2.8 – Explainability

Practical guidance –automotive

Authors: Daniel Omeiza and Dr Lars Kunze, University of Oxford

Explanations are key for building trust in autonomous systems such as autonomous vehicles (AVs). We identify different types of explanations in challenging driving scenarios. Moreover, we characterize several dimensions for explanations and identify different stakeholders for which explanations are relevant. Furthermore, we develop methods and provide guidance for generating explanations using vehicle perception and action data in dynamic driving scenarios.

The need for explainability

Transparency and Accountability

One generally agreed upon notion of accountability is associated with the process of being called 'to account' to some authority for one's actions [1]. In the human and machine context, accountability is conceptualised as the ability to determine whether the decision of a system was made in compliance with procedural and substantive standards, and importantly, to hold one responsible when there is a failure to meet the standards [2]. In autonomous driving, accountability becomes a challenging issue mainly because of the various operations involved (e.g. perception, localisation, planning, controls, system management, among others) that demand inputs from multiple stakeholders; responsibility gaps are not out of the common for such multifaceted processes.

Achieving accountability requires social interaction and exchange [3]. At one end, the requester of an account seeks answers and rectification while at the other end, the respondent or explainer responds and accepts responsibility if necessary. We expect autonomous systems to be able to provide an account in the form of an explanation that is intelligible to the requester to facilitate the assignment of responsibilities.

There have been debates on how responsibility should be allocated for certain AV accidents. Companies have stated the need to put legal frameworks in place in order to clarify where the responsibility lies in case of the occurrence of an accident after the realisation of fully automated driving [4]. Technical solutions are also being put forward. One such example is the proposal for the use of a 'blackbox', similar to a flight recorder in an aircraft, to facilitate investigations [5]. Shashua and Shalev-Shwartz [6] also advocated for the use of mathematical models to clarify faults in order to facilitate a conclusive determination of responsibility. The social aspect of accountability described by Mulgan [3], will demand that the aforementioned recommended approaches are able to plug into explanation mechanisms where causes and effects of actions can be communicated to the relevant stakeholders in intelligible ways.

Autonomous systems should be able to explain what they have 'seen' (perception), would do (plan), and have done (actions) when demanded. This is critical in accounting for actions that have resulted in undesired, discriminatory, and inequitable outcomes. This means that stakeholders such as passengers or auxiliary drivers who may not have direct involvement in

the management of the AVs should be able to instantaneously request accounts as intelligible explanations for such undesired actions when they occur.

Trust

Trust in the context of automation is considered as a social psychological concept that is important for understanding automation partnership [7]. It is the attitude that an agent or automation will help an individual to achieve their goals in a situation characterised by uncertainty and vulnerability. Trust in automation has been proved to have significantly influence in the acceptance of and reliance on automated systems [8] [9] [10]. Information about the functioning modes of an autonomous system at the user's disposal can help the user create a better understanding of the systems' behaviour, eventually adding to the user's knowledge base [11]. This process has been proved to be useful in calibrating trust. This information in context could be presented as explanations of the operational modes and behaviour of a complex system, such as an AV, especially when it acts outside the expectations of the user.

Trust can break down when there are frequent failures without adequate explanations, and regaining trust once lost can be challenging [12] [13]. For instance, previous reports on AV accidents may have a negative impact on calibrated trust in AVs. According to Hussainet al. [14], a serious challenge evident in intelligent transport systems is the lack of trust from the consumer's perspective. Trust is therefore imperative for achieving widespread deployment and use of AVs. It has been argued that trust is a substantial subjective predicting factor for the adoption of automated driving systems [9] [15] [16]. Trust formation and calibration in AVs have also been considered as a temporal process influenced by prior information or background knowledge [17] [18]. The provision of meaningful explanations to stakeholders (e.g. passengers, pedestrians and other road participants) over time as shown in ([19] [20] [21]]) is therefore an important way to build the necessary trust in AV technology.

Standards and regulations

Standards

Some of the standards provided in [22] are relevant to explainability in autonomous driving. For example, ISO/TR 21707:2008 which specifies a set of standard terminology for defining the quality of data being exchanged between data suppliers and data consumers in the ITS domain is very relevant to AV explainability although not originally intended for explainability. While the quality of data is important, the presentation style and language and the interfaces by which the data is provided are also critical for explanations in autonomous driving. We suggest that this standard and others (in [22]) be explored for the development of more AV explainability related ones, and should be made easily accessible.

Regulations

Regulations regarding the explainability of automated systems are being set by countries and regions. However, these regulations seem to be too general and do not directly specify requirements for specific technologies and stakeholders, especially in autonomous driving. Typical examples are the GDPR 'right to explanation' [23] and transparency act, and the UK's ethics, transparency and accountability framework for automated decision-making [24].

In the AV context, a preliminary consultation paper on autonomous vehicles [25], UK Law Commission states the recommendation of the National Physical Laboratory on explainability in autonomous driving:

It is recommended that the autonomous decision-making systems should be available, and able, to be interrogated post-incident. Similar to GDPR, decisions by automated systems must be explainable and key data streams stored in the run-up, during and after an accident.

While this is related to AVs, the paper failed to provide a more comprehensive guide on requirements.

