and assisting the car in controlling its trajectory along the journey. The review results show that explanations are generated and explored mainly for the control and navigation-related tasks, i.e., longitudinal control (speed regulation) and lateral control (steering operation). Having that the driving operations are interconnected, the explanation generation process for control and navigation operation is coupled with perception and planning tasks. Explanations are explored for each and every driving action, such as move, stop and lane change in urban areas [108, 186]. From the technical perspective, the aim of generating explanations for each action taken by AV is to understand how the model would perform under unseen environments and unexpected situations [15, 69, 120]. From the user-centric perspective, explaining every action helps build the user's initial trust and acceptance that the AV is making reasonable decisions aligned with their expectations. In summary, our review reveals that research has primarily focused on control and perception tasks, particularly for higher-level automation while lacking in-depth studies for explanations for planning and localisation tasks.

#### 4.1.3 *Function of Explanations.*

The function of explanation directly supports the purpose of the explanatory task. In addition, it is important to consider the level of autonomy inherent in the system to understand the function of explanations. While vehicles with higher levels of autonomy may sometimes require fewer explanations, the opposite can also be true—more agentic systems might need to provide more explanations due to reduced driver engagement. The greatest need for explanations may arise at intermediate levels of autonomy, where the driver is required to supervise or partly control the vehicle, but is less directly engaged in the driving task. This discussion highlights the function of explanations based on the level of autonomy, whether in collaborative driving or full autonomy scenarios.

In low-level automation (SAE Levels 1-2-3), the driver's role evolves across different levels. At SAE Level 1, the driver maintains partial control of the vehicle, either handling longitudinal tasks (accelerating, braking) or lateral tasks (steering), while the system assists with the other. In SAE Levels 2-3, the driver takes on the role of a supervisory controller, monitoring the performance of the driver assistance system and intervening when necessary. At these levels of autonomy, explanations are essential to help the driver understand when and how they need to interact with the AV in the context of shared control. The function of explanations in various scenarios within Levels 1-2-3 can be outlined as follows:

*Maintaining situation awareness*—Maintaining situation awareness is essential for the driver to take control at any point of the journey. The driver continuously monitors the driving behaviour and makes necessary adaptations if the current driving behaviour differs from their desired actions [250]. Explanation mechanism keeps the human driver informed in case of system uncertainty [129, 134, 136], perceived risk [22, 24, 164, 177, 231] and traffic complexity [2, 86].

*Safe transition to manual driving*—In complex traffic scenarios, a takeover request without context may not effectively convey the gravity of the situation, and the driver may take over control without fully comprehending the situation. This can lead to a decrease in the quality of takeover control [269]. A study by Chen et al. suggests that situation influences participants' decisions to take over control when observable cues are not available in the driving environment [30]. It is necessary to provide clear explanations to persuade the driver to take the necessary action to avoid potential hazards, e.g., the system sends an alerted message: 'Automated driving is about to be disabled due to the busy intersections, switch to manual driving mode and proceed with caution' [59, 181].

*Supporting trust calibration*—Another function of explanations is to convey the AV's performance limits to support trust calibration [132, 135]. Explanations provide a clear view of AV's abilities and operations to promote appropriate trust and discourage misuse of automation by over-reliance on it or disusing it due to under-trusting it. Goldman and Bustin show that the explanation presenting risk and the next plan reduces participants' willingness to take manual control, increasing enjoyment of automated driving and preventing disuse of automation [76]. In a study conducted by Helldin et al., the trust analysis showed that participants who received uncertainty information trusted the system less than those who did not, indicating a more proper trust calibration than in the control group [91].

In Level 4 automation, the vehicle can complete an entire journey without human intervention in specific conditions or environments; however, intervention may still be needed in exceptional situations with sufficient lead time.[3] Certain applications in this category may not necessitate a human driver. At Level 5 (full automation), no human intervention is required, and the vehicle can operate under all circumstances [72, 86]. In Level 4 and 5 automation, the function of explanations concerns both the internal and external stakeholders. The role of explanations includes:

*Maintaining situation awareness*—Situation awareness is also relevant in high-level automation, particularly in Level 4 automation, where occasional human intervention may be necessary. Providing explanations can facilitate a safe transition from automated to manual driving [136, 202]. Research has demonstrated that explanations effectively convey functional and system-wide uncertainties in diverse driving scenarios, such as construction zones, unclear lane markings, heavy traffic, animal crossings in wooded areas and sudden lane changes [58, 244]. Moreover, situation awareness cues can enhance engagement in NDRTs and increase the enjoyment of automated driving [70]. In full autonomy, explanations create a feeling of safety and control, elevating the passenger experience and trust, particularly in complex traffic situations such as intersections, road obstructions, construction sites, pedestrian zones, slow-moving bicycles and ride delays [43, 64, 86, 222, 223].

[3] Sufficient lead time should account for the driver's reaction time, the complexity of the driving task and the current driving conditions, to prevent any potential hazards that may emerge after the takeover request is issued.

*Increasing driving knowledge*—Commentary driving is a technique that involves drivers verbalising their observations, assessments and intentions. Explanations are explored as part of a driving commentary that mimics how a human instructor would explain driving behaviour [185]. It involves aspects such as driving speed, timing of overtaking another vehicle, distance to object, change plan and lane positioning. By articulating the driving actions out loud, novice drivers can gain a better understanding and awareness of their surroundings [207]. This would also help learners better understand the driving processes and increase their driving knowledge through conversational explanations. Moreover, research has shown that training in commentary driving improves responsiveness to hazards in a driving simulator [44].

*Safe interaction with external stakeholders*—Explanatory signals provided through visual displays play a crucial role in compensating for the absence of direct interaction with human drivers, typically involving gestures and eye contact with other road users. In scenarios such as four-way intersections and pedestrian crosswalks, where communication between AVs and other road users is essential, effective communication mechanisms become even more critical [9]. One way to improve interaction is by indicating the vehicle's driving mode, whether it is in automated or manual mode. This could foster transparency and help pedestrians understand what to expect from the AV and adjust their actions accordingly [109, 228]. It would reduce the frustration for internal stakeholders and make it clear to others around the vehicle that the onboard users are not actively controlling the vehicle and have no immediate agency over its actions [250]. External communication is also important in ridesharing scenarios, as it helps users identify the AV and understand its intent, especially when there are multiple AVs on the road [191].

Our analysis revealed that the literature predominantly addressed explanations related to Levels 4 and 5 automation, with limited coverage for Level 1-2-3 across all aspects (see Table 1).

## 4.2 RQ2: How Do Timing and Situation Criticality Influence the Explanatory Task?

According to our findings, explanations can be classified based on their timing as either proactive or reactive. Proactive explanations are given in anticipation of future needs, while reactive explanations are generated in response to a request after an event has occurred. Proactive explanations are particularly important in low-level automation, where the timing of warnings for potential interventions is a key concern. Studies have shown that providing explanations before the AV takes action promotes more trust than explanations provided after the event [88, 122]. Another aspect to consider is the criticality of the event, as it influences the optimal timing for delivering explanations [226].

Regarding criticality, we have identified two types of situations that require explanations: critical situations that threaten human health and life and non-critical situations that involve harmless but questionable driving behaviour and style. Adapted from Schneider et al., we discuss the timing and criticality of explanations for four different categories of driving situations: proactive and reactive explanations, which can be generated for critical or non-critical events [221]. It is essential for AVs to be capable of providing all these four categories of explanations in relevant contexts. Table 2 classifies the reviewed articles according to the timing and criticality of the explanations.

### *4.2.1 Proactive Explanations in Non-Critical Situations.*

This type of explanation aims to increase situation awareness for internal stakeholders by supporting their understanding of 'what is going on'. This includes providing information about the elements within the environment, their meaning, and a projection of their status in the near future [262]. External stakeholders also benefit from the proactive display of the driving mode of the AVs in inconsequential situations [67, 109]. Proactive

Table 2. Timing and Situations Requiring Explanations

| **Timing and Situations** | **SAE Level 1-2-3** | **SAE Level 4-5** |
|---|---|---|
| Proactive critical | User-centric [8, 9, 29, 30, 40, 52, 76, 84, 91, 96, 129, 131–135, 174, 227, 262, 268, 269] Technical-centric [50, 177] | User-centric [34, 36, 37, 39, 41, 42, 58, 68, 83, 110, 136, 139, 161, 182, 187, 191, 202, 228, 241, 244, 267, 271] Technical-centric [2, 115, 164, 193, 208, 210, 231, 232, 235] |
| Proactive non-critical | User-centric [38, 43, 78, 186, 216, 221, 254] Technical-centric [14, 17, 26, 49, 92, 98, 130, 150, 158, 166, 249] | User-centric [64, 70, 86, 109, 245] Technical-centric [18, 32, 97, 117, 137, 211, 239, 270] |
| Reactive critical | Technical-centric [24, 106] | User-centric [81, 222] Technical-centric [22, 146, 188, 213, 259, 265] |
| Reactive non-critical | Technical-centric [47, 48, 51, 56, 107, 120, 141, 144, 149, 195, 230, 255, 264, 266] | User-centric [35, 119, 223, 251] Technical-centric [1, 15, 19, 20, 23, 25, 31, 69, 108, 116, 118, 121, 152, 153, 165, 184, 185, 204, 218, 220, 224, 257] |

explanations in non-critical situations are intended to improve the overall user experience and comfort. Nevertheless, further research is needed to identify which non-critical situations require proactive explanations. Based on the literature, providing stakeholders with proactive explanations in non-critical situations could generate the following benefits:

*Acceptance and trust for new users*—As discussed in Section 2.1, acceptance and trust rely on an appropriate understanding of the technology. One way to achieve this is by expressing its internal processes proactively. Kuhn et al. and Omeiza et al. presented this type of explanation for normative driving actions, including move, stop and lane change, e.g., 'Car is stopping because the traffic light is not green on ego's lane', 'Stopping because cyclist stopped on my lane' [130, 186]. Ruijten et al. added anthropomorphic features to such explanations where AV perform a commentary driving (e.g., 'We're on a cobbled road with pedestrians, I'm slowing down.') [216]. Their findings show that adding a layer of human likeness to an explanatory agent made it perceived to be more intelligent and appealing. Although proactive explanations for non-critical situations help build initial trust and allow constant control over driving behaviour, they quickly become overwhelming and create a high workload for the user [70, 136]. Hartwich et al. suggested adapting this to the user-specific needs to gain broader acceptance rather than providing complete information permanently as a universal standard [86].

*Avert frustration and surprise*—In high-level automation, proactive explanations are necessary for onboard users who do not have any control over driving behaviour. This type of explanation helps to ease frustration and minimise the need for users to constantly ask about the AV's decisions [75]. For instance, in a scenario where the AV stops for longer than usual, it displays a message 'Yielding to pedestrians!' for the riders, giving a view of where the pedestrians are in the scene and how it's reacting [247].

4.2.2 *Proactive Explanations in Critical Situations.* Proactive explanations in critical situations play a pivotal role in ensuring the safety of everyone involved, particularly in cases where the driver is required to assume manual control. In such cases, the system should alert the driver regarding potential high-risk situations or the possibility of a failure [215]. For other road users, proactive communication of the AV's intention is crucial to ensure safe and efficient interaction with them. The purpose of proactive explanations in critical situations is twofold, as discussed below:

*Alert/prepare driver to takeover*—Proactive explanations for critical situations require drivers to pay more attention to the road situation and the recommended action to prepare for a potential takeover event [142]. Chen et al. proposed three types of critical situations that may lead to a takeover event: operational problems, system limitations and unexpected events [30]. Operational problems are related to the AV's internal functions (i.e., operating system problems involving sensors, computation and communication between hardware and software) that provide no visible cues within the driving environment. System limitations occur when the AV approaches its operation limits due to various factors such as road or environmental conditions, e.g., construction site, highway exit, traffic jam, foggy weather, road with bends or poor lane markings [174]. Unexpected situations include where the AV may be uncertain about its detection or classification of certain road elements and events (e.g., cut-in vehicle or stopped vehicle cause emergency manoeuvres [134], failure prediction [231] and accident anticipation [173]).

*Intent communication for external stakeholders*—Research has extensively examined the external communication of AVs in potentially unsafe situations, e.g., unsignaled intersection [67, 161], zebra crossing, parking lot [83] and four street crossing scenarios [182] involving pedestrians, cyclists, skateboarders, individuals with disabilities and other vehicles. Various design elements have been explored in online and simulated environments to facilitate unambiguous communication between AVs and other road users [37, 84, 139, 241]. However, it is still unknown how AVs should communicate proactively in critical situations and whether explanations are needed immediately after a critical situation [122].

*4.2.3 Reactive Explanations in Non-Critical Situations.* Reactive explanations are initiated by users in response to unexpected behaviour exhibited by the AV, especially when no apparent cause is evident in the environment. Research has explored various types of questions users may pose, including inquiries about what the system did, why it acted in a certain way, why it did not do something else and how it arrived at its decision. A study by Graefe et al. found that 'why' and 'how' questions are preferred for enhancing transparency, understandability and predictability [78]. Similarly, Wiegand et al. found that users expressed a desire for reactive explanations to influence the AV's driving behaviour interactively [250]. For instance, users may inform the vehicle that a pedestrian will not cross the street, enabling the AV to proceed safely. Such interaction could potentially correct the AV's driving behaviour. Another research has proposed the use of conversational agents, where AVs provide brief proactive explanations for unexpected driving behaviour, allowing users to inquire for further information about the situation [223].

Several technical-centric studies introduced *post hoc* explanation methods to justify a driving action retrospectively in response to a request [25, 164, 166]. However, much of this work is not comprehensible to non-technical users, which requires them to be integrated into HMIs and further assessed for their usability with human-subject studies.

*4.2.4 Reactive Explanations in Critical Situations.* Reactive explanations for critical situations involve events ranging from near misses to incidents compromising human health and lives. However, existing literature has not sufficiently explored reactive explanations concerning critical situations. Wiegand et al. conducted a study on unexpected driving scenarios that do not necessarily result in human injury, hardware damage, or an interruption to the regular operation but cause a great deal of frustration and dissatisfaction for the user [250]. Some of these scenarios in which the participants requested an explanation include unnecessary lane changes that almost lead to collisions, abrupt stops during turns, unexpected stops, sudden acceleration and deceleration and unclear interaction with pedestrians.

Moreover, reactive explanations provide evidence for post-incident forensic analysis, as they can answer questions that help determine liability-related issues arising from autonomous decisions [192, 208]. Explainability techniques are currently capable of providing answers to the following questions:

- *What happened?* This involves presenting a factual chain of events that demonstrates causation.
- *How did it happen?* This requires establishing an unbroken chain of causation between the defendant's negligence and the claimant's injury.
- *Why did it happen?* This involves identifying the breach of duty.

By answering these questions, XAI techniques can help explain the events and the chain of events to establish the factual and legal causation required by common and civil law systems. Consequently, such explanations can fulfil the obligations to establish causation, aiding in identifying the responsible party or the percentile of responsibility.

## 5 Explanatory Information

Explanatory information refers to the information provided by an AV to make its behaviour and decision-making processes understandable to users. Explanatory information is a component of 'transparency', which refers more broadly to the extent to which the system makes its processes visible to the user [202]. In some contexts, AV actions may appear self-explanatory such as when the vehicle stops at a red light. However, in more complex scenarios, users may require deeper insight into the AV's decision-making process to feel confident in its actions [27]. For example, users may need explanations related to sensor inputs, system beliefs and anticipated outcomes to achieve a higher level of transparency. Consequently, transparency is achieved through multiple layers. In this section, we delve into the type of explanatory information involved in different layers of transparency and task factors that influence the explanatory information in different layers of transparency, addressing RQ3 and RQ4.

### 5.1 RQ3: How Is the Explanatory Information Structured across Different Layers of Transparency?

In exploring the literature on transparency in autonomous systems, one notable resource is the IEEE standard, which defines transparency requirements at progressive levels [253]. At the lower levels, transparency is established by providing documentation about the system prior to any interaction. This includes detailing the system's general operating principles, expected behaviour and information on sensor data collection and usage. As discussed in Section 2.1.1, pre-existing knowledge significantly influences conceptual trust, often leading to a more positive outlook and a greater inclination to trust AV technology [111, 225]. Therefore, the initial level of transparency is critical in fostering a clear understanding of the technology and setting realistic expectations for safe interactions. User-centric studies typically adopt this approach by offering pre-interaction briefing, training and surveys, which can be interpreted as a representation of a low level of transparency [83, 91, 134, 186].

