Cars that Explain: Building Trust in Autonomous Vehicles through Explanations and Conversations

Balint Gyevnar
School of Informatics
University of Edinburgh
Edinburgh, UK
balint.gyevnar@ed.ac.uk

***Abstract*—Autonomous vehicles are subject to skepticism from the general public due to reports of fatal accidents and a lack of trust in the technology. Yet these vehicles are predicted to have important advantages over human drivers, which means their common adoption could help contribute to solutions to road safety, traffic jams, or energy consumption. I believe that achieving public recognition comes through transparency and accessibility. I propose a system called Explainable Autonomous Vehicle Intelligence (XAVI), that will give relevant and accurate explanations about its behaviour in a natural, conversational manner. By minimising the black-box nature of current AV techniques and giving easy-to-understand explanations, XAVI can make autonomous vehicles trustworthy and promote the advantages that their general adoption could entail.**

***Index Terms*—transparency, trust, autonomous vehicles, intelligent transport systems**

I. Introduction

The general public feels skeptical about vehicle autonomy [1]. Despite impressive advances in recent years in fields such as computer vision, intention prediction, and navigation, with autonomous vehicles (AVs) covering millions of kilometers of distance, accidents involving AVs seem to have reinforced distrust in people about their capabilities [2].

Therefore, we must build *trust* and *understanding* in AVs [3], [4]. It is my vision for the future to create a trustworthy and accessible intelligent transport system, that provides high level of autonomy without driver input while being completely *transparent* about its internal processes. I envision such a system as being able to explain any of its decisions to external inquirers in natural conversations, much like how humans would discuss their choices with someone else. I will show through examples how this system would interact with a passenger and propose an initial approach to creating such a complex, integrated system. I call this system the Explainable Autonomous Vehicle Intelligence, or XAVI for short.

Autonomous vehicles are predicted [5] to outperform human drivers in several aspects. Amongst their many advantages, AVs may reduce the number of accidents, provide accessible car-travel to people with disabilities, and decrease emissions. Integrating AVs with transparent explainable AI will create a system that is simpler to comprehend and by extension more trustworthy [6]. By allowing for a conversational approach we can appeal to the social nature of people and achieve better knowledge-transfer that matches the expectations of users [7].

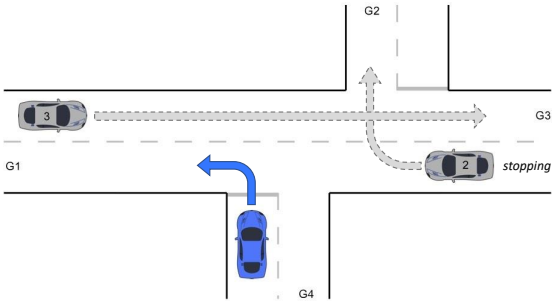

Fig. 1. Visual explanation showed by XAVI (in blue). Oncoming Vehicle 2 is stopping waiting for Vehicle 1 to pass and is predicted to turn right. Therefore, XAVI can turn left earlier using this time gap.

II. XAVI in Action

To illustrate why XAVI is powerful and well suited to build trust, imagine someone that uses a XAVI-powered mobility-on-demand service to commute to their workplace every day. During one morning trip three scenarios take place, that prompts interaction from the commuter.

At one time the vehicle breaks very suddenly even tough the passenger could not see anything in particular. After asking for a clarification, XAVI may show a recording of a small child chasing after a ball right in the direction of the car. The vehicle may also project an arrow extended towards the location of the averted collision to signal why breaking was necessary.

A little later, the car decides to take a route unfamiliar to the passenger, who asks for an explanation. The vehicle could reply, that based on traffic information it determined the chosen route to be the fastest. The commuter may interject saying they would have known a better route, but the vehicle could explain that it saw traffic diversions there just an hour ago which started to cause huge delays.

Finally, at a junction two other vehicles arrive as shown in Figure 1. XAVI chooses to take a left turn even though another vehicles was approaching on a priority lane from the right. The passenger could ask why XAVI thought this was a safe maneuver. XAVI may explain that the oncoming vehicle was likely trying to turn right and was giving way to the vehicle going straight, otherwise stopping on the road would be irrational for other goals. This gave XAVI enough time to turn left. The car may also display Figure 1 for clarification.

III. Why XAVI?

The example scenarios show explanations generated by XAVI, in which the passenger is reassured that the system is safe and fully in control of the situation by being able to react quickly and accurately to incidents, thereby contributing to a formation of trust [8]. Furthermore, the conversational style ensures that the commuter's doubts are freely expressed and readily addressed, as in the second scenario.

By being explainable, we allow our systems to be not only trustworthy but also accountable. This could mean attribution of responsibility in an incident is more easily determined [9]. Such attribution is not only an essential part of understanding accidents, but it also opens up our systems for scrutiny regarding its internal biases or unfairness [10]. Providing explanations that reveal information about these aspects is likely to boost social confidence in XAVI. In addition, transparency not only means explaining the decisions of XAVI to people, but it will enable people to provide more meaningful feedback to XAVI. The conversational approach not only helps users express their doubts or curiosity, but can also provide a feedback loop which we could use to optimise XAVI.

