2023

INVESTIGATING COLLABORATIVE EXPLAINABLE AI (CXAI)/SOCIAL FORUM AS AN EXPLAINABLE AI (XAI) METHOD IN AUTONOMOUS DRIVING (AD)

Tauseef Ibne Mamun
Michigan Technological University, tmamun@mtu.edu

INVESTIGATING COLLABORATIVE EXPLAINABLE AI (CXAI)/SOCIAL FORUM AS AN EXPLAINABLE AI (XAI) METHOD IN AUTONOMOUS DRIVING (AD)

By

Tauseef Ibne Mamun

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Applied Cognitive Science and Human Factors

MICHIGAN TECHNOLOGICAL UNIVERSITY

2023

This dissertation has been approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY in Applied Cognitive Science and Human Factors.

Department of Cognitive and Learning Sciences

Dissertation Advisor: *Shane T. Mueller*

Committee Member: *Kelly S. Steelman*

Committee Member: *Erich Petushek*

Committee Member: *Nathan L. Tenhundfeld*

Department Chair: *Kelly S. Steelman*

Table of Contents

List of Figures

List of Tables

Author Contribution Statement

Three chapters have been published/in press. The chapters are Chapter 3 - Linja, A., Mamun, T. I., & Mueller, S. T. (2022). When self-driving fails: Evaluating social media posts regarding problems and misconceptions about tesla's fsd mode. *Multimodal Technologies and Interaction*, *6*(10), 86. and Chapters 4 and 5 - Mamun, T. I., & Mueller, S. T. (2023, October). The Use of Social Forums to Train Users About Shortcomings of Tesla Full Self-Driving (FSD). In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (p. 21695067231193644). Sage CA: Los Angeles, CA: SAGE Publications.

Definitions

1. **Social QA/Social Forums/SQA:** A social forum refers to an online platform or portal that facilitates the convergence of individuals from diverse backgrounds and interests, enabling them to actively participate in dialogues, disseminate knowledge, and interchange thoughts on a broad spectrum of subjects or a single subject. The Collaborative Explainable AI (CXAI) system (Mamun, Hoffman, et al., 2021) resembles a social forum akin to Stack Exchange or Stack Overflow.

List of Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
CXAI	Collaborative Explainable AI
EI	Emotional Intelligence
FAQ	Frequently Asked Questions
XAI	Explainable AI
AD	Autonomous Driving
SQA	Social QA/Social Forums/ Social Question and Answer Platform
AV	Autonomous Vehicle
ICAP	Active Learning: Interactive, Constructive, Active, and Passive
SA	Situational Awareness
AP	Auto Pilot
FSD	Full Self-Driving
NADS	National Advanced Driving Simulator
NASA-TLX	Cognitive Workload Questionnaire
PPQ	Perceived Personalization Questionnaire

Abstract

Explainable AI (XAI) systems primarily focus on algorithms, integrating additional information into AI decisions and classifications to enhance user or developer comprehension of the system's behavior. These systems often incorporate untested concepts of explainability, lacking grounding in the cognitive and educational psychology literature (S. T. Mueller et al., 2021). Consequently, their effectiveness may be limited, as they may address problems that real users don't encounter or provide information that users do not seek.

In contrast, an alternative approach called Collaborative XAI (CXAI), as proposed by S. Mueller et al (2021), emphasizes generating explanations without relying solely on algorithms. CXAI centers on enabling users to ask questions and share explanations based on their knowledge and experience to facilitate others' understanding of AI systems. Mamun, Hoffman, et al. (2021) developed a CXAI system akin to a Social Question and Answer (SQA) platform (S. Oh, 2018a), adapting it for AI system explanations. The system successfully passed evaluation based on XAI metrics Hoffman, Mueller, et al. (2018), as implemented in a master's thesis by Mamun (2021), which validated its effectiveness in a basic image classification domain and explored the types of explanations it generated.

This Ph.D. dissertation builds upon this prior work, aiming to apply it in a novel context: users and potential users of self-driving semi-autonomous vehicles. This approach seeks to unravel communication patterns within a social QA platform (S. Oh, 2018a), the types of questions it can assist with, and the benefits it might offer users of widely adopted AI systems.

Initially, the feasibility of using existing social QA platforms as explanatory tools for an existing AI system was investigated. The study found that users on these platforms collaboratively assist one another in problem-solving, with many resolutions being reached (Linja et al., 2022). An intriguing discovery was that anger directed at the AI system drove increased engagement on the platform.