At higher levels, transparency requires the system to be responsive to user-initiated queries and provide situational explanations in real-time [253]. This transparency at the interaction level is the focus of our review, specifically in the context where individuals are directly impacted by the decisions made by AVs, as discussed in Section 4. At the interaction level, the layers of transparency

Table 3. Influence of Task and Context on Explanatory Information Concerning Internal and External Stakeholders

| Transparency | Internal Stakeholders | External Stakeholders |
|---|---|---|
| Layer 1: Goal | Driving state/mode<br>Driving action<br>Trajectory planning | Driving mode<br>Intent communication<br>Information + warning |
| Layer 2: Reasoning[a] | Driving action<br>Manoeuvre planning | *not addressed* |
| Layer 3: Projection | Uncertainty communication<br>System limitations | *not addressed* |

Adapted from [27]. [a]Although the literature suggests that agents typically provide explanatory information proactively at any level, there are also instances where reasoning-level explanations are given in response to users' queries.

the system requires vary depending on the task and context. A widely accepted protocol proposed by Chen et al. comprises three hierarchical layers of transparency [28]. Each layer outlines the necessary information that an agent must communicate to uphold a transparent interaction with the user.

- Layer 1: Goal—The first layer of transparency concerns the goals, actions and current status of the agent (e.g., 'AV stopping!'). Research on AV-user interaction, such as by [86, 96, 202], suggests that providing this basic level of explanatory information increases user comfort and situational awareness.
- Layer 2: Reasoning—The second layer of transparency contains the reasoning process, beliefs and motivations towards the current task (e.g., 'AV stopping because the traffic light is red!'). Studies, including [222, 262, 267], show that explaining the reasoning behind actions enhances user trust and safe mode transition in takeover situations.
- Layer 3: Projection—On the last layer of transparency, the agent is expected to predict future outcomes while managing uncertainty and potential limitations. Explanations might include statements such as 'AV stopping because the traffic light is red. It might take longer to take off due to a crowded pedestrian crossing'. Current research primarily focuses on explanatory information at this level in takeover scenarios, where the AV reaches the limits of its operational driving domain and alerts the driver to take control [29, 30].

The following research question investigates how explanatory information varies across different transparency layers, considering the diverse needs of internal and external stakeholders in various driving operations.

## 5.2 RQ4: How Do Task and Context Influence the Explanatory Information across Transparency Layers?

We reviewed literature across three layers of transparency, with each layer increasing in complexity in delivering explanatory information in relation to the explanatory task factors (i.e., stakeholder, driving condition, the function of explanation), as summarised in Table 3. Several driving conditions were identified for each transparency layer and stakeholder group. For example, in Layer 1 (Goal), the most basic explanatory information for both internal and external stakeholders involves communicating the current driving mode to prevent confusion in either environment. In the following, we delve into the explanatory information across transparency layers concerning different driving conditions for internal and external stakeholders.

*5.2.1 Layer 1 Transparency: Providing Current State, Goal and Action.* This layer of transparency entails the display of basic driving status cues to internal stakeholders, such as estimated arrival time, navigation and traffic conditions [70, 86]. In collaborative driving settings, the driver must be aware of the current mode and status to avoid confusion. During autonomous driving, the driver must be informed that the vehicle is following established driving norms and traffic regulations, such as indicating the current speed and speed limit. This layer of transparency needs no descriptive explanations; it only sends proactive, non-critical short messages to support situation awareness, e.g., 'Construction site!' [221]. Moreover, the driver can detect and understand the actions being performed by the vehicle through the visual arrows/carpet shown in HMI displays, indicating the direction and route without providing detailed reasoning [202, 268]. Concerning the planning task, several approaches have been suggested, such as bird's eye view displays for scene comprehension to identify the most influential objects in planning a trajectory [19, 23, 152, 211, 266].

Regarding external stakeholders, as per the Turing Red Flag law, AVs must be designed to clearly identify themselves as such and indicate their presence when interacting with external agents [253]. Research has explored the effect of **external HMIs (eHMIs)** to display the vehicle's driving mode [68, 109, 228]. Furthermore, literature has investigated the use of eHMIs in ambiguous crossing and unsignaled intersections for intent communication to facilitate safety crossing [39, 43, 139].

*5.2.2 Layer 2 Transparency: Providing Reasoning Behind AV's Actions and Decisions.* In this layer of transparency, it is essential for users to comprehend the reasoning behind a vehicle's action. This involves understanding what the AV is perceiving, analysing and making decisions. The user could trigger this function to produce an immediate explanation of why the system acted in a certain way in a given situation [78, 223]. Alternatively, the system itself could initiate a brief explanation of its ongoing activities [58, 78, 268]. Explanations generated for driving action could have different levels of specificity (abstract or specific) depending on the user's desire to influence the driving activity [186]. Furthermore, such explanations can facilitate interactive exploration of hypothetical situations, enabling users to inspect 'what if' scenarios.

Research has yet to delve into the communication of reasoning-level explanations and beyond with external stakeholders, particularly in the **vehicle-to-everything (V2X)** context. Such information is difficult to direct to individuals; it is, per the definition, broadcasted [53]. AV reasoning is often conveyed through its driving behaviour, which constitutes implicit communication. Further research is needed to identify scenarios necessitating explicit communication for AV reasoning. Additionally, there is a concern regarding handling communication with the outside in case eHMIs malfunction or display misleading information [96]. Given the limitations of LED-based eHMIs, alternative methods, such as personalised messages through smart and wearable devices, may be more appropriate for conveying this layer of transparency to other road users.

*5.2.3 Layer 3 Transparency: Conveying Uncertainty and Predictions of Failure.* In this layer of transparency, the driver must know the AV's capabilities and the conditions under which it operates. The driver should be informed of any unexpected outcomes, uncertainties or limitations when the AV approaches its operational boundaries due to road or environmental conditions, such as roads with curves, unclear markings or heavy traffic [30]. This can be achieved through a feedback mechanism that indicates an AV's proximity to its limits and control transition [174]. Studies have shown that informing drivers of the AV's uncertainty can better prepare them to switch to manual control when required [91]. Shull et al. implemented a system that offers pre-emptive feedback to the user when the AV fails to detect lane lines, thereby alerting them before it veers out of its designated lane [227]. Likewise, Kruger et al. and Kunze et al. conducted research on the impact of conveying reduced sensor reliability and machine certainty in diverse weather conditions, such as fog and rain [129, 135].

Table 4. Overview of the Communication Methods for Internal and External HMIs

| | | Communication Method | Presentation | Interactivity |
|---|---|---|---|---|
| Internal HMI | Display | Head-down displays (HDDs) | [30, 38, 121, 134] | [38, 134] |
| | | Head-up displays (HUDs) | [41, 43, 52, 86, 91, 132, 136, 174, 202, 221, 222, 251] | [52, 86, 136, 202, 221] |
| | | Head-mounted displays (HMDs) | [70] | [70] |
| | | Light band | [36, 41] | [36] |
| | | Vibrotactile feedback | [129, 133] | [129, 133] |
| | | Auditory voice message | [58, 186, 262, 267] | [58, 186, 262, 267] |
| | | Sonification | [29, 30] | [29, 30] |
| | Mult. | Audiovisual | [8, 64, 76, 216, 227, 244, 269] | [64, 216, 227, 244, 269] |
| | | LED and vibrotactile | [135, 268] | [135, 268] |
| | Features | Visuals/icons | [30, 52, 64, 91, 97, 132, 202, 216, 222, 269] | [52, 91, 136, 202, 222, 223] |
| | | Informative (text) | [69, 78, 81, 108, 119, 130, 177, 185, 188, 223] | [78, 81, 188] |
| | | Saliency maps | [1, 48, 120, 165, 218, 231, 264] | ø |
| | | Graphs/plots/trees | [17, 24, 25, 47, 144, 193, 220, 232, 266] | ø |
| | | Bird's eye view | [19, 23, 152, 166] | ø |
| External HMI | Display | LED bands | [34, 68, 83, 139, 271] | [34, 68, 83, 139, 271] |
| | | Forward-facing display | [40, 42, 96, 109, 161, 191, 228, 241] | [40, 109, 161, 228] |
| | | Optical projection | [37, 42, 182, 191] | [37, 42, 182, 191] |
| | Mult. | Display and projection | [39] | [39] |
| | | Audiovisual | [84] | [84] |
| | Features | Instructional | [35, 42, 109, 161, 191, 228] | [35, 42, 109, 161, 191, 228] |
| | | Symbolic | [9, 34, 37, 68, 83, 84, 96, 109, 139, 182, 228, 271] | [68, 83, 96, 139, 271] |
| | | Metaphorical | [37] | [37] |
| | | Anthropomorphic | [37, 241] | [37, 241] |

Research has shown that communicating potential uncertainties can improve trust calibration and situation awareness, leading to safer takeovers [134]. Issuing a warning in such circumstances is crucial, although the level of detail required may vary depending on the level of situational awareness required. In some cases, presenting visual information alone may be sufficient [40].

## 6 Explanatory Information Communication

HMIs facilitate communication between the AV, onboard occupants and surrounding road users. Bengler et al. and Murali et al. proposed a taxonomy for different types of HMIs in AVs [16, 176]. These HMIs are categorised as either 'internal' or 'external,' encompassing all interfaces within the vehicle's interior and exterior. We organised these interfaces and their components into communication methods, as outlined in Table 4. The table highlights studies that present explainability and explore interactive engagement through user evaluations. Articles focusing solely on technical-centric explainability without any presentation are omitted. Research involving interactivity mainly focuses on one-way interaction where communication flows from the AV to the user. More research is needed on bidirectional interactivity for both internal and external stakeholders, where users actively engage with the system, request additional information, or provide feedback to the system. In this section, we explore the different modalities and methods for delivering explanations to internal and external stakeholders, addressing RQ5 and RQ6.

### 6.1 RQ5: How Is the Explanatory Information Communicated to Internal Stakeholders?

Internal HMIs, which include all interfaces within the vehicle's interior, are designed to convey explanatory information to onboard users. They communicate the system's status, intentions and motion decisions in a transparent manner, promoting mode awareness and facilitating smooth transitions between driving modes for a safer, more comfortable journey. The communication of explanatory information is discussed in terms of display modalities, multimodality and features (i.e., the type of information conveyed).

### *6.1.1 Display Methods for Internal Users.*

Display modalities enable the communication of explanatory information. The internal HMI modalities are categorised into visualisations, audio and vibrotactile feedback. The proper modality of HMI must be activated by taking into account the context, audience and information content to effectively communicate relevant information.

In the realm of in-vehicle *visualisations*, feedback devices such as **head-down displays (HDDs)**, **head-up displays (HUDs)**, **head-mounted displays (HMDs)** and light bands/strips are commonly studied. While HDDs offer the benefit of not obstructing the users' view of the outside world, they can still be a source of distraction as users may become too engrossed in the display instead of the road ahead [175]. On the other hand, HUDs are designed to minimise driver distraction by presenting essential information relative to the positional and temporal environment within the driver's primary field of vision. Research suggests that HUDs offer better driving experiences over HDDs, resulting in decreased cognitive load, higher usability and better performance control [40, 134, 229]. AR-HUDs as **windshield displays (WSDs)** are the next step in developing HUDs by covering the entire windshield. These tools have demonstrated a remarkable capacity to enhance a driver's intuitive cognition and foster an efficient driving experience, particularly in challenging driving scenarios [40, 175]. However, to realize these benefits, HUD information must be thoughtfully designed in terms of content, timing and placement to ensure it complements rather than conflicts with real-world information.

HMDs offer the same advantages as HUDs. Optical HMDs allow users to access necessary information while remaining aware of their surroundings. Discrete HMDs provide a fully immersive **virtual reality (VR)** experience, which can replace the physical world with a virtual environment. This has the potential to reduce distractions and increase productivity for riders. Fereydooni et al. suggested incorporating short visual cues in the VR world during a ride, which was found to be helpful by passengers [70]. However, these devices are limited in the market and require further research to determine their suitability for in-vehicle applications. Given the technical difficulties and significant construction challenges in realising AR-HUDs and HMDs, current displays often come in the form of prototypes, concepts or demonstration videos. While AR-HUD proposes a more granular possibility of highlighting all relevant traffic objects, one of the concerns is that visual clutter may cause driver distraction and negatively impact driving performance. Kunze et al. argued that using the instrument cluster to visualise uncertainty information can increase mental workload [135]. Therefore, less distracting visualisation methods such as *light bands* (light strips) are used as a peripheral cue to bring attention to the objects of importance in the scene. Some works proposed lighting certain parts of the band depending on the position of the crossing pedestrian to indicate intention and perception [41, 89, 252].

*Auditory feedback* serves as another effective communication channel with internal stakeholders. Researchers have explored different strategies, such as speech, blended sonification and alerting sounds. In a study conducted by Avetisian et al., audio messages are used to direct users' attention to critical objects in the traffic to indicate their significance for the AV's decisions [7]. The study by Eimontaite et al. revealed that sound-based feedback improved the ease of operation and journey experience for elderly users when examining the impact of different modalities [64]. The research by Zhang et al. demonstrated the differences in the effectiveness of audio explanations across different age groups [267]. They discovered that voice-based explanations provided before action was taken resulted in the highest trust among elderly drivers. Additionally, blended sonification, which manipulates background music, was proposed to convey the reliability level to keep drivers informed. This strategy effectively increased monitoring behaviour and reduced response time to takeover requests, as suggested in studies by [29, 30].

*Vibrotactile feedback* is explored as an unobtrusive and intuitive means of communicating uncertainty to users. Kruger et al. conducted a study on the effects of using a vibrotactile belt to communicate sensor uncertainty information [129]. Their results showed that participants were able to understand the encoded information and perceived it as a meaningful communication medium. However, further investigation is needed to determine the usability of vibrotactile feedback for the ageing population, as it is known that age is associated with sensory decline. Kunze et al. suggested using vibration motors mounted in the vehicle seat to convey uncertainty [135]. While vibrotactile feedback alone can be helpful, it can also cause sudden spikes in attention, making it less effective for conveying dynamic information. Kunze et al.'s work suggests that combining vibrotactile feedback with other modalities may be more effective in communicating such information than relying solely on it [135].

*6.1.2 Multimodality for Internal Users.* Multiple feedback modalities enable the perception of a large amount of information without overwhelming a single sensory input. Multimodality is also necessary to convey warnings in case human intervention is needed [263]. This can increase the sense of urgency and lead to quicker response times [16]. However, it is essential to consider the form and content of multimodality depending on the context. These modalities should complement each other and aim to reduce information overload and clutter. Research has shown that a combination of visual and auditory resources is an effective feedback strategy to increase situational awareness [7, 37, 266]. Shull et al. studied how uncertainty feedback, representing current capability, influenced the transition of control to manual driving [227]. Their results showed that providing auditory-visual feedback led to longer HMI viewing times and early takeovers. In contrast, Zhang et al. found that auditory-visual explanations decreased decision-making performance [269]. Alternatively, visual-haptic feedback is introduced to raise the driver's attention when distracted by other tasks. As proposed in the study by Kunze et al., combining vibrotactile feedback with peripheral light can enhance the effectiveness of the feedback, as both can be adjusted based on the levels of uncertainty [135]. Nevertheless, further research is needed to assess the practicality of this approach.

*6.1.3 Display Features in Internal HMIs.* Explanations feature various formats, such as icons, text and bird's eye view to convey different types of information. Several visual presentation strategies were employed to support visibility and comprehension. Kunze et al. suggested using manipulable and abstract signs to communicate uncertainties by adjusting visual variables in size and transparency based on the level of ambiguity [132, 136]. Zhang et al. proposed using AR carpet visualisations to display lane availability, with arrows indicating lane-changing suggestions [268]. Monsaingeon et al. implemented an HDD interface that contains the instrument cluster, distance to the followed vehicle, detected road markings, system activation and a small area for textual messages [174]. Colley et al. presented semantic segmentation visualisations of the most relevant objects (e.g., vehicles, pedestrians) projected on the windshield as AR-HUD to promote user trust and situational awareness [43].

Research on the technical-centric perspective employs various presentation methods, including natural language, saliency maps, bounding boxes, feature importance plots, decision trees and bird's eye views. Natural language explanations provide detailed information about the vehicle's current actions [69, 130, 189]. Saliency maps and bounding boxes highlight areas influencing the system's decision-making [24, 48, 117, 120]. Graphs, plots and decision trees visually represent data and decision processes, helping users understand the rationale behind the system's actions [25, 47, 193]. The bird's eye view offers a top-down perspective, often illustrating the vehicle's surroundings and intended path [19, 23, 152]. However, despite their usefulness, these presentations are typically not integrated into HMIs and are rarely evaluated through human-subject studies.

### 6.2 RQ6: How Is the Explanatory Information Communicated to External Stakeholders?

External stakeholders receive explanatory information through external HMIs, which are installed on or projected from the vehicle's surface. While standardised eHMIs, such as indicators and brake lights, are already in use, new forms of eHMIs are being explored to enable seamless communication with other road users. The following section discusses various display modalities, multimodal approaches and the types of information conveyed by different modalities.