Explanations can help in assessing and rationalising the actions of an AV (*outcome-based*), and in providing information on the governance of the AV across its design, deployment, and management (*processed-based*). This is in line with the information commission office (ICO) guidelines [26] for general AI systems. We suggest that more concrete regulatory guidelines for AV explainability should be set in line with these two goals.

Aim	Standard and description	Stakeholder
Human Safety	ISO 19237:2017 Pedestrian detection and collision mitigation systems	**Class B and C** AV Developers, Regulators, System auditors, Accident investigators, Insurers
	ISO 22078:2020 Bicyclist detection and collision mitigation systems	
	ISO 26262:2011: Road vehicles – Functional safety. An international standard for functional safety of electrical and/or electronic (E/E) systems in production automobiles (2011). It addresses possible hazards caused by the malfunctioning behaviour of E/E safety-related systems, including the interaction of these systems.	
	ISO 21448:2019: Safety Of The Intended Functionality (SOTIF). Provides guidance on design, verification and validation measures. Guidelines on data collection (e.g. time of day, vehicle speed, weather conditions) (2019). (complementary to ISO 26262).	
	UL 4600: Standard for Safety for Evaluation of Autonomous Products. a safety case approach to ensuring autonomous product safety in general, and self-driving cars in particular.	
	SaFAD: Safety First for Automated Driving. White paper by eleven companies from the automotive industry and automated driving sector about frameworks for development, testing and validation of safe automated passenger vehicles (SAE Level 3/4).	
	RSS (Intel) / SFF (NVIDIA): Formal Models & Methods to evaluate safety of AV on top of ISO 26262 and ISO 21448 (proposed by companies).	
	IEEE Initiatives: “Reliable, Safe, Secure, and Time-Deterministic Intelligent Systems (2019)”; “A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems” (2019); “Assessment of standardization gaps for safe autonomous driving (2019)”.	

	The Autonomous: Global safety reference, created by the community leading automotive industry players, which facilitates the adoption of autonomous mobility on a grand scale (2019).	
Information/ data exchange	ISO/TR 21707:2008: Integrated transport information, management, and control— Data quality in intelligent transport systems (ITS). "specifies a set of standard terminology for defining the quality of data being exchanged between data suppliers and data consumers in the ITS domain" (2018).	**Class A and C** Passengers, Auxiliary Drivers, Pedestrians, Regulators, System auditors, Accident investigators Insurers
	ISO 13111-1:2017: The use of personal ITS station to support ITS service provision for travellers. "Defines the general information and use cases of the applications based on the personal ITS station to provide and maintain ITS services to travellers including drivers, passengers, and pedestrians" (2017).	
	ISO 15075:2003: In-vehicle navigation systems—Communications message set requirements. "Specifies message content and format utilized by in-vehicle navigation systems" (2003).	
	ISO/TR 20545:2017: Vehicle/roadway warning and control systems. "Provides the results of consideration on potential areas and items of standardization for automated driving systems" (2017).	
	ISO 17361:2017: Lane departure warning.	
	ISO/DIS 23150: Data communication between sensors and data fusion unit for automated driving functions.	

Table I: selected standards for autonomous vehicles. These standards underline the importance of safe, transparent, and explainable AVs.

Stakeholders

1. **Class A: End-users**
    - Passenger: this is the in-vehicle agent who may interact with the explanation agency in the AV but is not responsible for any driving operation.
    - Auxiliary driver: This is a special in-vehicle passenger who may also interact with the explanation agency in the AV and can also participate in the driving operations. This kind of participant may mainly exist in SAE level 3 and 4 vehicles.
    - Pedestrian: this is the agent outside the AV (external agent) who may interact with the AV to convey intentions either through gestures or an external human-machine interface (eHMI).
    - Pedestrian with Reduced Mobility (PRM): this is the agent outside the AV (external agent) who may interact with the AV to convey intentions either through gestures or an external human-machine interface (eHMI) but have reduced mobility capacity (e.g., pedestrian in a wheelchair).
    - Other road participants: these are other agents outside the AV (external agent) who may interact with the AV to convey intentions either through gestures or an external human-machine interface (eHMI) (e.g. cyclists, other vehicles).

2. **Class B: Developers and technicians**
    - AV developer: the agent who develops the automation software and tools for AVs.
    - Automobile technicians: the agent who repairs and maintains AVs.
3. **Class C: Regulators and insurers**
    - System auditor: the agent who inspects AV design processes and operations in order to ascertain compliance with regulations and guidelines.
    - Regulator: the agent who sets guidelines and regulations for the design, use, and maintenance of AVs.
    - Accident investigator: the agent who investigates the cause of an accident in which an AV was involved.
    - Insurer: the agent who insures the AV against vandalism, damage, theft, and accidents.

Explanation categorisations

We provide a categorisation of explanations based on the different methodologies identified in the explanation literature for the design and development of explanations techniques. Explanation techniques that are mainly based on the researcher's experience without further user studies to justify claims are categorised under unvalidated guidelines (UG). Those that adopted a user study to elicit users experience are categorised as empirically derived (ED), and those that built on psychology theories as are categorised under psychological constructs from formal theories (PC). Other dimensions for the categorisation include causal filter, explanation style, interactivity, dependence, system, scope, stakeholders, and operation.