The subsequent phase leverages observations from social QA platforms in the autonomous driving (AD) sector to gain insights into an AI system within a vehicle. The dissertation includes two simulation studies employing these observations as training materials. The studies explore users' Level 3 Situational Awareness (Endsley, 1995) when the autonomous vehicle exhibits abnormal behavior. These investigate detection rates and users' comprehension of abnormal driving situations. Additionally, these studies measure the perception of personalization within the context of the training process (Zhang & Curley, 2018), cognitive workload (Hart & Staveland, 1988), trust, and reliance (Körber, 2018) concerning the training process. The findings from these studies are mixed, showing higher detection rates of abnormal driving with training but diminished trust and reliance.

The final study engages current Tesla FSD users in semi-structured interviews (Crandall et al., 2006) to explore their use of social QA platforms, their knowledge sources during the training phase, and their search for answers to abnormal driving scenarios. The results reveal extensive collaboration through social forums and group discussions, shedding light on differences in trust and reliance within this domain.

1 Introduction

The emergence of autonomous driving (AV) has initiated a novel phase in transportation, with the potential for enhanced safety and efficiency on roadways, decreased traffic congestion, and improved accessibility for those with diverse abilities. At the core of achieving this vision are autonomous vehicle systems empowered by Artificial Intelligence (AI) technology. Nevertheless, as the complexity of these AI systems continues to grow and their integration into our everyday routines becomes more prevalent, concerns regarding transparency, accountability, and safety come to the forefront. The comprehension of the mechanisms and rationales behind the decision-making process of an AI system holds significant importance. This pertains not only to the advancement and governance of self-driving vehicles but also to the establishment of public confidence and reliance.

The concept of Explainable AI (XAI) - Hoffman et al. (2018) has become a significant focal point within the realm of autonomous driving. The inclusion of the idea of "explainability" is deemed essential and anticipated as it serves a crucial function in augmenting the level of transparency in the decision-making process of AI models. In less complex AI applications, such as the diagnosis of health issues through symptom analysis, the attainment of explainability is rather uncomplicated. Nevertheless, with the increasing need to attain levels of accuracy comparable to human capabilities and the growing integration of AI into different facets of our everyday existence, such as driving, we are confronted with complex situations that necessitate AI systems to make timely and accurate judgments, also explaining its decision to users.

There is an expectation that AVs should possess the capability to provide explanations for their *observations*, *actions*, and *potential future actions* within the areas in which they are deployed (Omeiza et al., 2021). In the explanation realm of AV, these explanations are given through the utilization of neural networks and deep learning techniques (Fujiyoshi et al., 2019; Zablocki et al., 2022). However, these decision-making applications frequently result in findings that are complex and less readily interpretable (S. T. Mueller et al., 2021). In alternative terms, the utilization of these sophisticated technologies leads to an increased complexity in comprehending the rationale and outcomes of judgments guided by AI. Moreover, Zablocki et al. (2022) argue that the need for explainability is complex and depends on various factors, including the person seeking explanations, their level of competence, and the available time for examining the explanation for explanation generation techniques using neural networks and deep learning. Given that autonomous vehicles are still in their early stages of commercialization and the technology is relatively nascent, the individuals operating these new technologies naturally lack experience. In high-pressure, split-second decision-making situations, there are doubts about the amount of time these drivers will have to analyze explanations while driving.

As observed, the present explanations are challenging to comprehend, and the feasibility of comprehending these 'hard-to-grasp' explanations within a short timeframe is uncertain. The primary objective of this dissertation is to introduce an innovative explanatory framework for autonomous driving and evaluate its efficacy using a range of experimental methodologies. This system aims to offer pre-drive instruction to users of autonomous vehicles, providing them with knowledge about *potential future actions* based on

collaborative insights obtained through specialized social forums dedicated to the AI system. It is important to acknowledge that automation is frequently distinguished by its dependence on pre-established norms and obedience to planned instructions. On the other hand, artificial intelligence, commonly referred to as AI is an extension of automation, exhibits the capacity to learn information from data and employ that knowledge to make well-informed choices. A clear differentiation between the two entities; the concept of automation revolves around the notion of purpose, while artificial intelligence (AI) centers on purpose and algorithmic processes.

In the current dissertation, the terms were utilized interchangeably, however, in both instances they referred to artificial intelligence (AI).

2 Literature Review

The literature review will summarize some research that forms the precursor and precedent for explaining AI systems, especially in autonomous driving. The review will try to establish a non-algorithmic approach (S. T. Mueller et al., 2021) to train and explain novice users of AI systems for autonomous driving. This training falls into the category of pre-explanation; one's reporting of mainly the 'what' or 'why' of an incident regarding the AI to help others to understand the AI.

Autonomous vehicles have to make quick decisions based on how they classify the objects in the scene in front of them. If an autonomous vehicle acts abnormally because of some misclassification problem, the consequence can be dangerous. Recently, a self-driving Uber killed a woman in Arizona, USA. It was the first known fatality involving a fully autonomous vehicle (Adadi & Berrada, 2018). The information reported by anonymous sources claimed that the vehicle's sensors did detect the pedestrian before the collision. However, the car's software reportedly failed to accurately classify the pedestrian as an object that required immediate evasive action. It treated it in the same way it would a plastic bag or tumbleweed carried on the wind (McFarland, 2018).