*6.2.1 Display Methods for External Users.* The external HMI modalities are categorised into light bands, forward-facing displays and optical projections. *Light band* exudes light patterns, such as running from one side to another or flashing to convey a message. The advantage of light band eHMI is its technical feasibility compared to projections or forward-facing displays. Furthermore, light bands are visible to multiple pedestrians and are not subject to the constraints of textual messages. Various lighting colours and patterns have been suggested to describe vehicle mode and intent [83]. One idea proposed by Lanzer et al. is to use a light strip that displays the vehicle's intent using different light patterns, such as yellow dots or flashed turquoise, to communicate whether the car is moving or intends to stop [139]. Faas et al. implemented a slowly flashing blue-green light above the windshield to signal yielding intent, which helped pedestrians calibrate their trust accordingly [67, 161]. Similarly, Zhanguzhinova et al. used red and green lights that switched between moving and stopping actions, resulting in smooth crossing behaviour and increased trust [271].

For *forward-facing displays*, various concepts have been proposed for positioning screens or panels on the vehicle's surface, including the bumper, grille, hood, windows and windshield. These display concepts have been studied to convey information for intended communication and increase road safety. Singer et al. proposed using symbols to communicate the vehicle's intentions to other road users, such as the vehicle's parking intention or direction of movement [228]. Research has also explored the use of forward-facing displays for presenting personal information when picking up passengers, indicating driving modes and signalling hazards to foster awareness and safety between the vehicle and external users [42, 109, 191]. Additionally, forward-facing displays have been studied for two-way communication, allowing pedestrians to respond to the vehicle's messages with gestures or waves [35].

*Optical projections* are an alternative means of conveying information by projecting messages—trajectories, stopping points, intentions and directions—onto the road surrounding the vehicle. This medium allows for the highest amount of information to be communicated, as the interface can be designed in detail [16, 53]. In an earlier study, Nguyen et al. proposed visualisations displayed on-road projections to indicate the intention of AV, such as stopping, slowing down, or proceeding [182]. Although this information helped participants adjust their actions, they demanded a display of timing during the crossing to know when the AV would start moving. Colley et al. compared the effectiveness of on-road projections versus information displayed on a car's surfaces, such as the bumper and windows [40]. The results showed that projecting the information on the sidewalk did not lead to increased trust. However, this concept was better at increasing situational awareness and warning distracted pedestrians about oncoming traffic. In another study by Colley et al., street projection was among the best-performing concepts as it was the only display communicating where the AV will stop [37]. Despite its potential benefits, on-road projections have certain limitations. They require near-perfect environmental conditions, including lighting, weather and road surface, to work properly. A powerful projector would also be necessary since the eHMI is projected over a significant distance.

*6.2.2 Multimodality for External Users.* In the realm of eHMIs, while the literature provides instances of combining two or more modalities, the exploration of multimodality has generally been less extensive compared to single-modality approaches [39, 57]. One line of research explored multimodality to address the inadequacy of current concepts for stakeholders with varying disabilities [84]. They compared auditory-visual modalities for individuals with and without **intellectual disability (ID)** to assess the inclusiveness of current eHMI concepts. The study involved emitting an audio message from the front of the vehicle for pedestrians and displaying a visual signal on the hood to indicate yielding intention. This implementation of multimodal eHMIs has shown positive results in terms of ID inclusivity. Regardless, it is essential to consider other factors, such as visual impairment, distracted road users and language barriers, as relying solely on visual displays may hinder safe interaction with various road users. There is no 'best' modality, as acceptability depends on various factors and is not universal, which calls for further research in mixed-modality eHMIs [219].

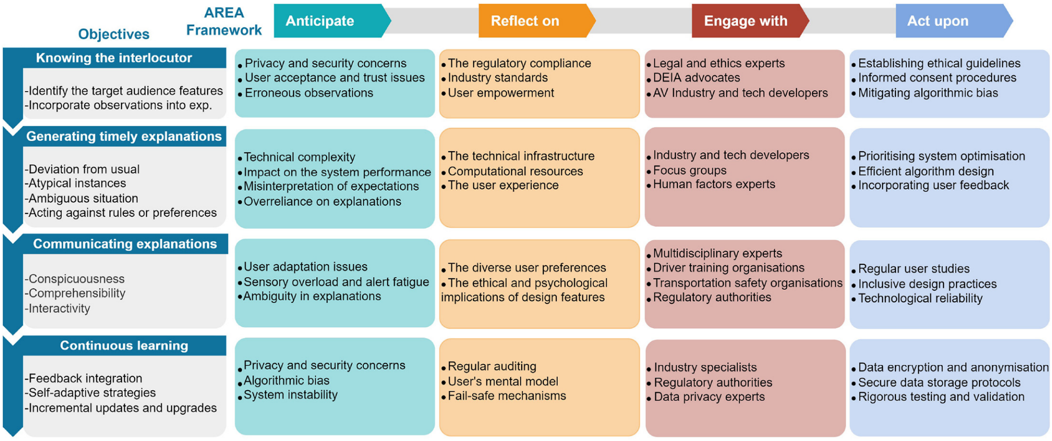
*6.2.3 Display Features in External HMIs.* External display signals information in four ways: instructional, symbolic, anthropomorphic and metaphorical [241]. Instructional signals are usually presented as text, providing advisory and informative messages. Advisory signals are not favoured for safety reasons (e.g., AV shows 'Cross', but another vehicle in the adjacent lane may not stop) [83]. Instead, intention-based informative messages (e.g., 'Stopping') are recommended by ISO [100]. Colley et al. considered instructional signals as the bidirectional communication between the AV and the pedestrian [35]. The study found that participants appreciate it when an AV responds to their hand gesture of 'Thank you!' with the feedback 'You are welcome!'. They described it as more human-like, friendly and clear. As useful as it is, there could be other limitations to textual signals, including language barrier, illegibility due to distance and vision disability.

The symbolic visual signals feature patterns (e.g., traffic symbols) that would animate over the LED display. Singer et al. explored the impact of animated symbols, like the P sign, arrow and flashing hand, on driving behaviour and intent recognition [228]. The results show that a combination of signals helps a better perception of the AV's intention and increases the perceived safety. Avetisyan et al. focused on eHMI for communication with human drivers in conventional vehicles in mixed traffic situations [9]. The visual messages displayed on the front of the AV warned drivers of uncertain situations, leading to increased awareness and trust.

The anthropomorphic signals comprised facial expressions such as animated eyes that would blink or gaze in different directions to indicate the vehicle's intention [214]. Colley et al. experimented with an expressive eHMI for AVs [37]. They utilised a bumper display shaped like a mouth, which would remain neutral with a horizontal line when the AV was not yielding and turn into a smile when the AV was yielding to pedestrians. Finally, metaphorical pictographs use animated allegorical narratives (e.g., 'walking man' at pedestrian crossings) to convey pedestrians to exercise caution, wait, or cross the street. Metaphorical pictographs and anthropomorphic signals seem most easily understood, especially as the characters explicitly display gestures and movements to indicate the future steps pedestrians could follow [241].

## 7 RRI Grounded Roadmap

As shown in our review (Sections 4–6), the explainability of AVs has been widely addressed in the research community. However, there is currently a lack of a systematic approach to the design of explanations, which may lead to suboptimal design of explanatory information or even contradictory outcomes. To address this gap, we suggest a roadmap that designers of explanatory



| Objectives / AREA Framework | Anticipate | Reflect on | Engage with | Act upon |
|---|---|---|---|---|
| **Knowing the interlocutor**<br>-Identify the target audience features<br>-Incorporate observations into exp. | • Privacy and security concerns<br>• User acceptance and trust issues<br>• Erroneous observations | • The regulatory compliance<br>• Industry standards<br>• User empowerment | • Legal and ethics experts<br>• DEIA advocates<br>• AV Industry and tech developers | • Establishing ethical guidelines<br>• Informed consent procedures<br>• Mitigating algorithmic bias |
| **Generating timely explanations**<br>-Deviation from usual<br>-Atypical instances<br>-Ambiguous situation<br>-Acting against rules or preferences | • Technical complexity<br>• Impact on the system performance<br>• Misinterpretation of expectations<br>• Overreliance on explanations | • The technical infrastructure<br>• Computational resources<br>• The user experience | • Industry and tech developers<br>• Focus groups<br>• Human factors experts | • Prioritising system optimisation<br>• Efficient algorithm design<br>• Incorporating user feedback |
| **Communicating explanations**<br>-Conspicuousness<br>-Comprehensibility<br>-Interactivity | • User adaptation issues<br>• Sensory overload and alert fatigue<br>• Ambiguity in explanations | • The diverse user preferences<br>• The ethical and psychological implications of design features | • Multidisciplinary experts<br>• Driver training organisations<br>• Transportation safety organisations<br>• Regulatory authorities | • Regular user studies<br>• Inclusive design practices<br>• Technological reliability |
| **Continuous learning**<br>-Feedback integration<br>-Self-adaptive strategies<br>-Incremental updates and upgrades | • Privacy and security concerns<br>• Algorithmic bias<br>• System instability | • Regular auditing<br>• User's mental model<br>• Fail-safe mechanisms | • Industry specialists<br>• Regulatory authorities<br>• Data privacy experts | • Data encryption and anonymisation<br>• Secure data storage protocols<br>• Rigorous testing and validation |

Fig. 3. RRI grounded roadmap for advancing explainable AV systems.

information in the context of AV could utilise to ensure that all key design elements are captured, as illustrated in Figure 3. The roadmap is grounded in RRI principles outlined in the AREA framework (see Section 3).

The proposed roadmap outlines four key objectives: knowing the interlocutor, generating timely explanations, communicating human-friendly explanations and continuous learning. The first three objectives correspond to each reach topic respectively, highlighting the essential components extracted from the explanatory task, explanatory information and explanatory information communications. In contrast to previous discussions (e.g., [6, 189, 256]), our roadmap also integrates the element of continuous learning and improvement. This objective encompasses the other three, emphasising the need for regular revisions and adaptations of explanation concepts as new methods emerge and user needs evolve. The development of each objective is informed by conclusions drawn from the literature survey and identifying gaps in each research topic. These lead to extracting subareas that are less explored but integral to advancing AV systems. Each objective is then further elaborated, incorporating AREA principles by considering the challenges highlighted in the literature and the broader context of AV development and design research.

### 7.1 Knowing the Interlocutor

An interlocutor in this context refers to the Class A users who are directly influenced by the actions taken by an AV (see Section 4.1.1). Previous research has discussed the integration of driver's physical and cognitive state information into in-vehicle interaction applications [103, 176]. However, these observations have not been incorporated into explanations, as noted in Section 4. This highlights the need for vehicle interaction technologies to identify the target interlocutor, i.e., the audience, whether the internal or external stakeholder and observe and recognise their physical and cognitive state to generate timely and meaningful explanations. As shown in Section 4.2, there are several scenarios where incorporating observations into explanations might be necessary, such as:

- The observations of stakeholders can be incorporated into explanations when the automation raises an intervention due to detecting deficient states of the driver. This approach can involve either reducing or increasing automation agency based on the level of autonomy the system has (e.g., AV takeover control due to detecting alcohol through transdermal sensors or alerting a drowsy driver for situation awareness).

- Observation can help offer proactive explanations in situations where there is a risk that the passenger(s) might feel discomfort, confusion, or frustration by a driving-related situation, especially when no visible reason exists in the scene.
- Observations can help to justify an action that resulted from detecting an unusual situation (e.g., AV pulls over and contacts an operator due to detecting the passenger's medical emergency).
- Understanding external stakeholders' attributes is essential to provide explainable trajectory planning for onboard users, especially when other road users' intentions are not clear but could influence the behaviour of the AV.
- Detecting relevant characteristics of safety-critical road users (e.g., child, elderly, pet/animal, distracted pedestrians, e-scooter user), if there exists any impairment (e.g., cognitive, vision, hearing, mobility) is necessary to properly calibrate the external communication mode and efficiently interact with them.

While an explainable AV with these capabilities can significantly improve user experience and safety, it also poses certain risks and implications related to the effectiveness and ethical use of the data. The literature evaluated several risks and implications concerning such technology, which relates to the explanatory task as follows:

*Data privacy and security concerns*—Incorporating stakeholder observations for explanations raises privacy concerns, especially when sensitive information about drivers or passengers is involved. Proper data handling and anonymisation protocols must be in place, and clear guidelines should be established for collecting, storing and processing stakeholder observations to protect the privacy and confidentiality of the individuals involved [128, 176].

*User acceptance and trust*—Users may have concerns about the collection and use of personal data [103, 176]. Therefore, it is essential to empower users to have control over their data and the ability to select which part of observations to opt in or out of the system. It is important to make sure that users are informed about the benefits and implications of incorporating observations in enhancing the AV system's operations.

*Erroneous observations*—There is a risk of misinterpreting stakeholder behaviour, leading to incorrect or inappropriate responses. To minimise this risk, it is necessary to have robust AI models capable of accurately understanding and responding to complex human behaviours and intentions [162]. Implementing robust sensor technologies and data fusion techniques can help ensure the accuracy and reliability of stakeholder observations, reducing the risk of false positives.

*Legal and regulatory compliance*—Incorporating stakeholder observations for decision-making raises questions about legal liability and compliance with existing regulations. The legal implications of AV interventions and data usage must be thoroughly evaluated and addressed to ensure adherence to relevant laws and regulations [16]. Additionally, clear ethical guidelines and informed consent procedures should be established for the collection and use of stakeholder observations in AV decision-making processes.

*Algorithmic bias and fairness*— Integrating stakeholder observations may reduce some biases while potentially introducing others into the AV's decision-making processes, especially if the observations are not sufficiently representative or if the algorithms are not properly calibrated. Implementing bias detection mechanisms, algorithmic fairness assessments and regular audits can help identify and mitigate potential biases, ensuring fair decision-making by the AV system [80, 128].

Addressing these risks and implications requires a collaborative and inclusive approach that combines robust technical solutions, ethical guidelines and legal compliance. Researchers have been engaging with various relevant parties to enable the responsible and ethical deployment of AV technology through industrial collaborations and workshops [78, 157, 178]. We identified some key interest groups to involve in the process, including:

*Legal authorities and ethics committees*—Collaborating with legal experts can help establish ethical guidelines, informed consent frameworks and regulatory compliance measures for integrating stakeholder observations in AV decision-making processes.

*Diversity, Equity, Inclusion and Accessibility advocates*—Involving accessibility, diversity and inclusion advocates can help address the diverse needs of users. Discussions around the importance of fair representation, assessments and bias detection mechanisms will ensure that the AV system's decision-making processes are inclusive and fair [199].

*AV industry and technology developers*—Collaborating with the AV industry and technology developers can provide insights into best practices, such as sensor technologies and data fusion techniques, for integrating stakeholder observations in AV systems. This knowledge exchange would also foster consensus on technical specifications and protocols, leading to standardised frameworks.

*Urban planners and representatives of other road users*—Engaging urban planners and representatives from various road user groups (e.g., pedestrians, cyclists) is crucial for understanding the broader context of AV deployment. Their input helps address urban infrastructure needs, safety concerns and the integration of AVs into existing transportation networks.

## 7.2 Generating Timely Explanations

Providing explanations for every action, even in routine situations, may result in information overload for users, ultimately diminishing the value of explanations and reducing their effectiveness [65]. Prioritising event criticality for explanations while ensuring a balance between transparency and simplicity is essential to avoid overwhelming users with unnecessary details. While this has been partially addressed in the literature, further research is needed to focus on timing for proactive explanations (see Sections 4.2.1 and 4.2.2). Timely explanations should address discrepancies between what the user expects the system to do and what it does, thereby promoting a seamless experience. Motivated by the work of Gervasia et al. [75], we suggest a set of guidelines for generating timely explanations that address expectation violation as follows:

*Deviation from usual*—AV should identify if an action differs from past behaviour in similar situations through statistical analysis of performance logs and semantic models. The explanation should acknowledge the atypical action and explain the reason when it exhibits unusual behaviour (e.g., AV not yielding to pedestrians).

*Atypical instances*—AV should generate explanations for actions taken in unusual situations causing atypical action. The AV can identify these situations by their frequency of occurrence and describe the unusual situation to the user (e.g., sharp brake due to an unidentified object—a deer jumping on the road).

*Ambiguous situation*—In some situations, the apparent cause of the action may be unavailable to the user, and the user might assume that the information used for the decision comes only from the vehicle's vision sensors. AV should explain the potential mismatch and indicate decision criteria in such situations, e.g., in a ride-sharing scenario, AV divert to an unusual route, justifying that it will take a new rider on the way.