XAVI is conversational but explanations need not be in words. Various modes of explanations such as audio cues or visual imagery can enable a higher degree of fidelity and accessibility for everyone as explanations may be provided about any aspect of the journey. Besides being more accessible, this fusion of media can ensure that the optimal level of user understanding is reached during interactions.

Integration of explainable AI into AVs could resolve a range of issues around black-box models as well [11]. For example, if we used an interpretable model such as IGP2 [12], we could make system debugging easier, while performance evaluation, model comparison, and hyper-parameter search could become more straightforward. Verifiability of these models would mean that rigorous proofs could be given for a given decision as in the GRIT system [13], while the white-box nature means we can also reason about the extent of knowledge transfer to various unseen driving scenarios.

Furthermore, general acceptance of AVs through XAVI could entail other positive aspects derived from the efficiency of autonomous driving. AVs are predicted [5] to increase travel safety significantly through faster reaction times, thereby reducing road fatalities and in turn increasing overall trust. Additionally, AVs will enable people with disabilities to benefit from car-travel, which combined with the multimedia design of XAVI would create an accessible travel form.

IV. How to Build XAVI?

Building XAVI involves solving and integrating a wide range of tasks from a variety of distinct fields, such as motion planning and prediction, cognitive modelling, and natural language processing. To better understand and structure how such a system could look, I suggest that XAVI use three distinct modules for processing as depicted in Figure 2.

The AV module is responsible for the actual operation of the car. The global planner combines relevant map, traffic, and weather data to generate a route to a given destination according to some user specified criteria (e.g. shortest distance, lowest emissions, etc.). This task can be solved by a range of route-finding algorithms [14], which may also be extended to include recent memory of traffic experiences to optimise route search. Using plan-explanation methods we could also generate justifications for our selected routes [15]. Following the global route, the local planner generates shorter term action-plans that dynamically optimise the car's behaviour based on the immediate perceived and predicted state of the environment. Interpretability of these systems is a key factor, as they form intelligible structures our explanations can be based on. For example, the recent prediction and planning system IGP2 [12] would predict both the optimal and worst-case maneuver-sequences, which could enable the creation of contrastive explanations and more efficient driving behaviour. Finally, low-level controllers of the car would execute the commands of maneuvers, while collecting feedback-signals.

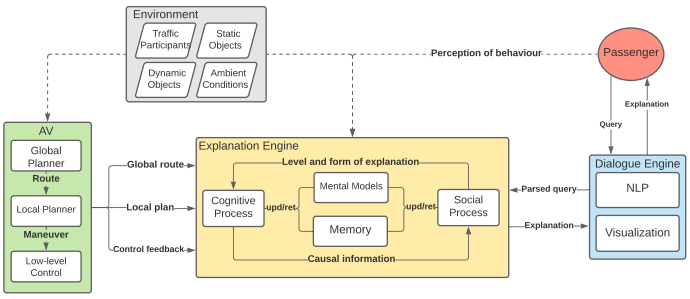

Fig. 2. Proposed XAVI system structure. Dashed lines mean perceptual inputs.

The primary module of XAVI is the explanation engine. Its main task is to synthesise information from other modules of XAVI and generate relevant explanations for the passengers. Following [16], its core is formed by an interaction-loop between the cognitive process that selects relevant causal information for the social process, which determines the kind of explanations required while also managing the incoming and outgoing communications with the passenger. To provide context-relevant and useful explanations, these processes update and retrieve information from a shared memory space. In addition, they may maintain and revise cognitive models e.g. based on criteria of explanation-seeking curiosity [17], which allow explanation selections that are more engaging and more in line with the expectations of passengers [18].

The direct communication with users is handled by the dialogue engine. This module parses the incoming queries of the passengers, and generates explanations based on the commands of the social process. Natural language interactions may be managed through semantics-oriented dialogue modelling [19], [20] where semantic information is extracted by the explanation engine from causal information. This information may then be displayed on screen or converted to sound for audio cues. Additional visual outputs could combine a range of data, such as simple displays of perceptual recordings or dynamically generated images like Figure 1.

V. Regulations and Ethical Concerns

XAVI fits into a broader trend of information management regulations and movements, such as the EU's GDPR [21] or the US Algorithmic Accountability Act of 2019 [22]. XAVI by design is capable of fulfilling legal obligations stemming from "the right to explanation", which could make it attractive for existing companies to adopt. Furthermore, full transparency of XAVI means not just explanations on-the-spot, but also complete access and management of the collected and processed data of passengers. This data may then be integrated into an event data recorder infrastructure, such as the EU's *eCall* system [23] to provide timely help to passengers and relevant explanations to the authorities in case of an incident.