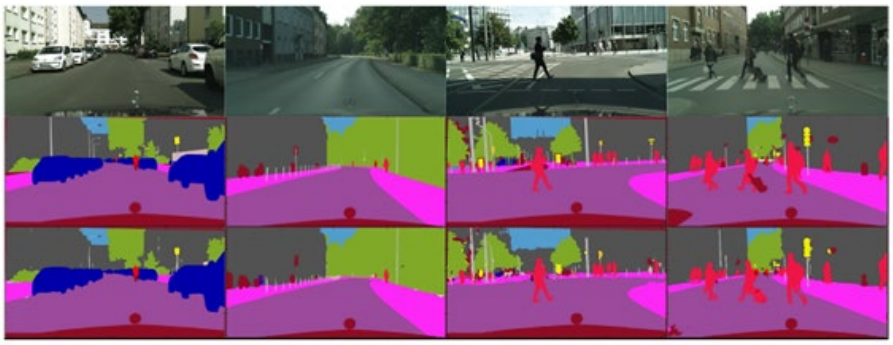
Autonomous vehicles also on the road generate unconventional driving scenarios for drivers in non-autonomous vehicles. Dixit et al. (2016) summarized the autonomous vehicle accidents from September 2014 to February 2016. Despite the progress made in autonomous vehicle technology, the report's timeline might make it seem outdated. However, autonomous vehicles are still in the initial phases of being introduced for commercial use and widespread adoption. The challenges outlined in this report may

continue to affect autonomous commercial vehicles as users of these vehicles might encounter a distinct array of issues. In that report, in most of the cases where the non-autonomous vehicle was determined to be at fault, the underlying cause tells that the driver of the other vehicle expects the autonomous vehicle to behave differently from what they would have normally expected. So, the roots of the problems regarding human-AI interaction in autonomous driving can be connected to imperfect AI and AI's abnormal nature seen from the outside. The latter problem will be investigated within the scope of this report, considering both the perspective of the autonomous vehicle operator and whether they continue to use AI despite AI's non-human traits.

Crafting an effective explanation of AI within an autonomous vehicle context can present challenges due to the unconventional nature of the scenario. Unlike traditional scenarios where explanations are provided after an incident, the complexity of autonomous driving means that waiting until a dangerous event occurs might not be feasible. Moreover, the limitations of AI impact not only the occupants of the self-driving vehicle but also pedestrians and other road users, creating a dual interaction between AI and different categories of humans.

Presently, the majority of explanations in autonomous driving situations fall under the category of post-explanations, usually delivered to the human driver after an event. This is due to the reactive nature of explaining AI decisions, which are based on actions that have already occurred. As the development of autonomous technology continues, finding ways to communicate AI behaviors and shortcomings before an event becomes increasingly important to ensure the safety and understanding of all stakeholders involved in

autonomous driving scenarios. Figure 1 shows post-explanation given through video and textual means, in one case highlighting pedestrians in the intersection through heat maps and in the other case, texts showing why an action was taken (e.g., continue next exit – must be in the roundabout and not in outer-lane) in the route. Considering the incident in Arizona involving Uber's self-driving vehicle, it's crucial to emphasize that real-time explanations of this sort could potentially lead drivers to be misled owing to misclassification. In such circumstances, the driver's explanation may differ from what the This can be created using the contents of social forum Pre-explanation can be helpful for



Macro action:	Additional applicability condition:	Maneuver sequence (maneuver parameters in brackets):
Continue	—	*lane-follow* (end of visible lane)
Continue next exit	Must be in roundabout and not in outer-lane	*lane-follow* (next exit point)
Change left/right	There is a lane to the left/right	*lane-follow* (until target lane clear), *lane-change-left/right*
Exit left/right	Exit point on same lane ahead of car and in correct direction	*lane-follow* (exit point), *give-way* (relevant lanes), *turn-left/right*
Stop	There is a stopping goal ahead of the car on the current lane	*lane-follow* (close to stopping point), *stop*

TABLE I: Macro actions used in our system. Each macro action concatenates one or more maneuvers and automatically sets their parameters.

Figure 1. Top: Visual explanation in AD (Hofmarcher et al., 2019); Bottom: Textual explanation in AD (Albrecht et al., 2021).

the large community of autonomous vehicles to overcome safety issues and understand shortcomings of the AI; reporting on these cases can be stored in SQA sites (S. Oh, 2018a) like the CXAI system (Mamun, Hoffman, et al., 2021), then humans in an autonomous vehicle would have a chance to understand abnormal driving conditions beforehand and

takeover. This can also be a source of knowledge for drivers not operating autonomous vehicles.