*Acting against rules or preferences*—In typical driving conditions, AVs seek to satisfy the established traffic rules, regulations and preferences; nevertheless, various factors (e.g., to prevent a potentially more severe incident, time frame, physical restrictions) may cause a violation, leading to the agent operating contrary to its set-up. Explanation, in this case, involves acknowledging the violated directive or preference and providing why that occurred (e.g., changing the preferred driving style or stopping at the ‘No stopping’ sign).

Implementing an explainable AV system with the outlined capabilities can significantly improve user trust and experience. However, it is crucial to consider and act on several key aspects to ensure the effectiveness of the explanations. One such aspect is technical complexity and system performance. Generating context-sensitive explanations requires a robust technical infrastructure and algorithms without impacting the overall system performance and responsiveness [261]. Therefore, it is crucial to prioritise system optimisation and efficient algorithm design to ensure that the explanation generation process does not compromise the overall performance and computational resources of the AV system [5, 243]. Another aspect to consider is the risk of misinterpreting user expectations, leading to misaligned explanations that fail to address the user’s concerns adequately. To mitigate this risk, incorporating robust user feedback mechanisms and conducting regular user studies can help align the AV system with user expectations more effectively [223]. Additionally, there is a risk of users becoming overly dependent on the explanations, which may lead to a lack of critical thinking and reduced user engagement with the driving task [262].

Academics and industrial partners have been working together to address these concerns and fuse their expertise to design AV systems that are both robust and reliable [183, 244]. In addition, collaborating with human factor experts in the field can offer valuable insight into users’ cognitive processes, behaviours and preferences [53]. Such collaborations can inform the explanation generation process and user-centred design principles to ensure the successful deployment of explainable AV technology with these capabilities.

### 7.3 Communicating Human-Friendly Explanations

In the literature, there is a consensus that explanations should convey relevant information in a clear, concise and friendly manner, revealing the reasoning behind a decision, the current state of affairs and the potential outcomes in the future [22, 129, 185]. For external stakeholders, communication should compensate for the absence of human-driver interaction. While discussions on these aspects exist, the lack of a comprehensive set of guidelines remains a gap. Our approach introduces three key properties for communicating human-friendly explanations in the context of AVs: conspicuousness (easy to perceive), comprehensibility (easy to understand) and interactivity (easy to engage with). We proposed a thorough approach with a detailed map outlining how to take on each one properly.

*7.3.1 Conspicuousness.* This refers to the degree to which the resulting explanation is clear and easy to notice through the visual, auditory, or tactile channel. The communication channel should be adjusted based on the degree of priority of the message.

*Low-priority messages*—These refer to the messages conveyed in the first layer of transparency that describe the current state of the system when it is running smoothly (see Section 5.2.1). Low-priority messages vary across levels of autonomy, with SAE Levels 2 and 3 focusing on driver awareness and readiness, while higher levels provide more general information that does not require immediate attention from the driver. To avoid disrupting users while they perform NDRTs, one could consider presenting low-priority messages as visual feedback incorporated into AR WSDs [212]. Visual feedback may include information about other vehicles in close proximity to the ego vehicle, such as their orientation, size, movement and intention [254]. This may also include information about detecting partially obstructed objects in the environment, as well as large objects that may block the view of potential hazards [232]. As discussed in

Section 4.2.1, this information should be readily available, especially during the introductory phase, to enhance the perception of safety. Concerning external road users, the low-priority messages involve the display of driving mode when the vehicle is operated in autonomous driving (see Section 5.2.1). Previously, marker lamps have been recommended by SAE [101]. Most recently, Mercedes-Benz became the first automaker to receive permits allowing the use of turquoise-coloured lights on the outside of the vehicle to indicate autonomous driving mode [236].

*High-priority messages*—Conspicuousness is essential for messages of higher importance. It is advisable to avoid displaying such information visually, as it may result in drivers overlooking significant changes [142]. This is because drivers frequently switch focus between the road, displays and NDRTs [134]. High-priority messages might be presented through multimodal interfaces to capture the user's attention as uncertainties escalate, e.g., when a takeover becomes more probable [181, 269]. As shown in Section 6.1.2, audible verbal messages accompanied by visual cues have been proven to be more effective than relying solely on visual signals when communicating takeover requests [40]. However, they may not be sufficient in situations where the driver is listening to music, conversing with passengers, or dealing with loud noises [171]. As discussed in Section 6.1.2, auditory and vibrotactile alerts could be utilised to adapt to the urgency of the message. Regardless of the modality, the warning messages should direct the user's attention towards the source of danger and persuade them to take action promptly. In regard to external stakeholders, high-priority messages involve situations ranging from intent communication to warning signals [67, 147, 241]. Such explanations should take into account the diverse road user requirements, such as those with impairments, language barriers, or those who are distracted.

*7.3.2 Comprehensibility.* Comprehensibility is the capacity of an explanatory agent to represent its knowledge in a human-understandable way [94]. For the message to be comprehensible, it must be easily understood with minimal cognitive effort. This property highly depends on the audience, and the context since comprehensibility is a subjective concept. We have gathered four factors that support comprehensibility:

*Clarity*—refers to the degree to which the resulting explanation is explicit. This property is particularly relevant for high-priority messages. The interface should adopt commonly used icons and filter out irrelevant traffic features to mitigate the risk of ambiguity [156]. Additionally, explanations should refrain from utilising highly abstracted information for situation awareness [40]. As highlighted in Section 5.2.2, AV needs to communicate its intentions explicitly to external stakeholders, leaving no room for interpretation. Moreover, it is crucial to avoid confusing or misleading those who are not the intended recipients. For instance, a passenger does not need to understand the information that is intended for the driver. Similarly, if the message targets pedestrians, it should not perplex other drivers who may happen to see it.

*Selection and refinement*—pertain to the capacity of an explanatory agent to focus solely on the critical causes that are sufficient to explain the situation. Typically, humans do not expect an explanation to contain the complete list of causes of a decision but rather seek an explanation that conveys the most critical information supporting the decision [172]. When presenting driving-related information, displays should offer minimal features necessary to justify the situation [181]. As pointed out in Section 6.2.1, external communication should also consider non-visual, more inclusive eHMI concepts for the safety of road users who do not have access to visuals [84, 271].

*Informativeness*—relates to the ability of an explanatory agent to provide relevant information to the user without causing cognitive workload [268]. As per the discussion in Section 6.1.2, when there is an unusual amount of information to convey, multimodality could be a solution to not overwhelming users by loading one sensory modality.

*Simplicity/Parsimony*—refers to the complexity of the resulting explanation. A parsimonious explanation is a simple explanation. The optimal degree of parsimony may vary depending on the user. For example, novice drivers may benefit from more detailed semantic information, while experienced or elderly drivers may find a simpler message more helpful [156]. Therefore, the explanation system should be adaptable to accommodate the specific needs and preferences of both the driver and the occupants.

*7.3.3 Interactivity.* Despite its significance, bidirectional interactivity remains a relatively less explored feature within the realm of explanation communication involving HMIs, as revealed in Table 4. The practical methodologies in two-way interactivity are mainly limited to technical-centric works that explore visual question-answering [4, 167, 203]. However, bidirectional interactivity involves gaining insight into the target audience, including their goals, needs and preferences [27]. By knowing the users, the interactive experience can be tailored to suit their context, behaviour and personality. We have identified three elements to consider for an interactive explanatory agent: engagement, empathy and anthropomorphism.

*Engaging*—Designing for engagement means making it responsive and adaptive to the users' inputs, actions and contexts. This involves an explanation method to reason about prior interactions to interpret and respond to users' follow-up questions. This is important for creating a rapport with the users, which is also beneficial for the initial trust-building process. Then, the users should be able to control the level of engagement they prefer [43]. Regarding other road users, communication should provide a form of acknowledgement, showing the vehicle is aware of them, as described in Section 6.2.1. This could include a mechanism for bidirectional communication by responding to gestures [40]. In a V2X context, messages could be sent to HMIs of other agents to enable interactive communication.

*Empathic*—Empathic design involves designing for emotion, making explanations expressive and personality-driven. This involves action, emotion and gesture recognition and utilising physiological inputs (e.g., thermal, olfactory, gustatory, cerebral, or cardiac signals) to synthesise data, adapt suitable modalities and refine interactions [103]. The explanatory agent should identify additional relevant user characteristics (e.g., age, gender, cultural background) to address empathic interface requirements. This could also be beneficial to the conceptual trust-building process (See Section 2.1.1).

*Anthropomorphic*—By leveraging anthropomorphic features such as attractiveness, personalisation, ethnicity and facial similarity, explanatory agents can create more engaging and user-friendly experiences. Implementing elements such as voice, tone, language, humour and emotion recognition can contribute to a human-like relatable experience [80, 216, 262]. According to the analysis in Section 6.2.3, anthropomorphic signals such as facial expressions or pictographs could be adapted to communicate the vehicle's intention for external stakeholders [37, 241].

Implementing an explainable AV that can adjust conspicuousness based on message priority requires thorough consideration during the design and deployment phases. Moreover, the emphasis on comprehensibility and interactivity adds further complexity to the implementation process. Therefore, there are several important risks and implications associated with these tasks that should be kept in mind, as follows:

*User adaptation and familiarity*—Adjusting the communication channel's conspicuousness may require users to adapt to new modes of information processing and response [16]. Ensuring that users are adequately trained and familiar with the AV's communication interfaces and protocols is crucial to facilitate quick and appropriate responses to messages of varying priorities [58]. The system could integrate training modules to guide users through the communication interfaces and protocols. Supplementary training resources, such as user manuals, interactive tutorials and simulation exercises, can be provided to ensure comprehensive understanding and readiness.

*Sensory overwhelm and alert fatigue*—Using multimodal communication in AVs can lead to sensory overwhelm and alert fatigue [168]. Overexposure to alerts may result in habituation, causing desensitisation to urgent alerts and delayed reactions in critical takeover scenarios [73]. Drawing from Section 4.2.1, continuous exposure to messages may overwhelm the user or even foster overreliance on the notifications, leading to complacency and reduced engagement with the driving environment. Striking a balance between the urgency of the message, sensory tolerance and the user's cognitive load is crucial to prevent undue stress without compromising the user's ability to maintain situational awareness.

*Ambiguity in explanations*—Explanations must align with the AV's driving behaviour. Inconsistencies between the provided explanations and the AV's actions can result in confusion—relates to both internal and external communication. This may lead users to draw wrong conclusions or take incorrect actions. To avoid confusion and misunderstanding of the system's status, various visual feedback strategies and symbols have been proposed [53, 181]. Nonetheless, further research is required to explore the use of multimodal communication, especially in high-pressure or noisy driving conditions, to determine if it contributes to ambiguity.

*User diversity*—As discussed in Section 2.1.1, there are differences in how people of different ages, genders and cultures perceive and trust AV technology [63, 155, 206, 274]. Further research requires consideration of these factors, along with digital literacy, cognitive abilities, and so forth, to ensure the information is easily understandable and identifiable by the intended stakeholders—to whom it was addressed.

*Technological reliability and redundancy*—Depending on the communication channel's reliability, there may be a risk of technical failures or malfunctions, leading to a lack of timely and effective message delivery [67, 96]. Implementing robust redundancy mechanisms and regular system checks are essential to minimise the risk of communication failures and ensure that critical messages reach the user promptly and reliably.

*Empathic and anthropomorphic design challenges*—Incorporating empathic design elements that rely on accurate emotion and gesture recognition presents technical challenges, emphasising the need for reliable systems to prevent misinterpretation of user emotions [145, 159]. Additionally, implementing anthropomorphic features to improve the user experience can raise ethical and psychological concerns related to humanising the AV system. This necessitates a balance between creating relatable experiences and avoiding potential user discomfort or confusion [248].

The literature reviewed in Sections 5 and 6 shows that most user-centric approaches have actively engaged with human-subject studies to understand potential users' expectations and concerns regarding the explanation generation process [9, 30, 37, 186, 269]. Moreover, engaging with community representatives and advocacy groups can help understand local communities' specific needs and preferences. Their guidance can inform the implementation of inclusive design practices and accessible communication modalities for a broad spectrum of users. However, there is a need to balance localisation and standardisation. While localisation addresses specific community needs, standardisation ensures consistency and avoids confusion when users move between different vehicles or regulatory environments. Ideally, a combination of both approaches should be pursued, where universal standards are implemented to promote clarity and safety while allowing for some degree of customisation to address local and individual requirements. In addition, collaboration with multidisciplinary experts, including human factors specialists, cognitive psychologists, UX designers and HCI specialists, is essential in optimising AV interface design and developing effective communication systems [53, 61, 222].

Other significant bodies to engage with include driver training and transportation safety organisations. They can facilitate the development of effective training programs and guidelines for users to familiarise themselves with the AV's communication interfaces and response protocols. Consultation with transportation safety regulators and authorities is necessary to ensure that the communication features align with existing safety standards and guidelines. Other regulatory authorities and standards organisations also play a significant role in ensuring compliance with industry regulations and safety standards for AV communication systems.

### 7.4 Continuous Learning and Improvement

The review provided in Sections 4–6 lacked discussions on how existing explanation concepts address the ongoing evolution of AV systems alongside the changing user needs. A continuous learning mechanism is necessary to ensure the system's explanation generation improves over time. User feedback integration, self-adaptive strategies and incremental updates are some essential aspects to take into account for continuous learning and improvement.

*Feedback integration*—The continuous learning process involves integrating user feedback into the learning process. By understanding user responses, concerns and suggestions, the system can adjust its explanation strategies to meet users' evolving needs and preferences and enhance the system's overall performance [13, 78]. Recent advancements in integrating large language models into AVs set a paradigm shift, demonstrating the potential for continuous learning and personalised engagement [45, 46, 237]. Thereby, users can enjoy a more seamless and intuitive interaction with AVs.

*Self-adaptive strategies*—As discussed in Section 2.1.1, sociocultural context conditions human cognition and perception, and thus, the explanation needs of users may differ across cultures [124]. As AVs traverse cultural boundaries, they may need to adjust their behaviour to align with the norms and expectations of each culture encountered [170, 201]. With continuous learning, the system would adapt strategies in response to changing environments by monitoring shifts in user behaviour and updating knowledge with fresh data points in real-time.

*Incremental updates and upgrades*—Continuous learning also involves the integration of incremental updates to the system's architecture, algorithms, explanation techniques and data processing [143]. By doing so, the system remains up to date with the latest advancements in the field and can leverage SotA techniques to improve its performance and capabilities. However, these enhancements may not be available immediately, as manufacturers typically release updates and upgrades at set intervals, ranging from several months to a year. A key challenge is also the lifespan of vehicles—often up to 20 years—requiring manufacturers to provide ongoing support for updates over extended periods. Ensuring compatibility will be crucial to maintaining performance throughout the vehicle's lifecycle.

Integrating continuous learning mechanisms in an explainable AV system has several benefits but also poses implications related to data processing and model performance. Collecting and analysing user feedback data may raise concerns about privacy and security. Ensuring robust data encryption, anonymisation techniques and secure data storage protocols are essential to protect sensitive user information from unauthorised access and misuse [234]. Related to this, regular auditing of the learning models and data sources is important to identify and mitigate any algorithmic bias resulting from continuous learning [12, 114, 151]. On another note, continuous learning may impact trust if users perceive the system's decisions as unpredictable or inconsistent. Providing clear and transparent explanations for the system's continuous learning processes and decision-making criteria can help users build a better mental model of the system [60, 74]. Concerning the model performance, continuous learning may introduce system instability, leading to unexpected behaviours or unintended consequences that could compromise user safety and trust. Implementing rigorous testing and validation procedures, along with effective fail-safe mechanisms, is crucial to ensure the stability and reliability of the AV system throughout the continuous learning process [123]. Finally, it is worth noting that engaging with relevant stakeholders, including data privacy and security experts, regulatory authorities and industry specialists, is important to mitigate the potential risks and implications.

## 8 Conclusions

In this study, we conducted a literature review to assess the SotA research in explainable autonomous driving. The study addressed several research questions related to the explanatory task, information and means of communication. Our analysis revealed that several factors, including stakeholders, driving operations and level of autonomy, influence the explanatory task. We discussed situations requiring explanations by contrasting proactive and reactive explanations based on their criticality—either situations that pose an immediate danger or harmless but questionable driving behaviours. Regarding explanatory information, we identified three layers of transparency: goal, reasoning and projection, elucidating them further based on the level of autonomy and stakeholder needs. Finally, we evaluated how the explanatory information is conveyed to internal and external stakeholders. We highlighted three critical design considerations: conspicuousness, comprehensibility and interactivity. While current approaches facilitate one-way interaction, our analysis underscores the need for further research into bidirectional interactivity, allowing users to actively engage with the system and provide feedback. Considering the lack of a systematic design approach, we suggested a roadmap to address all key design elements. By considering RRI principles, our roadmap emphasises the importance of understanding diverse requirements for explanations. It pinpoints critical design elements related to explanatory tasks, information and communication, as identified in the review: understanding the audience, generating timely explanations and communicating human-friendly explanations. Notably, our roadmap integrates continuous learning aspects, a dimension often overlooked in existing literature. The roadmap concerns both the internal and the external stakeholders; therefore, the suggestions can be adapted and applied to the research relating to both areas.