The description of the various dimensions of explanation is detailed in [22].

Explanation generation for AVs

Perception - vision-based explanations for AVs

Various methods have been proposed to explain neural networks which are fundamental structures for perception and scene understanding in AVs. Some of the prominent methods are *gradient-based*. Gradient-based or backpropagation methods are generally used for explaining convolutional neural network models. The main logic of these methods is dependent on gradients that are backpropagated from the output prediction layer of the CNN back to the input layer [44]. They are often presented in form of heatmaps. These methods mainly fall under the input influence explanation style in the explanation categorisation presented in Table 4. Many of the vision-based explanations for AVs (e.g. those from Table 3) stem from generic gradient-based methods explained above.

Perception - driving datasets for posthoc explanations

Several driving datasets have been made available for the purpose of training machine learning models for autonomous vehicles (see [45]). Some of these datasets have annotations—e.g. handcrafted explanations [27] [46], vehicle trajectories [47], human driver behaviour [48] [35] or anomaly identification with bounding boxes [30] [46]— that are helpful for posthoc driving behaviour explanation (See detail in [22].

References	Causal filter			Explanation style				Interactivity		Dependence		System		Scope		Method	Stakeholders Class	Operation
	Factual	Contrastive	Counterfactual	InputInfluence	Sensitivity	Case-based	Demographic	Conversational	Non-conversational	ModelAgnostic	ModelSpecific	Goal-Driven	Data-Driven	Local	Global			
Kim et al. [27]						✓			✓		✓		✓	✓		UG + ED	B, C	P, C
Chakraborti et al. [28]	✓			✓					✓		✓	✓			✓	UG	B & C	PL
Raman et al. [29]	✓			✓					✓		✓	✓			✓	UG	B & C	PL
Xu et al. [30]						✓			✓		✓		✓			UG	A & C	P
Kim & Canny [31]	✓		✓			✓			✓		✓		✓	✓		UG	B & A	P
Cultrera et al. [32]	✓			✓					✓		✓		✓	✓		UG	B & A	P
Schneider et al. [33]	✓			✓					✓				✓	✓		ED	A & C	P
Rahimpour et al. [34]				✓					✓		✓		✓	✓		UG	B & C	P
Shen et al. [35]	✓					✓			✓		✓		✓	✓		ED	B	P
Ben-Younes et al. [36]	✓					✓			✓		✓		✓	✓		UG	A & C	P
Nahata et al. [37]	✓		✓	✓				✓		✓			✓	✓		UG	B	PL
Ha et al. [20]	✓			✓					✓			✓				ED + PC	A & B	P
Koo et al. [19]	✓			✓					✓			✓				ED + PC	A & B	P
Bojarski et al. [38]	✓					✓			✓		✓		✓	✓		UG	B & C	P
Mori et al. [39]	✓			✓					✓		✓		✓	✓		UG	B & C	P
Liu et al. [40]	✓			✓					✓			✓		✓		ED	A	P
Omeiza et al. [21]	✓	✓	✓	✓					✓			✓				ED	A	P

Rizzo et al. [41]	✓			✓					✓		✓		✓	✓		UG	B	P
Liu et al. [42]	✓			✓					✓		✓		✓	✓		UG	B & C	P
Omeiza et al. [43]	✓	✓	✓	✓					✓			✓				ED	A	P

Table 2: Summary of explanations categories. The table includes a subset of the reviewed papers where each or a subset of the explanation categories was mentioned in the context of autonomous embodied agents.

Stakeholders: Class A—Passenger (PA), Pedestrian (PE), Pedestrian with Reduced Mobility (PRM), Other Road Participants (ORP), Auxiliary Driver (AD). Class B—Developer (DV), Auto-Mechanic (AM). Class C—System Auditor (SA), Regulator (RG), Insurer (IN), Accident Investigator (AI).

Methods: Unvalidated Guidelines (UG), Empirically Derived (ED), Psychological Constructs from Formal Theories (PC). Operations: Perception (P), Localisation (L), Planning (PL), Control (C), System Management (M)

Localisation

Precise and robust localisation is critical for AVs in complex environments and scenarios [49]. For effective planning and decision making, the position and orientation information is required to be precise in all weather and traffic conditions. Safety is often considered the most important design requirement and it is critical in the derivation of requirements for AVs [50]. Hence, communicating position over time and with justifications as explanations is crucial to expose increasing error rates in a timely manner before they cause an accident. For instance, the position errors can be transmitted continuously through a wireless channel to an operation centre from which the AV is managed. An interface that displays this information (e.g., a special dashboard or mobile application as shown in [33]) is provided and it is able to trigger an alarm for immediate action (e.g. safe parking) when the error margin is exceeded. There seems to be limited work on explainability in this area. However, explanations would be helpful, especially for Class B stakeholders.

Planning

Through AI planning and scheduling, the sequence of actions required for an agent to complete a task are generated. These action sequences are further utilised in influencing the agent's online decisions or behaviours with respect to the dynamics of the environment it operates in [51]. Often, the amount of data (e.g. descriptions of objects, states, and locations) that the AV processes per time is larger than such that a human may be able to process, and continuously and accurately keep track of. Hence, a stakeholder riding in an AV may be left in a confused state when the AV updates its trajectory without providing an explanation. Explainable planning can play a vital role in supporting users and improving their experiences when they interact with autonomous systems in complex decision-making procedures [52]. Relevant work includes XAI-PLAN [53], WHY-PLAN [54], refinement-based planning (RBP) [55], plan explicability and predictability [56], and plan explanation for model reconciliation [28] [57].