To establish CXAI or SQA as an explanation medium in autonomous driving, the literature review will mainly look into the current methods for explanation in autonomous driving and what makes an explanation in autonomous driving. The concept map in Figure 2 shows how the MS literature (Mamun, 2021) is connected to the Ph.D. literature and overall Ph.D. literature.

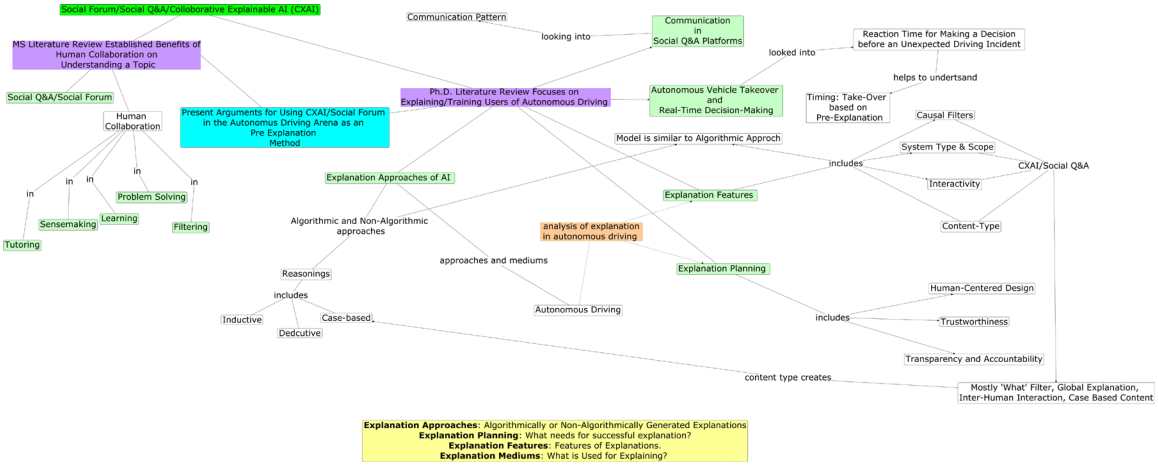


Figure 2. Concept Map for the Ph.D. Literature Review (For a better view: https://tinyurl.com/mwu6jsth)

The concept map illustrates a structure for assessing a non-algorithmic explanation approach (S. Mueller et al., 2021) within the context of autonomous driving. This evaluation will initially delve into both algorithmic (Das & Rad, 2020) and non-algorithmic (S. Mueller et al., 2021) methodologies within explainable AI (XAI) systems, and it will identify the prevailing approach in the realm of explainable AI for autonomous driving. Subsequently, the analysis will delve deeply into the outcomes stemming from XAI in the context of autonomous driving, as well as explore the various dimensions of explanations.

These sections of the literature review will present arguments that align with the integration of CXAI or SQA as a preliminary explanatory platform for XAI in autonomous driving scenarios. Following this, the literature review will examine the potential duration of drivers' responses to unconventional driving situations in terms of situational awareness. This exploration aims to provide insights into the appropriate timing for drivers to take action in an abnormal driving situation based on acquired pre-explanations and their awareness of the situation, a facet that can be empirically tested within this Ph.D. dissertation. Lastly, the review will scrutinize the communication aspect within the SQA framework to ascertain its support in the formation of explanations. The review will try to answer 3 ‘WHAT’ questions regarding XAI in AD.

1. ‘WHAT’ are the explanations?
2. ‘WHAT’ is the timing for effective communication (to explain) between Human-AI before an event?
3. ‘WHAT’ are the aspects of communication in an SQA platform?

2.1 Explanation Approaches of AI

This literature review will first discuss the XAI approaches (way to generate explanations) that provide explanations using different explanation mediums (platforms like visual, textual, etc.) for AI systems. Investigating the approaches will help to establish each approach’s pros and cons in the XAI of autonomous driving. The choice of mediums is determined by the explanatory approach, which encompasses algorithmic and non-algorithmic methodologies (S. T. Mueller et al., 2021). Lipton (2016) states that algorithmic approaches are common that describe how a model behaves and why usually

including verbal (natural language) explanations, visualizations, or explanations by example. Based on this, the algorithmic approach refers to a process for generating explanations by investigating the underlying workings of an AI model or system. The goal is to uncover the patterns, linkages, and computations that occur during the AI's decision-making process, and to provide insights into how the AI arrives at its conclusions. Techniques such as feature importance analysis, visualization of model internals, and sensitivity analysis are frequently used in algorithmic explanations to highlight the model's behavior and decision drivers. This method seeks to improve transparency and interpretability by exposing the "black box" nature of AI models and making their decision-making more human-comprehensible. So, model-based explanations can be condition-dependent and give an explanation based on action. Dependency on model-based explanatory systems depends on AI's architecture for any change (Das & Rad, 2020). So, this is likely to make the exploratory system rigid and may not keep up with new AI versions.