## Acknowledgement

We would like to thank Daniel Omeiza and the anonymous reviewers for their valuable feedback on the manuscript.

## References

[1] Mohanad Abukmeil, Angelo Genovese, Vincenzo Piuri, Francesco Rundo, and Fabio Scotti. 2021. Towards explainable semantic segmentation for autonomous driving systems by multi-scale variational attention. In *Proceedings of the IEEE International Conference on Autonomous Systems (ICAS '21)*. IEEE, 1–5.

[2] Stefano V. Albrecht, Cillian Brewitt, John Wilhelm, Balint Gyevnar, Francisco Eiras, Mihai Dobre, and Subramanian Ramamoorthy. 2021. Interpretable goal-based prediction and planning for autonomous driving. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '21)*. IEEE, 1043–1049.

[3] Imran Ali. 2019. Personality traits, individual innovativeness and satisfaction with life. *Journal of Innovation & Knowledge* 4, 1 (2019), 38–46.

[4] Shahin Atakishiyev, Mohammad Salameh, Housam Babiker, and Randy Goebel. 2023. Explaining autonomous driving actions with visual question answering. In *Proceedings of the IEEE 26th International Conference on Intelligent Transportation Systems (ITSC '23).* IEEE, 1207–1214.

[5] Shahin Atakishiyev, Mohammad Salameh, and Randy Goebel. 2024. Safety implications of explainable artificial intelligence in end-to-end autonomous driving. arXiv:2403.12176. Retrieved from https://arxiv.org/abs/2403.12176

[6] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. 2021. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. arXiv:2112.11561. Retrieved from https://arxiv.org/abs/2112.11561

[7] Lilit Avetisian, Jackie Ayoub, and Feng Zhou. 2022. Anticipated emotions associated with trust in autonomous vehicles. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 66. SAGE Publications, Los Angeles, CA, 199–203.

[8] Lilit Avetisyan, Jackie Ayoub, and Feng Zhou. 2022. Investigating explanations in conditional and highly automated driving: The effects of situation awareness and modality. *Transportation Research Part F: Traffic Psychology and Behaviour* 89 (2022), 456–466.

[9] Lilit Avetisyan, Aditya Deshmukh, X. Jessie Yang, and Feng Zhou. 2023. Investigating HMIs to foster communications between conventional vehicles and autonomous vehicles in intersections. arXiv:2305.17769. Retrieved from https://arxiv.org/abs/2305.17769

[10] Jackie Ayoub, Lilit Avetisyan, Mustapha Makki, and Feng Zhou. 2021. An investigation of drivers' dynamic situational trust in conditionally automated driving. *IEEE Transactions on Human-Machine Systems* 52, 3 (2021), 501–511.

[11] Hebert Azevedo-Sa, Huajing Zhao, Connor Esterwood, X. Jessie Yang, Dawn M. Tilbury, and Lionel P. Robert Jr. 2021. How internal and external risks affect the relationships between trust and driver behavior in automated driving systems. *Transportation Research Part C: Emerging Technologies* 123 (2021), 102973.

[12] Andrew Bae and Susu Xu. 2022. Discovering and understanding algorithmic biases in autonomous pedestrian trajectory predictions. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 1155–1161.

[13] Matthew Barker, Emma Kallina, Dhananjay Ashok, Katherine M. Collins, Ashley Casovan, Adrian Weller, Ameet Talwalkar, Valerie Chen, and Umang Bhatt. 2023. FeedbackLogs: Recording and incorporating stakeholder feedback into machine learning pipelines. arXiv:2307.15475. Retrieved from https://arxiv.org/abs/2307.15475

[14] Rolando Bautista-Montesano, Rogelio Bustamante-Bello, and Ricardo A. Ramirez-Mendoza. 2020. Explainable navigation system using fuzzy reinforcement learning. *International Journal on Interactive Design and Manufacturing (IJIDeM)* 14, 4 (2020), 1411–1428.

[15] Hédi Ben-Younes, Éloi Zablocki, Patrick Pérez, and Matthieu Cord. 2022. Driving behavior explanation with multi-level fusion. *Pattern Recognition* 123 (2022), 108421.

[16] Klaus Bengler, Michael Rettenmaier, Nicole Fritz, and Alexander Feierle. 2020. From HMI to HMIs: Towards an HMI framework for automated driving. *Information* 11, 2 (2020), 61.

[17] Frédéric Bouchard, Sean Sedwards, and Krzysztof Czarnecki. 2022. A rule-based behaviour planner for autonomous driving. In *Proceedings of International Joint Conference on Rules and Reasoning*. Springer, 263–279.

[18] Cillian Brewitt, Stefano V. Albrecht, John Wilhelm, Balint Gyevnar, Francisco Eiras, Mihai Dobre, and Ram Ramamoorthy. 2020. Autonomous driving with interpretable goal recognition and Monte Carlo tree search. In *Interaction and Decision-Making in Autonomous-Driving: A Virtual Workshop at RSS 2020*.

[19] Cillian Brewitt, Balint Gyevnar, Samuel Garcin, and Stefano V. Albrecht. 2021. GRIT: Fast, interpretable, and verifiable goal recognition with learned decision trees for autonomous driving. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '21).* IEEE, 1023–1030.

[20] Thibault Buhet, Emilie Wirbel, Andrei Bursuc, and Xavier Perrotton. 2021. Plop: Probabilistic polynomial objects trajectory prediction for autonomous driving. In *Proceedings of the Conference on Robot Learning*. PMLR, 329–338.

[21] Mirjam Burget, Emanuele Bardone, and Margus Pedaste. 2017. Definitions and conceptual dimensions of responsible research and innovation: A literature review. *Science and Engineering Ethics* 23 (2017), 1–19.

[22] Eduardo Candela, Olivier Doustaly, Leandro Parada, Felix Feng, Yiannis Demiris, and Panagiotis Angeloudis. 2023. Risk-aware controller for autonomous vehicles using model-based collision prediction and reinforcement learning. *Artificial Intelligence* 320 (2023), 103923.

[23] Sandra Carrasco, D. Fernández Llorca, and M. A. Sotelo. 2021. Scout: Socially-consistent and understandable graph attention network for trajectory prediction of vehicles and vrus. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV '21).* IEEE, 1501–1508.

[24] Zenon Chaczko, Marek Kulbacki, Grzegorz Gudzbeler, Mohammad Alsawwaf, Ilya Thai-Chyzhykau, and Peter Wajs-Chaczko. 2020. Exploration of explainable AI in context of human-machine interface for the assistive driving system. In *Proceedings of Asian Conference on Intelligent Information and Database Systems*. Springer, 507–516.

[25] Anas Charroud, Karim El Moutaouakil, Vasile Palade, and Ali Yahyaouy. 2023. XDLL: Explained deep learning LiDAR-based localization and mapping method for self-driving vehicles. *Electronics* 12, 3 (2023), 567.

[26] Dong Chen, Longsheng Jiang, Yue Wang, and Zhaojian Li. 2020. Autonomous driving using safe reinforcement learning by incorporating a regret-based human lane-changing decision model. In *Proceedings of the American Control Conference (ACC '20)*. IEEE, 4355–4361.

[27] Jessie Y. C. Chen, Shan G. Lakhmani, Kimberly Stowers, Anthony R. Selkowitz, Julia L. Wright, and Michael Barnes. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science* 19, 3 (2018), 259–282.

[28] Jessie Y. Chen, Katelyn Procci, Michael Boyce, Julia Wright, Andre Garcia, and Michael Barnes. 2014. Situation awareness-based agent transparency. *US Army Research Laboratory* (April 2014), 1–29.

[29] Kuan-Ting Chen and Huei-Yen Winnie Chen. 2021. Manipulating music to communicate automation reliability in conditionally automated driving: A driving simulator study. *International Journal of Human-Computer Studies* 145 (2021), 102518.

[30] Kuan-Ting Chen, Huei-Yen Winnie Chen, and Ann Bisantz. 2023. Adding visual contextual information to continuous sonification feedback about low-reliability situations in conditionally automated driving: A driving simulator study. *Transportation Research Part F: Traffic Psychology and Behaviour* 94 (2023), 25–41.

[31] Sikai Chen, Jiqian Dong, Runjia Du, Yujie Li, and Samuel Labi. 2021. Reason induced visual attention for explainable autonomous driving. arXiv:2110.07380. Retrieved from https://arxiv.org/abs/2110.07380.

[32] Yilun Chen, Chiyu Dong, Praveen Palanisamy, Priyantha Mudalige, Katharina Muelling, and John M. Dolan. 2019. Attention-based hierarchical deep reinforcement learning for lane change behaviors in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

[33] Jong Kyu Choi and Yong Gu Ji. 2015. Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction* 31, 10 (2015), 692–702.

[34] Mark Colley, Elvedin Bajrovic, and Enrico Rukzio. 2022. Effects of pedestrian behavior, time pressure, and repeated exposure on crossing decisions in front of automated vehicles equipped with external communication. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–11.

[35] Mark Colley, Jan Henry Belz, and Enrico Rukzio. 2021. Investigating the effects of feedback communication of autonomous vehicles. In *Proceedings of the 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications,* 263–273.

[36] Mark Colley, Julian Britten, Simon Demharter, Tolga Hisir, and Enrico Rukzio. 2022. Feedback strategies for crowded intersections in automated traffic—A desirable future? In *Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 243–252.

[37] Mark Colley, Julian Britten, and Enrico Rukzio. 2023. Scalability in external communication of automated vehicles: Evaluation and recommendations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 2 (2023), 1–26.

[38] Mark Colley, Benjamin Eder, Jan Ole Rixen, and Enrico Rukzio. 2021. Effects of semantic segmentation visualization on trust, situation awareness, and cognitive load in highly automated vehicles. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–11.

[39] Mark Colley, Tim Fabian, and Enrico Rukzio. 2022. Investigating the effects of external communication and automation behavior on manual drivers at intersections. *Proceedings of the ACM on Human-Computer Interaction* 6, MHCI (2022), 1–16.

[40] Mark Colley, Lukas Gruler, Marcel Woide, and Enrico Rukzio. 2021. Investigating the design of information presentation in take-over requests in automated vehicles. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, 1–15.

[41] Mark Colley, Svenja Krauss, Mirjam Lanzer, and Enrico Rukzio. 2021. How should automated vehicles communicate critical situations? A comparative analysis of visualization concepts. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–23.

[42] Mark Colley, Surong Li, and Enrico Rukzio. 2021. Increasing pedestrian safety using external communication of autonomous vehicles for signalling hazards. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, 1–10.

[43] Mark Colley, Max Rädler, Jonas Glimmann, and Enrico Rukzio. 2022. Effects of scene detection, scene prediction, and maneuver planning visualizations on trust, situation awareness, and cognitive load in highly automated vehicles. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–21.

[44] David Crundall, Ben Andrews, Editha Van Loon, and Peter Chapman. 2010. Commentary training improves responsiveness to hazards in a driving simulator. *Accident Analysis & Prevention* 42, 6 (2010), 2117–2124.

[45] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. 2024. Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles. *IEEE Intelligent Transportation Systems Magazine* 16 (2024), 81–94.

[46] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 958–979.

[47] Zhihao Cui, Meng Li, Yanjun Huang, Yulei Wang, and Hong Chen. 2022. An interpretation framework for autonomous vehicles decision-making via SHAP and RF. In *Proceedings of the 6th CAA International Conference on Vehicular Control and Intelligence (CVCI '22).* IEEE, 1–7.

[48] Luca Cultrera, Federico Becattini, Lorenzo Seidenari, Pietro Pala, and Alberto Del Bimbo. 2023. Explaining autonomous driving with visual attention and end-to-end trainable region proposals. *Journal of Ambient Intelligence and Humanized Computing* (2023), 1–13.

[49] Luca Cultrera, Lorenzo Seidenari, Federico Becattini, Pietro Pala, and Alberto Del Bimbo. 2020. Explaining autonomous driving by learning end-to-end visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 340–341.

[50] Mauro Da Lio, Riccardo Dona, Gastone Pietro Rosati Papini, and Kevin Gurney. 2020. Agent architecture for adaptive behaviors in autonomous driving. *IEEE Access* 8 (2020), 154906–154923.

[51] Shengzhe Dai, Zhiheng Li, Li Li, Nanning Zheng, and Shuofeng Wang. 2020. A flexible and explainable vehicle motion prediction and inference framework combining semi-supervised AOG and ST-LSTM. *IEEE Transactions on Intelligent Transportation Systems* 23, 2 (2020), 840–860.

[52] Henrik Detjen, Maurizio Salini, Jan Kronenberger, Stefan Geisler, and Stefan Schneegass. 2021. Towards transparent behavior of automated vehicles: Design and evaluation of HUD concepts to support system predictability through motion intent communication. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, 1–12.

[53] Debargha Dey, Azra Habibovic, Andreas Löcken, Philipp Wintersberger, Bastian Pfleging, Andreas Riener, Marieke Martens, and Jacques Terken. 2020. Taming the eHMI jungle: A classification taxonomy to guide, compare, and assess the design principles of automated vehicles' external human-machine interfaces. *Transportation Research Interdisciplinary Perspectives* 7 (2020), 100174.

[54] Debargha Dey, Marieke Martens, Chao Wang, Felix Ros, and Jacques Terken. 2018. Interface concepts for intent communication from autonomous vehicles to vulnerable road users. In *Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 82–86.

[55] Cyriel Diels and Simon Thompson. 2018. Information expectations in highly and fully automated vehicles. In *Proceedings of the AHFE 2017 International Conference on Human Factors in Transportation on Advances in Human Aspects of Transportation*. Neville A. Stanton (Ed.), Vol. 8. Springer, 742–748.

[56] Jiqian Dong, Sikai Chen, Shuya Zong, Tiantian Chen, and Samuel Labi. 2021. Image transformer for explainable autonomous driving system. In *Proceedings of the IEEE International Intelligent Transportation Systems Conference (ITSC '21).* IEEE, 2732–2737.

[57] Jinzhen Dou, Shanguang Chen, Zhi Tang, Chang Xu, and Chengqi Xue. 2021. Evaluation of multimodal external human–machine interface for driverless vehicles in virtual reality. *Symmetry* 13, 4 (2021), 687.

[58] Na Du, Jacob Haspiel, Qiaoning Zhang, Dawn Tilbury, Anuj K. Pradhan, X. Jessie Yang, and Lionel P. Robert Jr. 2019. Look Who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation Research Part C: Emerging Technologies* 104 (2019), 428–442.

[59] Na Du, Feng Zhou, Dawn Tilbury, Lionel Peter Robert, and X. Jessie Yang. 2021. Designing alert systems in takeover transitions: The effects of display information and modality. In *Proceedings of the 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 173–180.

[60] Yuemeng Du, Jingyan Qin, Shujing Zhang, Sha Cao, and Jinhua Dou. 2018. Voice user interface interaction design research based on user mental model in autonomous vehicle. In *Proceedings of the 20th International Conference on Interaction Technologies. Human-Computer Interaction (HCI International '18).* Springer, 117–132.

[61] Patrick Ebel, Christoph Lingenfelder, and Andreas Vogelsang. 2023. On the forces of driver distraction: Explainable predictions for the visual demand of in-vehicle touchscreen interactions. *Accident Analysis & Prevention* 183 (2023), 106956.

[62] EC. 2019. European Commission— Ethics Guidelines for Trustworthy AI. Retrieved March 13, 2024 from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[63] Aaron Edelmann, Stefan Stümper, and Tibor Petzoldt. 2021. Cross-cultural differences in the acceptance of decisions of automated vehicles. *Applied Ergonomics* 92 (2021), 103346.

[64] Iveta Eimontaite, Alexandra Voinescu, Chris Alford, Praminda Caleb-Solly, and Phillip Morgan. 2020. The impact of different human-machine interface feedback modalities on older participants' user experience of CAVs in a simulator environment. In *Proceedings of the International Conference on Human Factors in Transportation (AHFE '19)*, Vol. 10. Springer, 120–132.

[65] Fredrick Ekman, Mikael Johansson, and Jana Sochor. 2017. Creating appropriate trust in automated vehicle systems: A framework for HMI design. *IEEE Transactions on Human-Machine Systems* 48, 1 (2017), 95–101.

[66] Mohamed Elbanhawi, Milan Simic, and Reza Jazar. 2015. In the passenger seat: Investigating ride comfort measures in autonomous cars. *IEEE Intelligent Transportation Systems Magazine* 7, 3 (2015), 4–17.