Vehicle control

Control in an AV generally has to do with the manipulations of vehicle motions such as lane changing, lane-keeping, and car following. These manipulations are broadly categorised under longitudinal control (speed regulation with throttle and brake) and lateral control (i.e. automatic steering to follow track reference) [58]. Interfaces that come with Advanced Driving Assistance System (ADAS) now display rich digital maps [59], vehicle's position, and track related attributes ahead or around the vehicle. Stakeholders may issue investigatory queries when the AV makes a decision against their expectations. Other than existing in-

vehicle visual interfaces such as mixed reality (MR) visualisation [60], and other flexible (i.e. highly reconfigurable) dashboard panels [61], in-vehicle interfaces that support the exchange of messages between the stakeholder and the AV are crucial. The user should be able to query the interface and receive explanations for navigation and control decisions in an appropriate form; either through voice, text, visual, gesture or a combination of any of these options.

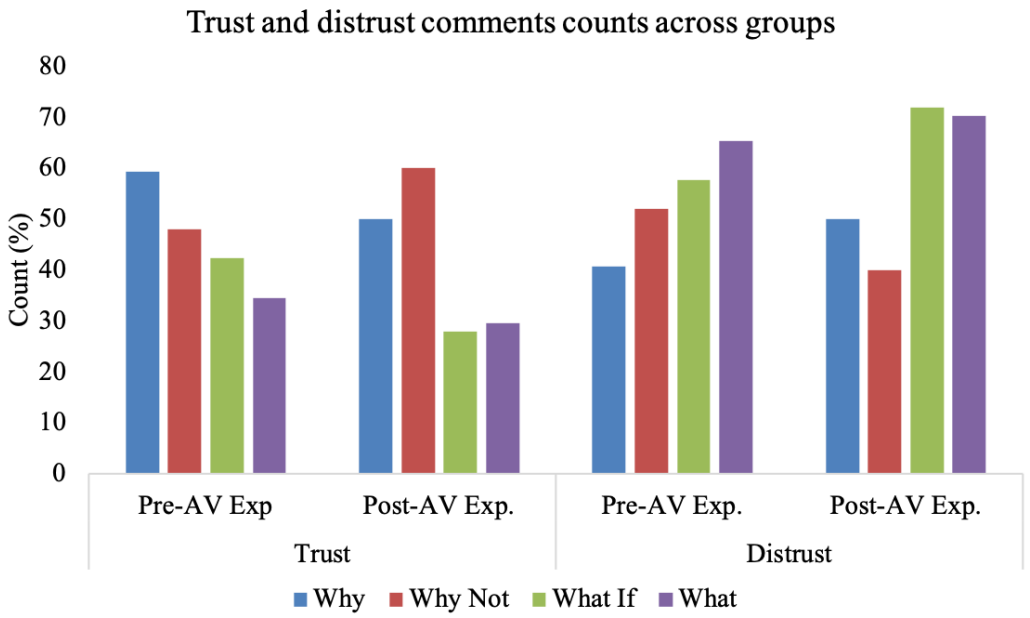

Figure 1: Example from [21] using the trust objective. Frequency of negative (distrust) and positive (trust) comments about trust in AVs. The y-axis indicates the frequency in percentage, while the x-axis indicates the pre-AV and post-AV experiences along with trust and distrust comments. Only the Why Not group had increased positive comments in the post-AV experience questionnaire. Why group received factual explanations, Why Not group received contrastive explanations, What-If group received counterfactual explanations, What group received a non-causal explanation (i.e. action description without the cause of the action).

Evaluating explanations

There are generally quantitative and qualitative means for assessing explanations. Quantitative approaches (e.g. correlating with existing supposed faithful explanation approaches [62] [63], local fidelity [64], change in log-odds [65], data staining [66], use of interpretable ground truths, and observing behaviour through input perturbation) are usually used to assess faithfulness of an explanation to the actual system workings. While qualitative approaches have mostly been applied in assessing the utility of explanations using the intelligibility, plausibility, accountability, confidence, trust, knowledge enhancement etc. objectives; they are usually carried out using user studies. In autonomous systems, especially autonomous driving, the qualitative approaches have been mostly adopted in assessing explanations (e.g. in [21] [43] [19] [20]).

Assessing accountability through explanations

Omeiza et. al [43] proposed a data representation approach to enhance accountability in autonomous driving through explanations. Explanations were constructed to make references to AV's actions, observations in scenes, and road rules. A tree-based (interpretable) approach was proposed to assess accountability along with a user study.