S. Mueller et al. (2021) propose a non-algorithmic approach to explanations to bypass this caveat. One of the options for a non-algorithmic approach is Collaborative XAI or CXAI (Mamun, Hoffman, et al., 2021). This option helps users explain things and help one another, providing global explanations about a system for novice users. The concept entails that the explanation will align with the iterations of the AI. This option can also be used as experiential training in the form of a user guide about the cognitive operations of AI (S. T. Mueller, 2009). The CXAI option is also connected with another non-algorithmic approach: Cognitive Tutorial. The Cognitive Tutorial recognizes that users will come to the AI with misconceptions about how it works–often assuming it works in the same way

a human would (S. T. Mueller, 2020). However, an AI may succeed or fail in unexpected ways. The goal of this tutorial is to use experiential training to help the user understand the competence boundaries of the system–along dimensions that include modeling/representation, algorithms, data, and output/visualization. The downside of the non-algorithmic approach is that without a few experts in AI, some complicated AI aspects may not be answered by only novices. Therefore, this dissertation will investigate whether mutually similar inexperienced users can assist one another in comprehending the actions of autonomous vehicles by utilizing the CXAI option within the non-algorithmic framework.

In the current XAI systems, visualization is the most frequently applied medium. Explanation medium can be a combination of more than one medium; the user could engage with an agent through a dialog system by interacting with visualization and stating questions in natural language to understand the decisions made by the model (Sevastjanova et al., 2018). In storytelling, a combination of text and visual elements like infographics are used in diverse formats to communicate the data effectively (J. Wang et al., 2019). AI researchers or developers primarily design this approach with their advanced understanding of the AI but lay users often demand AI explanations that are easily understandable to them (Liao et al., 2020). This creates a problem of non-user-centric explanations in these mediums (Mamun, Baker, et al., 2021). The CXAI option involves a blend of textual and visual mediums, incorporating content contributed by the lay users of the AI system.

The next subsection will look into the reasoning used in the approaches for the explanation as the CXAI lacks reasoning traces (Lim et al., 2009), according to Mamun, Baker, et al. (2021). So, the determination of the presence of reasoning in other ways must be examined.

2.1.1 Explanation Reasonings in the Approaches

Most explanations have reasoning traces (Lim et al., 2009), and common reasoning can be inductive and deductive (Sternberg & Sternberg, 2018). These two types of reasonings have been used in the past for the algorithmic approach (M. I. Alam et al., 2021; Hayes et al., 2010; Johnson-Laird, 1999; Kemp & Tenenbaum, 2009; Klauer & Phye, 2008; Rips, 1994). The inductive (bottom-up) strategy first explains minor and observable details, followed by complex relations. Deductive strategy (top-down) starts with the whole picture as an overview, and then more details get added and explained to show a complete view.

Another reasoning type is case-based reasoning. In a case-based system, solutions for problems are derived by retrieving the most similar cases from its memory and adapting them to fit the given problem (Doyle et al., 2003). The significant advantage of the case-based system comes from user acceptance. In a rule-based AI system (van der Waa et al., 2021), the system must explain its decisions using rules that the user may not fully understand or accept (Riesbeck, 1988). With a case-based system, an explanation can be given by simply presenting an actual prior case to the user as support for the prediction of the system (Doyle et al., 2003). There are many ways to retrieve cases for reasoning or explanation. One way is to form a complete problem description and use this to select a relevant case (Doyle et al., 2003). In the incremental approach, more discriminating features are used to discriminate between cases in a case base (Cunningham et al., 1995,

1998; Owens, 1992). However, the most common retrieval technique is to look for near-neighbor cases (Watson, 1998). Once the similarity between the target case and all of the stored cases has been calculated, the most similar cases are retrieved, i.e., the cases with the highest similarity value (Doyle et al., 2003). CARES (Ong et al., 1997) uses this matching process.

In some cases, a situation where the nearest neighbor might not be the best case to support an explanation arises when the nearest neighbor is further from the decision boundary than the target case. For example, if a decision is being made on whether to keep a sick 12-week-old baby in the hospital for observation, a similar example with a 14-week-old baby that was kept is more compelling than one with an 11-week-old baby - based on the notion that younger babies are more likely to be kept in (Doyle et al., 2004). Doyle et al. (2004) used explanation utility, a metric different from the similarity metric used for nearest neighbor retrieval. In the decision surface, if the nearest neighbor is on the wrong side of the query case relative to the decision boundary, the utility measure chooses/retrieves another case as an explanation that lies between the query case and the decision boundary. (In such cases, the query case is closer to the decision boundary than the nearest neighbor). Though case-based systems used the algorithmic approach till now, CXAI (a non-algorithmic approach) is based on case-based reasoning that shows similar cases based on keywords or topics/triggers (Mamun, Hoffman, et al., 2021).