[67] Stefanie M. Faas and Martin Baumann. 2021. Pedestrian assessment: Is displaying automated driving mode in self-driving vehicles as relevant as emitting an engine Sound in electric vehicles? *Applied Ergonomics* 94 (2021), 103425.

[68] Stefanie M. Faas, Lesley-Ann Mathis, and Martin Baumann. 2020. External HMI for Self-driving vehicles: Which information shall be displayed? *Transportation Research Part F: Traffic Psychology and Behaviour* 68 (2020), 171–186.

[69] Yuchao Feng, Wei Hua, and Yuxiang Sun. 2023. NLE-DM: Natural-language explanations for decision making of autonomous driving based on semantic scene understanding. *IEEE Transactions on Intelligent Transportation Systems* 24 (2023), 9780–9791.

[70] Nadia Fereydooni, Einat Tenenboim, Bruce N. Walker, and Srinivas Peeta. 2022. Incorporating situation awareness cues in virtual reality for users in dynamic in-vehicle environments. *IEEE Transactions on Visualization and Computer Graphics* 28, 11 (2022), 3865–3873.

[71] M. Lance Frazier, Paul D. Johnson, and Stav Fainshmidt. 2013. Development and validation of a propensity to trust scale. *Journal of Trust Research* 3, 2 (2013), 76–97.

[72] Anna-Katharina Frison, Philipp Wintersberger, and Andreas Riener. 2019. Resurrecting the ghost in the Shell: A need-centered development approach for optimizing user experience in highly automated vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour* 65 (2019), 439–456.

[73] Ernestine Fu, Mishel Johns, David A. B. Hyde, Srinath Sibi, Martin Fischer, and David Sirkin. 2020. 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*, 1–13.

[74] Francisco Javier Chiyah Garcia, David A Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie. 2018. Explainable autonomy: A study of explanation styles for building clear mental models. In *Proceedings of the 11th International Conference of Natural Language Generation 2018.* Association for Computational Linguistics, 99–108.

[75] Melinda T. Gervasio, Karen L. Myers, Eric Yeh, and Boone Adkins. 2018. Explanation to avert surprise. In *IUI Workshops*, Vol. 2068.

[76] Claudia V. Goldman and Ronit Bustin. 2022. Trusting explainable autonomous driving: Simulated studies. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV '22).* IEEE, 1255–1260.

[77] Noah Goodall. 2024. Normalizing crash risk of partially automated vehicles under sparse data. *Journal of Transportation Safety & Security* 16, 1 (2024), 1–17.

[78] Julia Graefe, Selma Paden, Doreen Engelhardt, and Klaus Bengler. 2022. Human centered explainability for intelligent vehicles—A user study. In *Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 297–306.

[79] Maria J. Grant and Andrew Booth. 2009. A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal* 26, 2 (2009), 91–108.

[80] Balint Gyevnar. 2022. Cars that explain: Building trust in autonomous vehicles through explanations and conversations. Retrieved April 12, 2022.

[81] Balint Gyevnar, Massimiliano Tamborski, Cheng Wang, Christopher G. Lucas, Shay B. Cohen, and Stefano V. Albrecht. 2022. A human-centric method for generating causal explanations in natural language for autonomous vehicle motion planning. arXiv:2206.08783. Retrieved from https://arxiv.org/abs/2206.08783

[82] Taehyun Ha, Sangyeon Kim, Donghak Seo, and Sangwon Lee. 2020. Effects of explanation types and perceived risk on trust in autonomous vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour* 73 (2020), 271–280.

[83] Azra Habibovic, Victor Malmsten Lundgren, Jonas Andersson, Maria Klingegård, Tobias Lagström, Anna Sirkka, Johan Fagerlönn, Claes Edgren, Rikard Fredriksson, Stas Krupenia, et al. 2018. Communicating intent of automated vehicles to pedestrians. *Frontiers in Psychology* 9 (2018), 1336.

[84] Mathias Haimerl, Mark Colley, and Andreas Riener. 2022. Evaluation of common external communication concepts of automated vehicles for people with intellectual disabilities. *Proceedings of the ACM on Human-Computer Interaction* 6, MHCI (2022), 1–19.

[85] Benjamin Hardin, Pericle Salvini, Marina Jirotka, and Lars Kunze. 2024. How well do drivers adapt to remote operation? Learning from remote drivers with on-road experience (2024).

[86] Franziska Hartwich, Cornelia Hollander, Daniela Johannmeyer, and Josef F. Krems. 2021. Improving passenger experience and trust in automated vehicles through user-adaptive HMIs: "The more the better" does not apply to everyone. *Frontiers in Human Dynamics* 3 (2021), 669030.

[87] Franziska Hartwich, Claudia Witzlack, Matthias Beggiato, and Josef F. Krems. 2019. The first impression counts—A combined driving simulator and test track study on the development of trust and acceptance of highly automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour* 65 (2019), 522–535.

[88] Jacob Haspiel, Na Du, Jill Meyerson, Lionel P. Robert Jr, Dawn Tilbury, X. Jessie Yang, and Anuj K. Pradhan. 2018. Explanations and expectations: Trust building in automated vehicles. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 119–120.

[89] Tobias Hecht, Stefanie Weng, Luca-Felix Kick, and Klaus Bengler. 2022. How users of automated vehicles benefit from predictive ambient light displays. *Applied Ergonomics* 103 (2022), 103762.

[90] Sabrina M. Hegner, Ardion D. Beldad, and Gary J. Brunswick. 2019. In automatic we trust: Investigating the impact of trust, control, personality characteristics, and extrinsic and intrinsic motivations on the acceptance of autonomous vehicles. *International Journal of Human–Computer Interaction* 35, 19 (2019), 1769–1780.

[91] Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. 2013. Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 210–217.

[92] Bassam Helou, Aditya Dusi, Anne Collin, Noushin Mehdipour, Zhiliang Chen, Cristhian Lizarazo, Calin Belta, Tichakorn Wongpiromsarn, Radboud Duintjer Tebbens, and Oscar Beijbom. 2021. The reasonable crowd: Towards evidence-based and interpretable models of driving behavior. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '21)*. IEEE, 6708–6715.

[93] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors* 57, 3 (2015), 407–434.

[94] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. arXiv:1812.04608. Retrieved from https://arxiv.org/abs/1812.04608

[95] Kai Holländer, Mark Colley, Enrico Rukzio, and Andreas Butz. 2021. A taxonomy of vulnerable road users for HCI based on a systematic literature review. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13.

[96] Kai Holländer, Philipp Wintersberger, and Andreas Butz. 2019. Overtrust in external cues of automated vehicles: An experimental investigation. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 211–221.

[97] Yijie Hou, Chengshun Wang, Junhong Wang, Xiangyang Xue, Xiaolong Luke Zhang, Jun Zhu, Dongliang Wang, and Siming Chen. 2021. Visual evaluation for autonomous driving. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 1030–1039.

[98] Christian Hubschneider, André Bauer, Jens Doll, Michael Weber, Sebastian Klemm, Florian Kuhnt, and J. Marius Zöllner. 2017. Integrating end-to-end learned steering Into probabilistic autonomous driving. In *Proceedings of the IEEE 20th International Conference on Intelligent Transportation Systems (ITSC '17)*. IEEE, 1–7.

[99] Muhammad Irfan and Munir Ahmad. 2021. Relating consumers' information and willingness to buy electric vehicles: Does personality matter? *Transportation Research Part D: Transport and Environment* 100 (2021), 103049.

[100] ISO/TR23049. 2018. Road vehicles: Ergonomic aspects of external visual communication from automated vehicles to other road users. Standard. International Organization for Standardization.

[101] J3134/201905. 2019. Automated driving system (ADS) Marker Lamp. SAE International. Retrieved March 13, 2024 from sae.org.

[102] Saurabh Jain, Adarsh Kumar, Keshav Kaushik, and Rajalakshmi Krishnamurthi. 2022. Autonomous driving systems and experiences: A comprehensive survey. *Autonomous and Connected Heavy Vehicle Technology* (2022), 65–80.

[103] Pascal Jansen, Mark Colley, and Enrico Rukzio. 2022. A design space for human sensor and actuator focused in-vehicle interaction based on a systematic literature review. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–51.

[104] Adam Sebastian Jardim, Alex Michael Quartulli, and Sean Vincent Casley. 2013. *A Study of Public Acceptance of Autonomous Cars*. Worcester Polytechnic Institute, Worcester, MA, 156.

[105] Suresh Kumaar Jayaraman, Chandler Creech, Dawn M. Tilbury, X. Jessie Yang, Anuj K. Pradhan, Katherine M. Tsui, and Lionel P. Robert Jr. 2019. Pedestrian trust in automated vehicles: Role of traffic signal and AV driving behavior. *Frontiers in Robotics and AI* 6 (2019), 117.

[106] Seonghoon Jeong, Sangho Lee, Hwejae Lee, and Huy Kang Kim. 2023. X-CANIDS: Signal-aware explainable intrusion detection system for controller area network-based in-vehicle network. *IEEE Transactions on Vehicular Technology* 73 (2023), 3230–3246.

[107] Xia Jiang, Jian Zhang, and Bo Wang. 2022. Energy-efficient driving for adaptive traffic signal control environment via explainable reinforcement learning. *Applied Sciences* 12, 11 (2022), 5380.

[108] Taotao Jing, Haifeng Xia, Renran Tian, Haoran Ding, Xiao Luo, Joshua Domeyer, Rini Sherony, and Zhengming Ding. 2022. Inaction: Interpretable action decision making for autonomous driving. In *Proceedings of the European Conference on Computer Vision*. Springer, 370–387.

[109] Philip Joisten, Emanuel Alexandi, Robin Drews, Liane Klassen, Patrick Petersohn, Alexander Pick, Sarah Schwindt, and Bettina Abendroth. 2020. Displaying vehicle driving mode–effects on pedestrian behavior and perceived safety. In *Proceedings of the 2nd International Conference on Human Systems Engineering and Design on Human Systems Engineering and Design II* : Future Trends and Applications *(IHSED '19)*. Springer, 250–256.

[110] Muhammad Monjurul Karim, Yu Li, and Ruwen Qin. 2022. Toward explainable artificial intelligence for early anticipation of traffic accidents. *Transportation Research Record* 2676, 6 (2022), 743–755.

[111] Sherrie-Anne Kaye, Ioni Lewis, Sonja Forward, and Patricia Delhomme. 2020. A priori acceptance of highly automated cars in Australia, France, and Sweden: A theoretically-informed investigation guided by the TPB and UTAUT. *Accident Analysis & Prevention* 137 (2020), 105441.

[112] Ryan Othniel Kearns. 2023. Contextual trust. arXiv:2303.08900. Retrieved from https://arxiv.org/abs/2303.08900

[113] Liam Kettle and Yi-Ching Lee. 2022. Augmented reality for vehicle-driver communication: A systematic review. *Safety* 8, 4 (2022), 84.

[114] Sakib Mahmud Khan, M. Sabbir Salek, Vareva Harris, Gurcan Comert, Eric Morris, and Mashrur Chowdhury. 2023. Autonomous vehicles for all? *Journal on Autonomous Transportation Systems* 1 (2023), 1–8.

[115] Majid Khonji, Jorge Dias, Rashid Alyassi, Fahad Almaskari, and Lakmal Seneviratne. 2020. A risk-aware architecture for autonomous vehicle operation under uncertainty. In *Proceedings of the IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR '20)*. IEEE, 311–317.

[116] Jinkyu Kim and John Canny. 2017. Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE International Conference on Computer Vision*, 2942–2950.

[117] Jinkyu Kim and John Canny. 2018. Explainable deep driving by visualizing causal attention. *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baró, Yağmur Güçlütürk, Umut Güçlü, Marcel van Gerven (Eds.), Springer Nature, 173–193.

[118] Jinkyu Kim, Anna Rohrbach, Zeynep Akata, Suhong Moon, Teruhisa Misu, Yi-Ting Chen, Trevor Darrell, and John Canny. 2021. Toward explainable and advisable model for self-driving cars. *Applied AI Letters* 2, 4 (2021), e56.

[119] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual explanations for self-driving vehicles. In *Proceedings of the European Conference on Computer Vision (ECCV '18)*, 563–578.

[120] Pawit Kochakarn, Daniele De Martini, Daniel Omeiza, and Lars Kunze. 2023. Explainable action prediction through Self-supervision on scene graphs. arXiv:2302.03477. Retrieved from https://arxiv.org/abs/2302.03477

[121] Suresh Kolekar, Shilpa Gite, Biswajeet Pradhan, and Abdullah Alamri. 2022. Explainable AI in scene understanding for autonomous vehicles in unstructured traffic environments on Indian roads using the inception U-Net model with grad-CAM visualization. *Sensors* 22, 24 (2022), 9677.

[122] Jeamin Koo, Jungsuk Kwac, Wendy Ju, Martin Steinert, Larry Leifer, and Clifford Nass. 2015. 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)* 9 (2015), 269–275.

[123] Philip Koopman. 2018. Practical experience report: Automotive safety practices vs. accepted principles. In *Proceedings of the 37th International Conference on Computer Safety, Reliability, and Security (SAFECOMP '18)*. Springer, 3–11.

[124] Hana Kopecka and Jose Such. 2020. Explainable AI for cultural minds. In *Workshop on Dialogue, Explanation and Argumentation for Human-Agent Interaction*.

[125] Moritz Körber, Eva Baseler, and Klaus Bengler. 2018. Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics* 66 (2018), 18–31.

[126] Moritz Körber and Klaus Bengler. 2014. Potential individual differences regarding automation effects in automated driving. In *Proceedings of the XV International Conference on Human Computer Interaction*, 1–7.

[127] Johannes Kraus, David Scholz, Dina Stiegemeier, and Martin Baumann. 2020. The more you know: Trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Human Factors* 62, 5 (2020), 718–736.

[128] Ioannis Krontiris, Kalliroi Grammenou, Kalliopi Terzidou, Marina Zacharopoulou, Marina Tsikintikou, Foteini Baladima, Chrysi Sakellari, and Konstantinos Kaouras. 2020. Autonomous vehicles: Data protection and ethical considerations. In *Proceedings of the 4th ACM Computer Science in Cars Symposium*, 1–10.

[129] Matti Krüger, Tom Driessen, Christiane B. Wiebel-Herboth, Joost C. F. de Winter, and Heiko Wersing. 2020. Feeling uncertain—Effects of a vibrotactile belt that communicates vehicle sensor uncertainty. *Information* 11, 7 (2020), 353.

[130] Marc Alexander Kühn, Daniel Omeiza, and Lars Kunze. 2023. Textual explanations for automated commentary driving. arXiv:2304.08178. Retrieved from https://arxiv.org/abs/2304.08178

[131] Alexander Kunze, Stephen J. Summerskill, Russell Marshall, and Ashleigh J. Filtness. 2017. Enhancing driving safety and user experience through unobtrusive and function-specific feedback. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct*, 183–189.

[132] Alexander Kunze, Stephen J. Summerskill, Russell Marshall, and Ashleigh J. Filtness. 2018. Augmented reality displays for communicating uncertainty information in automated driving. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 164–175.

[133] Alexander Kunze, Stephen J. Summerskill, Russell Marshall, and Ashleigh J. Filtness. 2018. Preliminary evaluation of variables for communicating uncertainties using a haptic Seat. In *Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 154–158.

[134] Alexander Kunze, Stephen J. Summerskill, Russell Marshall, and Ashleigh J. Filtness. 2019. Automation transparency: Implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics* 62, 3 (2019), 345–360.

[135] Alexander Kunze, Stephen J. Summerskill, Russell Marshall, and Ashleigh J. Filtness. 2019. Conveying uncertainties using peripheral awareness displays in the context of automated driving. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 329–341.

[136] Alexander Kunze, Stephen J. Summerskill, Russell Marshall, and Ashleigh J. Filtness. 2019. Function-specific uncertainty communication in automated driving. *International Journal of Mobile Human Computer Interaction (IJMHCI)* 11, 2 (2019), 75–97.

[137] Lars Kunze, Tom Bruls, Tarlan Suleymanov, and Paul Newman. 2018. Reading between the lanes: Road layout Reconstruction from partially segmented scenes. In *Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITSC '18).* IEEE, 401–408.

[138] Anton Kuznietsov, Balint Gyevnar, Cheng Wang, Steven Peters, and Stefano V. Albrecht. 2024. Explainable AI for safe and trustworthy autonomous driving: A systematic review. arXiv:2402.10086. Retrieved from https://arxiv.org/abs/2402.10086

[139] Mirjam Lanzer, Ina Koniakowsky, Mark Colley, and Martin Baumann. 2023. Interaction effects of pedestrian behavior, smartphone distraction and external communication of automated vehicles on crossing and gaze behavior. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18.