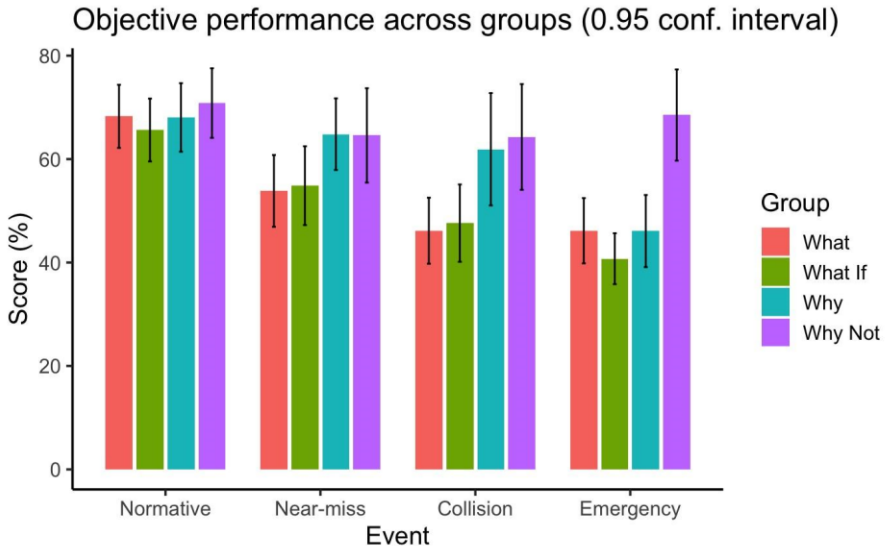

Figure 2: Example from [21] using the knowledge enhancement objective. Task performance in the different driving events. With the exception of the near-miss category, participants in the Why Not group consistently outperformed other groups. Impact of explanation types was greatest in collision and emergency events.

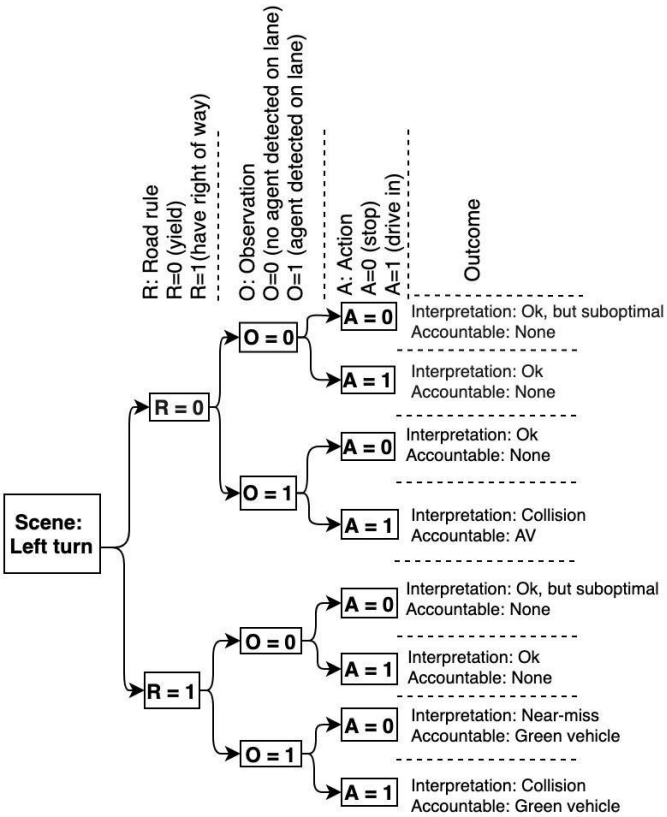

Figure 3: Assessing accountability through explanations. The figure illustrates the underlying tree-based representation used in explanation generation. The tree is constructed with key variables: road rules (R), observations (O), and actions (A). Different types of explanations are generated through different traversals of the tree. We manually interpreted the outcomes indicating accountability (especially in collision incidents) for each path in the tree representation one of the left turn scenarios used in the user study.

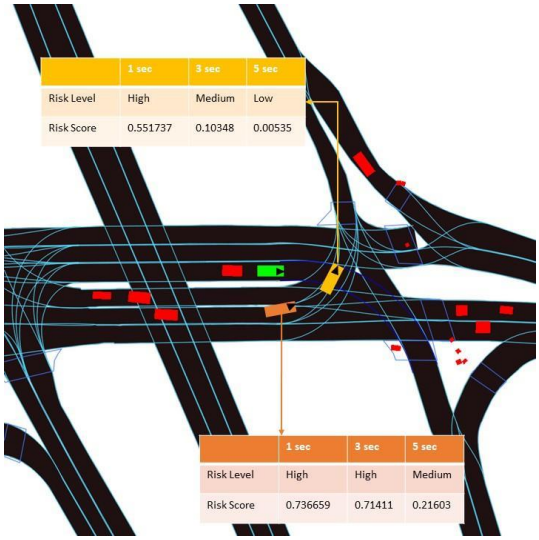

Figure 4: An instance of the Lyft Level5 dataset [47] that visualises a potential left-turn conflict. As before, the ego is depicted in green and two other agents in yellow and orange. The yellow agent is in the direct view of the ego vehicle and has a current high-risk value because the relative distance between them is small. However, due to looming, we realise that these cars will never actually meet since the yellow agent will pass the point before the ego vehicle arrives there, which is indicated by its decreasing risk. On the other hand, the orange vehicle, poses a very high risk to the ego as it might actually reach the point of collision at the same time as the ego, once the yellow vehicle has passed. Thus, this helps the ego to prioritize between different agents, to focus on the riskier one (here the orange agent) and to manoeuvre accordingly.