2.1.2 Explanation Medium in Autonomous Driving

Until now, the algorithmic approach (see Figure 1) using different mediums has been used to explain autonomous driving (Atakishiyev et al., 2021). The most common explanation

medium is visual (Bojarski et al., 2016; Hofmarcher et al., 2019; J. Kim & Canny, 2017; Suchan et al., 2019; C. Wang et al., 2021). Very few processes used textual (Albrecht et al., 2021; Corso & Kochenderfer, 2020; Kothawade et al., 2021) and a combination of visual and textual (Brewitt et al., 2021; J. Kim et al., 2021). However, the algorithmic process primarily educates autonomous vehicle developers rather than road users (Atakishiyev et al., 2021). The techniques employed in the process predominantly revolve around machine learning and deep learning.

A novel non-algorithmics approach focusing on involving users in explanations specific to their needs could be investigated for road users. A potential application of the non-algorithmic approach involves establishing a combination of textual and visual medium's repository of the autonomous vehicle's vulnerabilities that might not be readily available from the vehicle's manufacturer. This can be created using the contents of social forums. This knowledge base can serve as an educational resource for inexperienced users, enlightening them about takeover scenarios and enhancing their understanding.

Subsequent sections will delve into the factors that can yield dichotomous outcomes within autonomous driving and the features of the varying dimensions that algorithmic and non-algorithmic approaches employ when elucidating AI concepts to users.

2.2 Explanation Planning for Autonomous Driving

The provision of explanations for AI within the context of autonomous driving offers substantial advantages to its users. Given the relative novelty of this technology and the fact that even the most advanced autonomous vehicles, such as those from Tesla, have achieved only Level 2 autonomy, the establishment of explainability becomes crucial. This

move towards explainable autonomous driving has the potential to yield benefits that address concerns about the acceptance of new technologies.

Atakishiyev et al. (2021) and Omeiza et al. (2021) identified and elucidated three critical pillars—Trustworthiness, Human-Centered Design, and Transparency and Accountability—that collectively form the bedrock of a robust and proficient XAI Planning system in the domain of autonomous driving. These studies emphasize the critical importance of imbuing autonomous driving systems with the ability not only to make decisions but also to provide understandable explanations for those decisions, fostering not only trust and acceptance but also advancing safety and accountability in this cutting-edge technological landscape. These elements play a decisive role in shaping the outcomes of human-AI collaboration in autonomous driving scenarios. In the approaches of XAI, the integration of these three elements becomes essential. These factors carry significant weight in preventing potential consequences like the widespread rejection of AI technologies. This section aims to ascertain how the CXAI system implements these three elements within its approach.

2.2.1 Trustworthiness

Autonomous driving represents a relatively recent technological advancement. To ensure both the safety of the driver and bystanders, the driver must possess a comprehensive understanding of the vehicle's capabilities. This knowledge forms the basis for cultivating trust, a dynamic process, transitioning from the initial establishment of trust to its continuous reinforcement (W. Wang & Siau, 2018).

During the phase of initial trust-building, drivers must receive adequate training on the system and have access to an XAI system that aids in the ongoing cultivation of trust. The accuracy of AI models in critical scenarios holds paramount importance, as any biases or inaccuracies could potentially lead to the rejection of these models (L. Alam, 2020; S. T. Mueller et al., 2021). When trust in an AI system is compromised, restoring it becomes a challenging endeavor.

It's important to note that in cases where an AI system provides incorrect guidance, such as in the realm of Fintech AI, the repercussions can be financially significant for users. This might prompt users to discontinue using the service, despite its prior instances of overall success. This decision comes even though the AI could potentially enhance its performance by learning from previous failures (T. Kim & Song, 2021).

The establishment of trust can be facilitated through explanations. The empirical evidence presented by Holliday et al. (2016) demonstrates that supplying explanations and creating perceptible systems markedly enhances users' trust in AI systems. Within the algorithmic approach, Israelsen & Ahmed (2019) underscore the requirement for AI assurance to cultivate trust in human-autonomous systems. In a study involving 552 drivers, Choi & Ji (2015) identified three factors—system transparency, technical competence, and situation management—that positively influence the development of trust. While the study originally pertained to the adoption of an autonomous vehicle, the focus of the report will shift toward examining the adoption of an XAI system within the context of autonomous vehicles. Consider the scenario of Uber's self-driving vehicle incident. If a visual algorithmic explanatory medium, similar to the one depicted in Figure 1, had been in place, it might have effectively isolated pedestrians from the overall environment and

communicated this information to the driver via a heatmap visualization. However, crucially, this medium fell short of indicating the subsequent action that the vehicle would take. This lack of alignment between the explanation provided and the subsequent action created a notable disparity. It failed to achieve transparency and technical competence in its performance. By not apprising the driver of the unfolding situation, the system demonstrated an inability to effectively manage the scenario, thus compromising its situation management capability.