[140] Chaiwoo Lee, Carley Ward, Martina Raue, Lisa D'Ambrosio, and Joseph F. Coughlin. 2017. Age differences in acceptance of self-driving cars: A survey of perceptions and attitudes. In *Proceedings of the 3rd International Conference on Human Aspects of IT for the Aged Population. Aging, Design and User Experience (ITAP '17).* Springer, 3–13.

[141] Donsuk Lee, Yiming Gu, Jerrick Hoang, and Micol Marchetti-Bowick. 2019. Joint interaction and trajectory prediction for autonomous driving using graph neural networks. arXiv:1912.07882. Retrieved from https://arxiv.org/abs/1912.07882

[142] Sangwon Lee, Jeonguk Hong, Gyewon Jeon, Jeongmin Jo, Sanghyeok Boo, Hwiseong Kim, Seoyoon Jung, Jieun Park, Inheon Choi, and Sangyeon Kim. 2023. Investigating effects of multimodal explanations using multiple in-vehicle displays for takeover request in conditionally automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour* 96 (2023), 1–22.

[143] Timothée Lesort, Vincenzo Lomonaco, Andrei Stoian, Davide Maltoni, David Filliat, and Natalia Díaz-Rodríguez. 2020. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information Fusion* 58 (2020), 52–68.

[144] Meng Li, Yulei Wang, Hengyang Sun, Zhihao Cui, Yanjun Huang, and Hong Chen. 2023. Explaining a machine-learning lane change model with maximum entropy Shapley values. *IEEE Transactions on Intelligent Vehicles* 8, 6 (2023), 3620–3628.

[145] Wenbo Li, Yaodong Cui, Yintao Ma, Xingxin Chen, Guofa Li, Guanzhong Zeng, Gang Guo, and Dongpu Cao. 2021. A spontaneous driver emotion facial expression (defe) dataset for intelligent vehicles: Emotions triggered by video-audio clips in driving scenarios. *IEEE Transactions on Affective Computing* 14 (2021), 747–760.

[146] Xiao Li, Guy Rosman, Igor Gilitschenski, Brandon Araki, Cristian-Ioan Vasile, Sertac Karaman, and Daniela Rus. 2021. Learning an explainable trajectory generator using the automaton generative network (AGN). *IEEE Robotics and Automation Letters* 7, 2 (2021), 984–991.

[147] Yeti Li, Murat Dikmen, Thana G. Hussein, Yahui Wang, and Catherine Burns. 2018. To cross or not to cross: Urgency-based external warning displays on autonomous vehicles to improve pedestrian crossing safety. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 188–197.

[148] Zhizhong Li, Xiaohuan Zhou, Xiaoan Wang, and Zhongyin Guo. 2013. Study on subjective and objective safety and application of expressway. *Procedia-Social and Behavioral Sciences* 96 (2013), 1622–1630.

[149] Roman Liessner, Jan Dohmen, and Marco Wiering. 2021. Explainable reinforcement learning for longitudinal control. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART '21).* SciTePress, 874–881.

[150] Amarildo Likmeta, Alberto Maria Metelli, Andrea Tirinzoni, Riccardo Giol, Marcello Restelli, and Danilo Romano. 2020. Combining reinforcement learning with rule-based controllers for transparent and general decision-making in autonomous driving. *Robotics and Autonomous Systems* 131 (2020), 103568.

[151] Hazel Si Min Lim and Araz Taeihagh. 2019. Algorithmic decision-making in AVs: Understanding ethical and technical concerns for smart cities. *Sustainability* 11, 20 (2019), 5791.

[152] Sandra Carrasco Limeros, Sylwia Majchrowska, Joakim Johnander, Christoffer Petersson, and David Fernández Llorca. 2022. Towards explainable motion prediction using heterogeneous graph representations. arXiv:2212.03806. Retrieved from https://arxiv.org/abs/2212.03806

[153] Lei Lin, Weizi Li, Huikun Bi, and Lingqiao Qin. 2021. Vehicle trajectory prediction using LSTMs with spatial–temporal attention mechanisms. *IEEE Intelligent Transportation Systems Magazine* 14, 2 (2021), 197–208.

[154] Liangkai Liu, Sidi Lu, Ren Zhong, Baofu Wu, Yongtao Yao, Qingyang Zhang, and Weisong Shi. 2020. Computing systems for autonomous driving: State of the art and challenges. *IEEE Internet of Things Journal* 8, 8 (2020), 6469–6486.

[155] Peng Liu, Yawen Zhang, and Zhen He. 2019. The effect of population age on the acceptable safety of self-driving vehicles. *Reliability Engineering & System Safety* 185 (2019), 341–347.

[156] Weimin Liu, Qingkun Li, Zhenyuan Wang, Wenjun Wang, Chao Zeng, and Bo Cheng. 2023. A literature review on additional semantic information conveyed from driving automation systems to drivers through advanced in-vehicle hmi just before, during, and right after takeover request. *International Journal of Human–Computer Interaction* 39, 10 (2023), 1995–2015.

[157] Andreas Löcken, Andrii Matviienko, Mark Colley, Debargha Dey, Azra Habibovic, Yee Mun Lee, and Andreas Riener. 2022. Accessible automated automotive workshop series (A3WS): International perspective on inclusive external human-machine interfaces. In *Adjunct Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 192–195.

[158] Alexandre Lombard, Jocelyn Buisson, Abdeljalil Abbas-Turki, Stéphane Galland, and Abderrafiaa Koukam. 2020. Curvature-based geometric approach for the lateral control of autonomous cars. *Journal of the Franklin Institute* 357, 14 (2020), 9378–9398.

[159] Maria Paz Sesmero Lorente, Elena Magán Lopez, Laura Alvarez Florez, Agapito Ledezma Espino, José Antonio Iglesias Martínez, and Araceli Sanchis de Miguel. 2021. Explaining deep learning-based driver models. *Applied Sciences* 11, 8 (2021), 3321.

[160] James R. Lumpkin. 1985. Validity of a brief locus of control scale for survey research. *Psychological Reports* 57, 2 (1985), 655–659.

[161] Stefanie M. Faas, Johannes Kraus, Alexander Schoenhals, and Martin Baumann. 2021. Calibrating pedestrians' trust in automated vehicles: Does an intent display in an external HMI support trust calibration and safe crossing behavior? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–17.

[162] Nan Ma, Zhixuan Wu, Yiu-ming Cheung, Yuchen Guo, Yue Gao, Jiahong Li, and Beijyan Jiang. 2022. A survey of human action recognition and posture prediction. *Tsinghua Science and Technology* 27, 6 (2022), 973–1001.

[163] Jane F. Mackworth. 1968. Vigilance, arousal, and habituation. *Psychological Review* 75, 4 (1968), 308.

[164] Arnav Vaibhav Malawade, Shih-Yuan Yu, Brandon Hsu, Harsimrat Kaeley, Anurag Karra, and Mohammad Abdullah Al Faruque. 2022. Roadscene2vec: A Tool for extracting and embedding road scene-graphs. *Knowledge-Based Systems* 242 (2022), 108245.

[165] Harsh Mankodiya, Dhairya Jadav, Rajesh Gupta, Sudeep Tanwar, Wei-Chiang Hong, and Ravi Sharma. 2022. Od-XAI: Explainable AI-based semantic object detection for autonomous vehicles. *Applied Sciences* 12, 11 (2022), 5310.

[166] Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. 2022. Explainable sparse attention for memory-based trajectory predictors. In *Proceedings of the European Conference on Computer Vision*. Springer, 543–560.

[167] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. 2023. LingoQA: Video question answering for autonomous driving. arXiv:2312.14115. Retrieved from https://arxiv.org/abs/2312.14115

[168] Kimberly D. Martinez and Gaojian Huang. 2022. Exploring the effects of meaningful tactile display on perception and preference in automated vehicles. *Transportation Research Board* (2022).

[169] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *Academy of Management Review* 20, 3 (1995), 709–734.

[170] Wolfgang Messner. 2022. Improving the cross-cultural functioning of deep artificial neural networks through machine enculturation. *International Journal of Information Management Data Insights* 2, 2 (2022), 100118.

[171] Ondrej Miksik, I. Munasinghe, J. Asensio-Cubero, S. Reddy Bethi, S. T. Huang, S. Zylfo, X Liu, T. Nica, A. Mitrocsak, S. Mezza, et al. 2020. Building proactive voice assistants: When and how (not) to interact. arXiv:2005.01322. Retrieved from https://arxiv.org/abs/2005.01322

[172] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[173] Muhammad Monjurul Karim, Yu Li, and Ruwen Qin. 2021. Towards explainable artificial intelligence (XAI) for early anticipation of traffic accidents. arXiv:2108.00273. Retrieved from https://arxiv.org/abs/2108.00273

[174] Noé Monsaingeon, Yanna Carli, Loïc Caroux, Sabine Langlois, and Céline Lemercier. 2021. Indicating the limits of partially automated vehicles with drivers' peripheral vision: An online study. In *Proceedings of the International Conference on Applied Human Factors and Ergonomics*. Springer, 78–85.

[175] Lia Morra, Fabrizio Lamberti, F. Gabriele Pratticó, Salvatore La Rosa, and Paolo Montuschi. 2019. Building trust in autonomous vehicles: Role of virtual reality driving simulators in HMI design. *IEEE Transactions on Vehicular Technology* 68, 10 (2019), 9438–9450.

[176] Prajval Kumar Murali, Mohsen Kaboli, and Ravinder Dahiya. 2022. Intelligent in-vehicle interaction technologies. *Advanced Intelligent Systems* 4, 2 (2022), 2100122.

[177] Richa Nahata, Daniel Omeiza, Rhys Howard, and Lars Kunze. 2021. Assessing and explaining collision Risk in dynamic environments for autonomous driving safety. In *Proceedings of the IEEE International Intelligent Transportation Systems Conference (ITSC '21)*. IEEE, 223–230.

[178] Luca Nannini, Agathe Balayn, and Adam Leon Smith. 2023. Explainability in AI policies: A critical review of communications, reports, regulations, and standards in the EU, US, and UK. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1198–1212.

[179] Ilja Nastjuk, Bernd Herrenkind, Mauricio Marrone, Alfred Benedikt Brendel, and Lutz M Kolbe. 2020. What drives the acceptance of autonomous driving? An investigation of acceptance factors from an end-user's perspective. *Technological Forecasting and Social Change* 161 (2020), 120319.

[180] Frederik Naujoks, Simon Höfling, Christian Purucker, and Kathrin Zeeb. 2018. From partial and high automation to manual driving: Relationship between Non-driving related tasks, drowsiness and take-over performance. *Accident Analysis & Prevention* 121 (2018), 28–42.

[181] Frederik Naujoks, Katharina Wiedemann, Nadja Schömig, Sebastian Hergeth, and Andreas Keinath. 2019. Towards guidelines and verification methods for automated vehicle HMIs. *Transportation Research Part F: Traffic Psychology and Behaviour* 60 (2019), 121–136.

[182] Trung Thanh Nguyen, Kai Holländer, Marius Hoggenmueller, Callum Parker, and Martin Tomitsch. 2019. Designing for projection-based communication between autonomous vehicles and pedestrians. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 284–294.

[183] Fjollë Novakazi, Alexander Eriksson, and Lars-Ola Bligård. 2022. Design for perception-a systematic approach for the design of driving automation systems based on the users' perception. In *Adjunct Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 87–90.

[184] Tomasz Nowak, Michal R. Nowicki, Krzysztof Ćwian, and Piotr Skrzypczyński. 2019. How to improve object detection in a driver assistance system applying explainable deep learning. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV '19)*. IEEE, 226–231.

[185] Daniel Omeiza, Sule Anjomshoae, Helena Webb, Marina Jirotka, and Lars Kunze. 2022. From spoken thoughts to automated driving commentary: Predicting and explaining intelligent vehicles' actions. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV '22)*. IEEE, 1040–1047.

[186] Daniel Omeiza, Raunak Bhattacharyya, Nick Hawes, Marina Jirotka, and Lars Kunze. 2023. Effects of explanation specificity on passengers in autonomous driving. arXiv:2307.00633. Retrieved from https://arxiv.org/abs/2307.00633

[187] Daniel Omeiza, Konrad Kollnig, Helena Web, Marina Jirotka, and Lars Kunze. 2021. Why not explain? Effects of explanations on human perceptions of autonomous driving. In *Proceedings of the IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO '21)*. IEEE, 194–199.

[188] Daniel Omeiza, Helena Web, Marina Jirotka, and Lars Kunze. 2021. Towards accountability: Providing intelligible explanations in autonomous driving. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV '21)*. IEEE, 231–237.

[189] Daniel Omeiza, Helena Webb, Marina Jirotka, and Lars Kunze. 2021. Explanations in autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems* 23, 8 (2021), 10142–10162.

[190] Maria Oskina, Haneen Farah, Peter Morsink, Riender Happee, and Bart van Arem. 2023. Safety assessment of the interaction between an automated vehicle and a cyclist: A controlled field test. *Transportation Research Record* 2677, 2 (2023), 1138–1149.

[191] Chelsea Owensby, Martin Tomitsch, and Callum Parker. 2018. A framework for designing interactions between pedestrians and driverless cars: Insights from a ride-sharing design study. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction*, 359–363.

[192] Paulo Henrique Padovan, Clarice Marinho Martins, and Chris Reed. 2023. Black is the new orange: How to determine AI liability. *Artificial Intelligence and Law* 31, 1 (2023), 133–167.

[193] Rohan Paleja, Yaru Niu, Andrew Silva, Chace Ritchie, Sugju Choi, and Matthew Gombolay. 2022. Learning interpretable, high-performing policies for autonomous driving. arXiv:2202.02352. Retrieved from https://arxiv.org/abs/2202.02352

[194] Kristen Pammer, Helena Predojevic, and Angus McKerral. 2023. Humans vs. machines; Motorcyclists and car drivers differ in their opinion and trust of self-drive vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour* 92 (2023), 143–154.

[195] Hujie Pan, Zining Wang, Wei Zhan, and Masayoshi Tomizuka. 2020. Towards better performance and more explainable uncertainty for 3d object detection of autonomous vehicles. In *Proceedings of the IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC '20).* IEEE, 1–7.

[196] Ilias Panagiotopoulos and George Dimitrakopoulos. 2018. An empirical investigation on consumers' intentions towards autonomous driving. *Transportation Research Part C: Emerging Technologies* 95 (2018), 773–784.

[197] Eleonora Papadimitriou, Chantal Schneider, Juan Aguinaga Tello, Wouter Damen, Max Lomba Vrouenraets, and Annebel Ten Broeke. 2020. Transport safety and human factors in the era of automation: What can transport modes learn from each other? *Accident Analysis & Prevention* 144 (2020), 105656.

[198] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors* 39, 2 (1997), 230–253.

[199] Sabina M. Patel, Elizabeth H. Lazzara, Elizabeth Phillips, Alex Chaparro, Thomas J. Alicia, Jennifer Teves, and Ericka Rovira. 2023. Robotics at the crossroads: A discussion of ethical considerations, moral implications, and inclusive design. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 67. SAGE Publications, Los Angeles, CA, 519–522.

[200] William Payre, Julien Cestac, Nguyen-Thong Dang, Fabrice Vienne, and Patricia Delhomme. 2017. Impact of training and in-vehicle task performance on manual control recovery in an automated car. *Transportation Research Part F: Traffic Psychology and Behaviour* 46 (2017), 216–227.

[201] Anantha Pillai. 2017. Virtual reality based study to analyse pedestrian attitude towards autonomous vehicles. *KTH Royal Institute of Technology* (2017).

[202] Raissa Pokam, Serge Debernard, Christine Chauvin, and Sabine Langlois. 2019. Principles of transparency for autonomous vehicles: First results of an experiment with an augmented reality human–machine interface. *Cognition, Technology & Work* 21 (2019), 643–656.

[203] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. 2024. NuScenes-QA: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, 4542–4550.

[204] Zhiqian Qiao, Jeff Schneider, and John M. Dolan. 2021. Behavior planning at urban intersections through hierarchical reinforcement learning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '21).* IEEE, 2667–2673.

[205] Weina Qu, Hongli Sun, and Yan Ge. 2021. The effects of trait anxiety and the big five personality traits on self-driving car acceptance. *Transportation* 48 (2021), 2663–2679.

[206] Weina Qu, Jing Xu, Yan Ge, Xianghong Sun, and Kan Zhang. 2019. Development and validation of a questionnaire to assess public receptivity toward autonomous vehicles and its relation with the traffic safety climate in China. *Accident Analysis & Prevention* 128 (2019), 78–86.