Generating explanations

Algorithms have been proposed to generate explanations in visual forms (e.g. in [27]) and in natural language forms (e.g. in [37]). In [37], an interpretable algorithm was proposed to generate natural language explanations for collision risks models in autonomous driving. Figure 4 provides an example of the type of scenarios where the risk models were applied.

Explanation 1: RandomForest Regressor, 1*s*

Why: "The predicted risk for the provided agent's attributes is 0.4922 because important features such as 'beta6' has a value between 0.0$rads^{-1}$ and 16.0179$rads^{-1}$, 'agent vel' was below 5.2209ms^{-1}, 'ego vel' was below 0.0001ms^{-1}."

What-If (counterfactual inference): "To get the risk prediction below 0.3, the following conditions should be true: 'alpha6' should be greater than 0.0$rads^{-1}$, 'agent vel' should be above 6.794ms^{-1}."

Explanation 1 shows sample natural language explanations.

Decision-making

Deep learning models are being used to predict AVs' trajectories or high-level plans. In fact, some companies are implementing end-to-end deep learning models to handle core driving operations (perception, localisation, and planning) at a go. These methods need explainer models to generated explanations to relevant stakeholders. In [D0], a transparent algorithm is developed to predict an ego vehicle's actions and then generate intelligible explanations for the prediction based on the ROADS dataset [67]. Figure 5 summarises the entire process, from predicting to explaining.

Activities description

We did an extensive literature review on explanations in the context of autonomous driving [22], [D1], [D2] in which we have identified multiple types and dimensions of explanations as well as the requirements for different stakeholders. We reported on causal explanations (contrastive, non-contrastive, counterfactual) and non-causal explanations as well as how explanations are useful for different stakeholders such as end-users, technicians and engineers, regulators and insurers. In a user study we have evaluated different types of explanations in safety critical scenarios [21] [43]. In [21], we investigated how different types of explanations are perceived by humans in challenging driving scenarios (including near-miss events and accidents). In [43], we tested and reported on the accountability of explanations and evaluated to what extent explanations can help end-users to understand traffic rules.

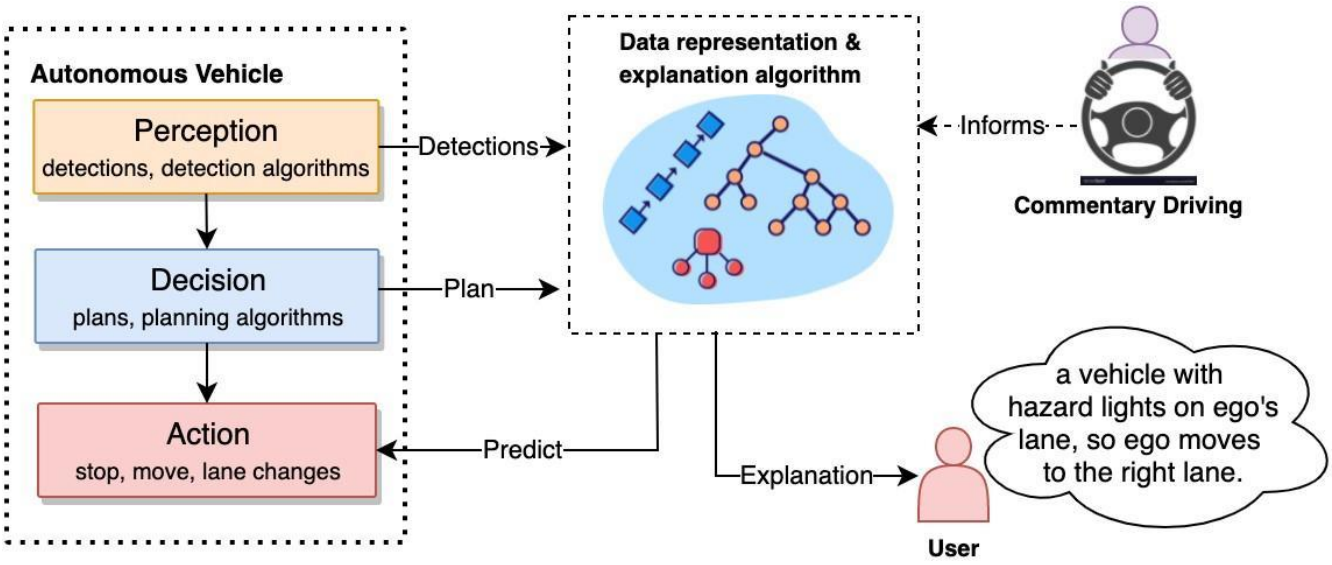

Figure 5: From commentary driving, requirements for explanations are gathered to inform the design of an explanation algorithm. The algorithm receives input data from the different autonomous driving operations, provides a structured representation, and generates intelligible explanations to stakeholders.

In [37], we investigated how explanations can help to explain risk assessments in real-world driving scenarios. We will report on explainable AI (XAI) techniques that allow us to explain important factors when assessing collision risk with other road users (incl. vehicles, pedestrians and cyclists).