In the context of a non-algorithmic approach, such as CXAI, it's possible to incorporate these same three factors highlighted by Choi & Ji (2015) for addressing the aforementioned vehicle incident. The intent behind utilizing CXAI is to educate drivers through pre-explanations before embarking on any drive. This approach aims to equip drivers with the ability to identify and respond to abnormal driving situations effectively. Within a CXAI framework, system transparency is achieved through the presentation of counterarguments or shared experiences related to a particular notion. For instance, Linja et al. (2022) demonstrated this phenomenon by analyzing data from Tesla's SQA sites. Users engaged in discussions and shared experiences about instances where the vehicle failed to adhere to legal driving practices. An illustrative case involved Tesla users discussing the vehicle's inability to recognize birds on the road, with other users corroborating and adding information about other animals the vehicle struggled to identify. This collaborative process serves the dual purpose of making users aware of potentially hazardous scenarios and facilitating an understanding of situations that might lead to fatal outcomes, all before embarking on a drive.

It's noteworthy that, within the context of CXAI, the concept of technical competence diverges from its conventional interpretation. Here, technical competence pertains to optimizing the system's usability and making it user-friendly, rather than focusing on intricate technical capabilities.

The CXAI system devised by (Mamun, Hoffman, et al., 2021) demonstrated that users engaging with this XAI system could indeed place trust in the explanations it generated (Ibne Mamun et al., 2022). Therefore, a platform resembling SQA can garner users' trust if employed as an explanatory system.

2.2.2 Human-Centered Design

Several studies have been conducted in the past that use human-centered XAI design in the forms of visual, audio, and textual information to transmit the instant decisions of a vehicle to the humans in the vehicle (Gang et al., 2018; Koo et al., 2015; Walch et al., 2015). To evaluate the XAI system to make it human-centric, Förster et al. (2020) propose a circular framework that contains nine steps under three phases:

A. Instantiation (the phase devoted to understanding the XAI system’s application domain and the target users as well as defining requirements for the XAI system),
B. Calibration (calibrating the XAI system according to iteratively refined requirements), and
C. Quality Control (devoted to continuous evaluation of the deployed XAI system under real-world conditions).

Human reasoning can also be used to develop human-centric XAI systems (D. Wang et al., 2019). Chromik & Butz (2021) proposes three design principles that comprise using the natural language, follow-ups on initial explanations, and the opportunity for offering multiple explanation methods and modalities to enable users to triangulate insights. Subsequent interactions that build upon initial explanations can be closely linked to the concept of continuous trust-building.

Drawing upon the design principles outlined by Chromik & Butz (2021), the utilization of familiar SQA platforms such as StackOverflow and StackExchange can be perceived as human-centric as they are under continuous evaluation under real-world conditions. The development of a CXAI system, modeled after these platforms, is guided by human-centric design principles, including heuristic evaluation and the think-aloud protocol (S. T. Mueller et al., n.d.).

Explanations generated via the StackOverflow/StackExchange style of XAI platforms demonstrate a notable degree of readability in natural language (Mamun, Baker, et al., 2021). These explanations are purposefully constructed to converge user insights, a process facilitated by user comments within an observational context (see Figure 4).

2.2.3 Transparency and Accountability

Martinho et al. (2021) describe accountability as a combination of liability and responsibility. Ehsan et al. (2021) put forward 'Social Transparency' (ST) for algorithmic-centric XAI systems. This perspective, integrates the socio-organizational context, taking into careful account of the social, organizational, and cultural factors that can influence the utilization of AI. In particular, Ehsan et al. (2021) identified social interaction that enhance

effective online communication and collaboration. As SQA sites are naturally open platforms and collaboration in this platforms is shown by Linja et al. (2022); what users post is transparent to others, and users know they are accountable for their posts as they may get challenged by others if they post something wrong or factually correct (Linja et al., 2022) – see Figure 3. One of the challenges for a SQA platform is its vast volume of different types of data, and the users are self-accountable for what they use from these mediums. Harper et al. (2009) showed that humans could reliably distinguish between conversational and informational questions in a SQA platform.