[207] Rudi Rankin. 2023. LINGO-1: Exploring natural language for autonomous driving—Wayve.AI. Retrieved December 21, 2023 from https://wayve.ai/thinking/lingo-natural-language-autonomous-driving/

[208] Nidhi Rastogi, Sara Rampazzi, Michael Clifford, Miriam Heller, Matthew Bishop, and Karl Levitt. 2022. Explaining RADAR features for detecting spoofing attacks in connected autonomous vehicles. arXiv:2203.00150. Retrieved from https://arxiv.org/abs/2203.00150

[209] Hongping Ren, Hui Gao, He Chen, and Guangzhen Liu. 2022. A survey of autonomous driving scenarios and scenario databases. In *Proceedings of the 9th International Conference on Dependable Systems and Their Applications (DSA '22).* IEEE, 754–762.

[210] Alessandro Renda, Pietro Ducange, Francesco Marcelloni, Dario Sabella, Miltiadis C. Filippou, Giovanni Nardini, Giovanni Stea, Antonio Virdis, Davide Micheli, Damiano Rapone, et al. 2022. Federated learning of explainable AI models in 6G systems: Towards secure and automated vehicle networking. *Information* 13, 8 (2022), 395.

[211] Katrin Renz, Kashyap Chitta, Otniel-Bogdan Mercea, A Koepke, Zeynep Akata, and Andreas Geiger. 2022. Plant: Explainable planning transformers via object-level representations. arXiv:2210.14222. Retrieved from https://arxiv.org/abs/2210.14222

[212] Andreas Riegler, Philipp Wintersberger, Andreas Riener, and Clemens Holzmann. 2019. Augmented reality windshield displays and their potential to enhance user experience in automated driving. *i-com* 18, 2 (2019), 127–149.

[213] Gaith Rjoub, Jamal Bentahar, and Omar Abdel Wahab. 2022. Explainable AI-based federated deep reinforcement learning for trusted autonomous driving. In *Proceedings of the International Wireless Communications and Mobile Computing (IWCMC '22).* IEEE, 318–323.

[214] Alexandros Rouchitsas and Håkan Alm. 2022. Ghost on the windshield: Employing a virtual human character to communicate pedestrian acknowledgement and vehicle intention. *Information* 13, 9 (2022), 420.

[215] Ericka Rovira, Anne Collins McLaughlin, Richard Pak, and Luke High. 2019. Looking for age differences in self-driving vehicles: Examining the effects of automation reliability, driving risk, and physical impairment on trust. *Frontiers in Psychology* 10 (2019), 800.

[216] Peter A. M. Ruijten, Jacques M. B. Terken, and Sanjeev N. Chandramouli. 2018. Enhancing trust in autonomous vehicles through intelligent user interfaces that mimic human behavior. *Multimodal Technologies and Interaction* 2, 4 (2018), 62.

[217] SAE. 2018. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. *SAE Int.* 4970, 724 (2018), 1–5.

[218] Vani Suthamathi Saravanarajan, Rung-Ching Chen, Cheng-Hsiung Hsieh, and Long-Sheng Chen. 2023. Improving semantic segmentation under hazy weather for autonomous vehicles using explainable artificial intelligence and adaptive dehazing approach. *IEEE Access* 11 (2023), 38194–38207.

[219] Anna Schieben, Marc Wilbrink, Carmen Kettwich, Ruth Madigan, Tyron Louw, and Natasha Merat. 2019. Designing the interaction of automated vehicles with other traffic participants: Design considerations based on Human needs and expectations. *Cognition, Technology & Work* 21 (2019), 69–85.

[220] Lukas M Schmidt, Georgios Kontes, Axel Plinge, and Christopher Mutschler. 2021. Can you trust your autonomous car? Interpretable and verifiably safe reinforcement learning. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV '21).* IEEE, 171–178.

[221] Tobias Schneider, Sabiha Ghellal, Steve Love, and Ansgar R. S. Gerlicher. 2021. Increasing the user experience in autonomous driving through different feedback modalities. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 7–10.

[222] Tobias Schneider, Joana Hois, Alischa Rosenstein, Sabiha Ghellal, Dimitra Theofanou-Fülbier, and Ansgar R. S. Gerlicher. 2021. Explain yourself! Transparency for positive UX in autonomous driving. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–12.

[223] Tobias Schneider, Joana Hois, Alischa Rosenstein, Sandra Metzl, Ansgar R. S. Gerlicher, Sabiha Ghellal, and Steve Love. 2023. Don't fail Me! The level 5 autonomous driving information dilemma regarding transparency and user experience. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 540–552.

[224] Maike Schwammberger. 2021. A quest of self-explainability: When causal diagrams meet autonomous urban traffic manoeuvres. In *Proceedings of the IEEE 29th International Requirements Engineering Conference Workshops (REW '21).* IEEE, 195–199.

[225] Moayad Shammut1a, Muhammad Imrana, and Faraz Hasanb. 2021. Autonomous mobilities and social meanings: The case of New Zealand. In *Proceedings of the Conference—Mobilities In Transition: Circulation, Appropriation, Globalization (T2M '21*).

[226] Yuan Shen, Shanduojiao Jiang, Yanlin Chen, and Katie Driggs Campbell. 2020. To explain or not to explain: A study on the necessity of explanations for autonomous vehicles. arXiv:2006.11684. Retrieved from https://arxiv.org/abs/2006.11684

[227] Emily M. Shull, John G. Gaspar, Daniel V. McGehee, and Rose Schmitt. 2022. Using human–machine interfaces to convey feedback in automated driving. *Journal of Cognitive Engineering and Decision Making* 16, 1 (2022), 29–42.

[228] Timo Singer, Jonas Kobbert, Babak Zandi, and Tran Quoc Khanh. 2020. Displaying the driving state of automated vehicles to other road users: An International, virtual reality-based study as a first step for the harmonized regulations of novel signaling devices. *IEEE Transactions on Intelligent Transportation Systems* 23, 4 (2020), 2904–2918.

[229] Missie Smith, Joseph L. Gabbard, and Christian Conley. 2016. Head-up vs. head-down displays: Examining traditional methods of display assessment while driving. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 185–192.

[230] Eduardo Soares, Plamen Angelov, Dimitar Filev, Bruno Costa, Marcos Castro, and Subramanya Nageshrao. 2019. Explainable density-based approach for self-driving actions classification. In *Proceedings of the 18th IEEE International Conference on Machine Learning and Applications (ICMLA '19).* IEEE, 469–474.

[231] Andrea Stocco, Paulo J. Nunes, Marcelo D'Amorim, and Paolo Tonella. 2022. Thirdeye: Attention maps for safe autonomous driving systems. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, 1–12.

[232] Jakob Suchan, Mehul Bhatt, and Srikrishna Varadarajan. 2021. Commonsense visual sensemaking for autonomous driving—On generalised neurosymbolic online abduction integrating vision and semantics. *Artificial Intelligence* 299 (2021), 103522.

[233] Lishan Sun, Zeyu Cheng, Dewen Kong, Yan Xu, Shangwu Wen, and Kangyu Zhang. 2023. Modeling and analysis of human-machine mixed traffic flow considering the influence of the trust level toward autonomous vehicles. *Simulation Modelling Practice and Theory* 125 (2023), 102741.

[234] Xiaoqiang Sun, F. Richard Yu, and Peng Zhang. 2021. A survey on cyber-security of connected and autonomous vehicles (CAVs). *IEEE Transactions on Intelligent Transportation Systems* 23, 7 (2021), 6240–6259.

[235] Chen Tang, Nishan Srishankar, Sujitha Martin, and Masayoshi Tomizuka. 2024. Grounded relational inference: Domain knowledge driven explainable autonomous driving. *IEEE Transactions on Intelligent Transportation Systems* 25 (2024), 10617–10635.

[236] Techspot. 2023. Mercedes-Benz Debuts Turquoise Exterior Lights to Indicate the Car is Self-Driving— techspot.com. Retrieved February 5, 2024 from https://shorturl.at/jmvF6

[237] Sule Tekkesinoglu and Lars Kunze. 2024. From feature importance to natural language explanations using LLMs with RAG. In *Proceedings of Multimodal, Affective and Interactive eXplainable AI (MAI-XAI '24)*, Vol. 3803, 114–132.

[238] UKRI. 2023. Framework for Responsible Research and Innovation—Ukri.org. Retrieved December 20, 2023 from https://shorturl.at/gnwNX

[239] Thomas van Orden and Arnoud Visser. 2021. End-to-end imitation learning for autonomous vehicle steering on a single-camera stream. In *Proceedings of the International Conference on Intelligent Autonomous Systems*. Springer, 212–224.

[240] J. Pablo Nuñez Velasco, Haneen Farah, Bart van Arem, and Marjan P. Hagenzieker. 2019. Studying pedestrians' crossing behavior when interacting with automated vehicles using virtual reality. *Transportation Research Part F: Traffic Psychology and Behaviour* 66 (2019), 1–14.

[241] Himanshu Verma, Guillaume Pythoud, Grace Eden, Denis Lalanne, and Florian Evéquoz. 2019. Pedestrians and visual signs of intent: Towards expressive autonomous passenger shuttles. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–31.

[242] Giulia Vilone and Luca Longo. 2020. Explainable artificial intelligence: A systematic review. arXiv:2006.00093. Retrieved from https://arxiv.org/abs/2006.00093.

[243] Chuyao Wang and Nabil Aouf. 2024. Explainable deep adversarial reinforcement learning approach for robust autonomous driving. *IEEE Transactions on Intelligent Vehicles* (2024).

[244] Chao Wang, Matti Krüger, and Christiane B. Wiebel-Herboth. 2020. "Watch out!": Prediction-level intervention for automated driving. In *Proceedings of the 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 169–180.

[245] Chao Wang, Thomas H. Weisswange, Matti Krueger, and Christiane B. Wiebel-Herboth. 2021. Human-vehicle cooperation on prediction-level: Enhancing automated driving with Human foresight. In *Proceedings of the IEEE Intelligent Vehicles Symposium Workshops (IV Workshops '21)*. IEEE, 25–30.

[246] Shenhao Wang and Jinhua Zhao. 2019. Risk preference and adoption of autonomous vehicles. *Transportation Research Part A: Policy and Practice* 126 (2019), 215–229.

[247] Waymo. 2023. Waypoint—The Official Waymo Blog: Driven by Waymo, Designed with Trust—waymo.com. Retrieved December 21, 2023 from https://waymo.com/blog/2019/08/driven-by-waymo-designed-with-trust/

[248] Adam Waytz, Joy Heafner, and Nicholas Epley. 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology* 52 (2014), 113–117.

[249] Ming Wei and Wei Ren. 2021. A lane changing model based on imitation learning and Gaussian velocity Fields. In *Proceedings of the IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI '21)*. IEEE, 146–153.

[250] Gesa Wiegand, Malin Eiband, Maximilian Haubelt, and Heinrich Hussmann. 2020. "I'd like an explanation for that!" Exploring reactions to unexpected autonomous driving. In *Proceedings of the 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, 1–11.

[251] Gesa Wiegand, Matthias Schmidmaier, Thomas Weber, Yuanting Liu, and Heinrich Hussmann. 2019. I drive-you trust: Explaining driving behavior of autonomous cars. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6.

[252] Marc Wilbrink, Anna Schieben, and Michael Oehl. 2020. Reflecting the automated vehicle's perception and intention: Light-based interaction approaches for on-board HMI in highly automated vehicles. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, 105–107.

[253] Alan Winfield, Eleanor Watson, Takashi Egawa, Emily Barwell, Iain Barclay, Serena Booth, Louise A. Dennis, Helen Hastie, Ali Hossaini, Naomi Jacobs, et al. 2022. *IEEE Standard for Transparency of Autonomous Systems*. Institute of Electrical and Electronics Engineers (IEEE).

[254] Philipp Wintersberger, Hannah Nicklas, Thomas Martlbauer, Stephan Hammer, and Andreas Riener. 2020. Explainable automation: Personalized and adaptive UIs to foster trust and understanding of driving automation systems. In

*Proceedings of the 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 252–261.

[255] Wilfried Wöber, Georg Novotny, Lars Mehnen, and Cristina Olaverri-Monreal. 2020. Autonomous vehicles: Vehicle parameter estimation using variational Bayes and kinematics. *Applied Sciences* 10, 18 (2020), 6317.

[256] Yang Xing, Chen Lv, Dongpu Cao, and Peng Hang. 2021. Toward human-vehicle collaboration: Review and perspectives on human-centered collaborative automated driving. *Transportation Research Part C: Emerging Technologies* 128 (2021), 103199.

[257] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. 2020. Explainable object-induced action decision for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9523–9532.

[258] Mark S. Young and Neville A. Stanton. 2002. Malleable attentional resources theory: A new explanation for the effects of mental underload on performance. *Human Factors* 44, 3 (2002), 365–375.

[259] Shih-Yuan Yu, Arnav Vaibhav Malawade, Deepan Muthirayan, Pramod P. Khargonekar, and Mohammad Abdullah Al Faruque. 2021. Scene-graph augmented data-driven risk assessment of autonomous vehicle decisions. *IEEE Transactions on Intelligent Transportation Systems* 23, 7 (2021), 7941–7951.

[260] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. 2020. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access* 8 (2020), 58443–58469.

[261] Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. 2022. Explainability of deep vision-based autonomous driving systems: Review and challenges. *International Journal of Computer Vision* 130, 10 (2022), 2425–2452.

[262] Jing Zang and Myounghoon Jeon. 2022. The effects of transparency and reliability of in-vehicle intelligent agents on driver perception, takeover performance, workload and situation awareness in conditionally automated vehicles. *Multimodal Technologies and Interaction* 6, 9 (2022), 82.

[263] Bo Zhang, Joost De Winter, Silvia Varotto, Riender Happee, and Marieke Martens. 2019. Determinants of take-over time from automated driving: A meta-analysis of 129 studies. *Transportation Research Part F: Traffic Psychology and Behaviour* 64 (2019), 285–307.

[264] Chenkai Zhang, Daisuke Deguchi, Yuki Okafuji, and Hiroshi Murase. 2023. More persuasive explanation method for end-to-end driving models. *IEEE Access* 11 (2023), 4270–4282.

[265] Ethan Zhang, Ruixuan Zhang, and Neda Masoud. 2023. Predictive trajectory planning for autonomous vehicles at intersections using reinforcement learning. *Transportation Research Part C: Emerging Technologies* 149 (2023), 104063.

[266] Kunpeng Zhang, Xiaoliang Feng, Lan Wu, and Zhengbing He. 2022. Trajectory prediction for autonomous driving using spatial-temporal graph attention transformer. *IEEE Transactions on Intelligent Transportation Systems* 23, 11 (2022), 22343–22353.

[267] Qiaoning Zhang, Xi Jessie Yang, and Lionel P. Robert Jr. 2021. Drivers' age and automated vehicle explanations. *Sustainability* 13, 4 (2021), 1948.

[268] Yiwen Zhang, Wenjia Wang, Xinyan Zhou, Qi Wang, and Xiaohua Sun. 2023. Tactical-level explanation is not enough: Effect of explaining AV's lane-changing decisions on drivers' decision-making, trust, and emotional experience. *International Journal of Human–Computer Interaction* 39, 7 (2023), 1438–1454.

[269] Yiwen Zhang, Xinyan Zhou, Wenjia Wang, Yuanda Hu, and Xiaohua Sun. 2023. Keeping in the lane! Investigating drivers' performance handling silent vs. alerted lateral control failures in monotonous partially automated driving. *International Journal of Industrial Ergonomics* 95 (2023), 103429.

[270] Zhengming Zhang, Renran Tian, Rini Sherony, Joshua Domeyer, and Zhengming Ding. 2022. Attention-based interrelation modeling for explainable automated driving. *IEEE Transactions on Intelligent Vehicles* 8, 2 (2022), 1564–1573.

[271] Symbat Zhanguzhinova, Emese Mako, Attila Borsos, Ágoston Pál Sándor, and Csaba Koren. 2023. Communication between autonomous vehicles and pedestrians: An experimental study using virtual reality. *Sensors* 23, 3 (2023), 1049.

[272] Siyuan Zhou, Xu Sun, Bingjian Liu, and Gary Burnett. 2021. Factors affecting pedestrians' trust in automated vehicles: Literature review and theoretical model. *IEEE Transactions on Human-Machine Systems* 52, 3 (2021), 490–500.

[273] Siyuan Zhou, Xu Sun, Qingfeng Wang, Bingjian Liu, and Gary Burnett. 2023. Examining pedestrians' trust in automated vehicles based on attributes of trust: A qualitative study. *Applied Ergonomics* 109 (2023), 103997.

[274] Jan C. Zoellick, Adelheid Kuhlmey, Liane Schenk, Daniel Schindel, and Stefan Blüher. 2019. Amused, accepted, and used? Attitudes and emotions towards automated vehicles, their relationships, and predictive value for usage intention. *Transportation Research Part F: Traffic Psychology and Behaviour* 65 (2019), 68–78.

Received 20 March 2024; revised 24 December 2024; accepted 9 January 2025