Application of approach

We are in the process of developing methods for the (online) generation of explanations (based on perception and action data) [D0] with the aim to integrate and demonstrate the techniques on the Oxford RoboCar dataset as well as the SAX dataset [D3].

References

[1] George W Jones. The search for local accountability. *Strengthening local government in the 1990s*, pages 49–78, 1992.

[2] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood.

Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*, 2017.

[3] Richard Mulgan. 'Accountability': An ever-expanding concept? *Public administration*, 78(3):555–573, 2000.

[4] Honda sustainability report (Tech. Rep.). Honda. Accessed: Jun. 25, 2021.

[5] Sustainable value report (Tech. Rep.). BMW. Accessed: Jun. 25, 2021.

[6] A Shashua and Shai Shalev-Shwartz. A plan to develop safe autonomous vehicles. and prove it. *Intel Newsroom*, page 8, 2017.

[7] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[8] David P Biros, Mark Daly, and Gregg Gunsch. The influence of task load and automation trust on deception detection. *Group Decision and Negotiation*, 13(2):173–189, 2004.

[9] Sebastian Hergeth, Lutz Lorenz, Roman Vilimek, and Josef F Krems. Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human factors*, 58(3):509–519, 2016.

[10] Bonnie M Muir and Neville Moray. Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3):429–460, 1996.

[11] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

[12] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. The role of trust in automation reliance. *International journal of human-computer studies*, 58(6):697–718, 2003.

[13] Poornima Madhavan and Douglas A Wiegmann. Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors*, 49(5):773–785, 2007.

[14] Rasheed Hussain and Sherali Zeadally. Autonomous cars: Research results, issues, and future challenges. *IEEE Communications Surveys & Tutorials*, 21(2):1275–1313, 2018.

[15] William Payre, Julien Cestac, and Patricia Delhomme. Fully automated driving: Impact of trust and practice on manual control recovery. *Human factors*, 58(2):229–241, 2016.

[16] Bako Rajaonah, Fran¸coise Anceaux, and Fabrice Vienne. Trust and the use of adaptive cruise control: a study of a cut-in situation. *Cognition, Technology & Work*, 8(2):146–155, 2006.

[17] Matthias Beggiato and Josef F Krems. The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation research part F: traffic psychology and behaviour*, 18:47–57, 2013.

[18] Johannes Maria Kraus, Yannick Forster, Sebastian Hergeth, and Martin Baumann. Two routes to trust calibration: effects of reliability and brand information on trust in automation. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 11(3):1–17, 2019.

[19] Jeamin Koo, Jungsuk Kwac, Wendy Ju, Martin Steinert, Larry Leifer, and Clifford Nass. Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 9(4):269–275, 2015.

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

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

[22] Daniel Omeiza, Helena Webb, Marina Jirotka, and Lars Kunze. Explanations in autonomous driving: A survey. *arXiv preprint arXiv:2103.05154*, 2021.

[23] Paul Voigt and Axel Von dem Bussche. The EU General Data Protection Regulation (GDPR). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[24] GOV.UK. Ethics, Transparency and Accountability Framework for Automated Decision-Making.

[25] Law Commission. Automated Vehicles: Analysis of Responses to the Preliminary Consultation Paper. Accessed: July, 5.

[26] Law Commission. What goes into an explanation? Accessed: July, 5.

[27] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for selfdriving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578, 2018.

[28] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. *arXiv preprint arXiv:1701.08317*, 2017.

[29] Vasumathi Raman, Constantine Lignos, Cameron Finucane, Kenton CT Lee, Mitchell P Marcus, and Hadas Kress-Gazit. Sorry Dave, I'm Afraid I Can't Do That: Explaining Unachievable Robot Tasks Using Natural Language. In *Robotics: Science and Systems*, volume 2, pages 2–1, 2013.

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

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

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

[33] Tobias Schneider, Joana Hois, Alischa Rosenstein, Sabiha Ghellal, Dimitra Theofanou-Fu¨lbier, and Ansgar RS Gerlicher. Explain yourself! transparency for positive ux in autonomous driving. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2021.

[34] Alireza Rahimpour, Sujitha Martin, Ashish Tawari, and Hairong Qi. Context aware road-user importance estimation (icare). In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pages 2337–2343, 2019.

[35] Yuan Shen, Shanduojiao Jiang, Yanlin Chen, Eileen Yang, Xilun Jin, Yuliang Fan, and Katie Driggs Campbell. To explain or not to explain: A study on the necessity of explanations for autonomous vehicles. *arXiv preprint arXiv:2006.11684*, 2020.

[36] H´edi Ben-Younes, Eloi Zablocki, Patrick P´erez, and Matthieu Cord. Driving behavior explanation with multi-´ level fusion. *arXiv preprint arXiv:2012.04983*, 2020.

[37] Richa Nahata, Daniel Omeiza, Rhys Howard, and Lars Kunze. Assessing and Explaining Collision Risk in Dynamic Environments for Autonomous Driving Safety. In *24th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2021.

[38] Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Larry J Ackel, Urs Muller, Phil Yeres, and Karol Zieba. Visualbackprop: Efficient visualization of CNNs for autonomous driving. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4701–4708, 2018.