An important consideration around XAVI is how we handle failure cases. While these cases may negatively affect trust [3], however a combination of explanations and an expression of regret was shown [3] to be able to sufficiently recover trust.

Ultimately, XAVI must also rely on some form of collected data, such as voice recordings or perceptual inputs to make the best predictions possible. A prevailing issue with data is the inherent and latent bias encoded in it [10]. This could cause distrust in the applicability of AVs in rare or novel scenarios, and would make the deployment of XAVI in various parts of the world a bigger challenge [2]. However, an advantage of XAVI is the potential ability to explain decisions in terms of social expectations, which could immediately shed light on systematic biases and prompt designers for corrections.

VI. Summary

In this article, I have outlined my vision for a trustworthy and transparent, fully autonomous self-driving system called Explainable Autonomous Vehicle Intelligence, which provides clear and relevant explanations using conversations with the aim to achieve general acceptance for AVs. The subsequent deployment of AVs brings with it a range of advantages, such as decreasing number of accidents, better traffic management, reduced emissions, and decreased travel times.

I argued that by building interpretable and intelligible systems such as XAVI, we can boost people's trust in AVs and therefore seek to achieve general acceptance for them. XAVI also naturally aligns with modern privacy regulations, as it guarantees the oversight of private data processing in a fully transparent way, further increasing user trust. Passengers will be able to interrogate every part of the car's systems. This should propagate accurate knowledge about the workings of these vehicles in society, that will help solidify public trust. To show the feasibility of my vision, I proposed an initial approach to building XAVI using a modular-approach with modules grounded in existing research.

References

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

[2] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "Explanations in autonomous driving: A survey," *ArXiv*, vol. abs/2103.05154, 2021.

[3] E. S. Kox, J. H. Kerstholt, T. F. Hueting, and P. W. de Vries, "Trust repair in human-agent teams: the effectiveness of explanations and expressing regret," *Autonomous Agents and Multi-Agent Systems*, vol. 35, no. 2, p. 30, Jun. 2021.

[4] N. Adnan, S. Md Nordin, M. A. bin Bahruddin, and M. Ali, "How trust can drive forward the user acceptance to the technology? in-vehicle technology for autonomous vehicle," *Transportation Research Part A: Policy and Practice*, vol. 118, pp. 819–836, 2018.

[5] D. J. Fagnant and K. Kockelman, "Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations," *Transportation Research Part A: Policy and Practice*, vol. 77, pp. 167–181, 2015.

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

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

[8] P. Launonen, A. O. Salonen, and H. Liimatainen, "Icy roads and urban environments. passenger experiences in autonomous vehicles in finland," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 80, pp. 34–48, 2021.

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

[10] D. Danks and A. J. London, "Algorithmic bias in autonomous systems." in *IJCAI*, vol. 17, 2017, pp. 4691–4697.

[11] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, Aug. 2018.

[12] S. V. Albrecht, C. Brewitt, J. Wilhelm, B. Gyevnar, F. Eiras, M. Dobre, and S. Ramamoorthy, "Interpretable goal-based prediction and planning for autonomous driving," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[13] C. Brewitt, B. Gyevnar, S. Garcin, and S. V. Albrecht, "GRIT: fast, interpretable, and verifiable goal recognition with learned decision trees for autonomous driving," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[14] H. "Bast, D. Delling, A. Goldberg, M. Müller-Hannemann, T. Pajor, P. Sanders, D. Wagner, R. F. Werneck, and P. Sanders, *Route Planning in Transportation Networks*. Cham: Springer International Publishing, 2016, pp. 19–80.

[15] M. Fox, D. Long, and D. Magazzeni, "Explainable planning," 2017.

[16] R. Dazeley, P. Vamplew, C. Foale, C. Young, S. Aryal, and F. Cruz, "Levels of explainable artificial intelligence for human-aligned conversational explanations," *Artificial Intelligence*, vol. 299, p. 103525, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000437022100076X

[17] E. G. Liquin and T. Lombrozo, "A functional approach to explanation-seeking curiosity," *Cognitive Psychology*, vol. 119, p. 101276, 2020.

[18] M. Westberg, A. Zelvelder, and A. Najjar, "A Historical Perspective on Cognitive Science and Its Influence on XAI Research," in *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 205–219.

[19] X. Bai, Y. Chen, L. Song, and Y. Zhang, "Semantic representation for dialogue modeling," 2021.

[20] G. Wirsching, M. Huber, C. Kölbl, R. Lorenz, and R. Römer, "Semantic dialogue modeling," in *Cognitive Behavioural Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 104–113.

[21] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[22] 116th Congress (2019-2020), "Algorithmic accountability act of 2019," *H.R.2231*, 2019.

[23] K. Pinter, Z. Szalay, and G. Vida, "Road Accident Reconstruction Using On-board Data, Especially Focusing on the Applicability in Case of Autonomous Vehicles," *Periodica Polytechnica Transportation Engineering*, vol. 49, no. 2, pp. 139–145, 2021, number: 2.