[39] Keisuke Mori, Hiroshi Fukui, Takuya Murase, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Visual explanation by attention branch network for end-to-end learning-based self-driving. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1577–1582, 2019.

[40] Tianwei Liu, Huiping Zhou, Makoto Itoh, and Satoshi Kitazaki. The impact of explanation on possibility of hazard detection failure on driver intervention under partial driving automation. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 150–155, 2018.

[41] Stefano Giovanni Rizzo, Giovanna Vantini, and Sanjay Chawla. Reinforcement learning with explainability for traffic signal control. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 3567–3572, 2019.

[42] Yi-Chieh Liu, Yung-An Hsieh, Min-Hung Chen, C-H Huck Yang, Jesper Tegner, and Y-C James Tsai. Interpretable self-attention temporal reasoning for driving behavior understanding. In *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2338–2342, 2020.

[43] Daniel Omeiza. Towards accountability: Providing intelligible explanations in autonomous driving. In *Proceedings of the 32nd IEEE Intelligent Vehicles Symposium*, 2021.

[44] Arun Das and Paul Rad. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *arXiv preprint arXiv:2006.11371*, 2020.

[45] Joel Janai, Fatma Gu¨ney, Aseem Behl, Andreas Geiger, et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends in Computer Graphics and Vision*, 12(1– 3):1–308, 2020.

[46] Tackgeun You and Bohyung Han. Traffic accident benchmark for causality recognition. In *European Conference on Computer Vision*, pages 540–556. Springer, 2020.

[47] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. *arXiv preprint arXiv:2006.14480*, 2020.

[48] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7699–7707, 2018.

[49] Liang Wang, Yihuan Zhang, and Jun Wang. Map-based localization method for autonomous vehicles using 3d-lidar. *IFAC-PapersOnLine*, 50(1):276–281, 2017.

[50] Tyler GR Reid, Sarah E Houts, Robert Cammarata, Graham Mills, Siddharth Agarwal, Ankit Vora, and Gaurav Pandey. Localization requirements for autonomous vehicles. *arXiv preprint arXiv:1906.01061*, 2019.

[51] F´elix Ingrand and Malik Ghallab. Deliberation for autonomous robots: A survey. *Artificial Intelligence*, 247:10–44, 2017.

[52] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. The emerging landscape of explainable automated planning & decision making. 2020.

[53] Rita Borgo, Michael Cashmore, and Daniele Magazzeni. Towards providing explanations for AI planner decisions. *arXiv preprint arXiv:1810.06338*, 2018.

[54] Raj Korpan and Susan L Epstein. Toward natural explanations for a robot's navigation plans. *Notes from the Explainable Robotic Systems Workshop, Human-Robot Interaction*, 2018.

[55] Julien Bidot, Susanne Biundo, Tobias Heinroth, Wolfgang Minker, Florian Nothdurft, and Bernd Schattenberg. Verbal plan explanations for hybrid planning. In *MKWI*, pages 2309–2320. Citeseer, 2010.

[56] Yu Zhang, Sarath Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. Plan explicability and predictability for robot task planning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1313–1320, 2017.

[57] Tathagata Chakraborti, Sarath Sreedharan, Sachin Grover, and Subbarao Kambhampati. Plan Explanations as Model Reconciliation–An Empirical Study. In *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 258–266, 2019.

[58] Alireza Khodayari, Ali Ghaffari, Sina Ameli, and Jamal Flahatgar. A historical review on lateral and longitudinal control of autonomous vehicle motions. In *International Conference on Mechanical and Electrical Technology*, pages 421–429, 2010.

[59] Tomtom launches map-based adas software platform virtual horizon. Accessed: Aug. 10, 2021.

[60] Shota Sasai, Itaru Kitahara, Yoshinari Kameda, Yuichi Ohta, Masayuki Kanbara, Yoichi Morales, Norimichi Ukita, Norihiro Hagita, Tetsushi Ikeda, and Kazuhiko Shinozawa. MR visualization of wheel trajectories of driving vehicle by seeing-through dashboard. In *IEEE International Symposium on Mixed and Augmented Reality Workshops*, pages 40–46, 2015.

[61] Luis Marques, Ver´onica Vasconcelos, Paulo Pedreiras, and Luis Almeida. A flexible dashboard panel for a small electric vehicle. In *6th Iberian Conference on Information Systems and Technologies (CISTI 2011)*, pages 1–4. IEEE, 2011.

[62] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[63] David Alvarez-Melis and Tommi S Jaakkola. Towards robust interpretability with self-explaining neural networks. *arXiv preprint arXiv:1806.07538*, 2018.

[64] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[65] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.

[66] Jacob Sippy, Gagan Bansal, and Daniel S Weld. Data staining: A method for comparing faithfulness of explainers. In *Proceedings of the 2020 ICML Workshop on Human Interpretability in Machine Learning (WHI 2020). International Conference on Machine Learning, Virtual*, volume 7, 2020.

[67] Gurkirt Singh, Stephen Akrigg, Manuele Di Maio, Valentina Fontana, Reza Javanmard Alitappeh, Suman Saha, Kossar Jeddisaravi, Farzad Yousefi, Jacob Culley, Tom Nicholson, et al. Road: The road event awareness dataset for autonomous driving. *arXiv preprint arXiv:2102.11585*, 2021.