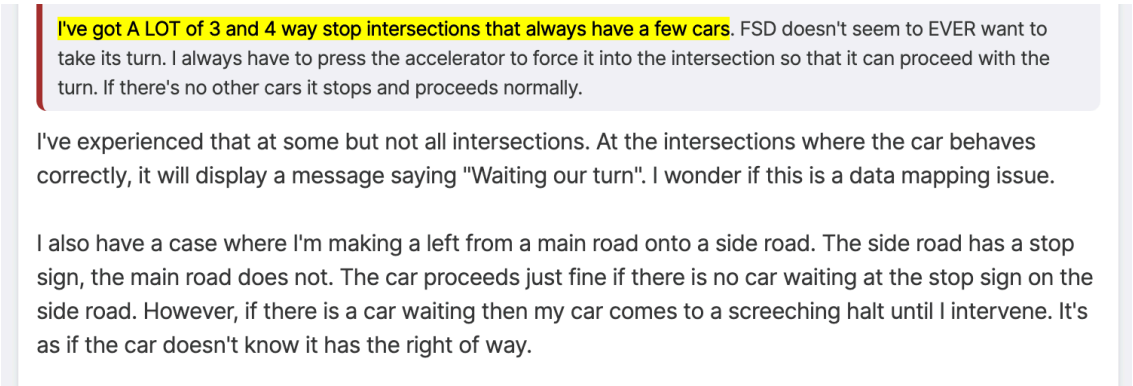


Figure 3. A Counterargument to an Initial Notion

The discussion shows that SQA/CXAI sites can be trustworthy, transparent, and designed for humans to understand a topic or situation collaboratively. This establishes that the use of the CXAI/SQA site will produce positive consequences for the AI that will lead to lesser rejection of the AI.

2.3 Explanation Features in Autonomous Driving

In the previous sections, explanation mediums and approaches have been discussed. Explanation mediums can be mainly visual or textual, or a combination of both. Explanation approaches (algorithmic or non-algorithmic) use these mediums to explain to users about AI. Mediums can contain one or more features of explanations. Omeiza et al. (2021) called these feature dimensions and proposed six dimensions in the context of autonomous driving: Causal Filters, Content-Type, Model, System Type, Scope, and Interactivity. The 'Model' dimension is the same as the algorithmic approach, but any approach can have one or more features of explanations.

2.3.1 Causal Filters

Omeiza et al. (2021) discussed causal filters as effective dimensions of explanation. By causal filters, they were referring to questions that Lim & Dey (2009) proposed - "why," "why not," "how to," and "what if", and 'what' that are used to explain the outcome of an event. The first four questions are reasoning traces (Lim et al., 2009) that have some reasons behind an explanation. This type of explanation can answer the questions like 'why a car is going in the middle lane?'.

These filters are related to the so-called explanation triggers identified by (Mamun, Baker, et al., 2021; S. T. Mueller et al., 2021). But most explanations fall into a 'what'-style explanation type; these 'what' explanations appear to have many different purposes, especially describing surprising results, warning others about mistakes, and advising how to handle certain cases. Notably, relatively few statements answer 'why' or 'why-not' questions—and these represent justification-style explanations that are probably the most

typical ones in current XAI systems. SQA sites like the CXAI system give different types of ‘what’ - style explanations that users understand the AI system (Ibne Mamun et al., 2022; Mamun, Baker, et al., 2021).

2.3.2 Content-Type

Explanations encompass a spectrum of knowledge components that are intricately entwined with the process of explanation and their mode of presentation. These encompass a variety of content types, each shedding light on distinct aspects of the explanation. For instance, concepts like input influence (the assessment and comprehension of the impact that various input variables or features have on the decisions and behavior of autonomous driving systems is referred to as input influence) and input sensitivity (input sensitivity refers to the extent to which changes or variations in input data influence the predictions or choices made by autonomous driving systems) delve into the degree to which different inputs affect the outcome of an explanation, highlighting the relative significance of each factor. In parallel, case bases provide concrete instances and scenarios that exemplify the principles underpinning the explanation, rendering it more comprehensible through real-world contexts.

Furthermore, the inclusion of demographic aspects in explanations is critical because it elucidates how various socioeconomic, cultural, and geographical factors contribute to the predictive behaviors stated in the explanation. Because of unique contextual considerations, these demographic elements might reveal nuanced patterns that demonstrate how certain groups or places may be more influenced by the explanation's predictions.

The SQA/CXAI platforms are primarily case based, as users share their experiences with autonomous vehicles (Linja et al., 2022).

2.3.3 System Type & Scope

Omeiza et al. (2021) propose explanations can be data-driven (i.e., explaining the outcome of a predictive model) or goal-driven (explaining an AI's behaviors based on achieving its goal in a predefined setting). The goal-driven explanation helps in the goal-driven learning process that applies at least four types of learning strategies: Learning by gathering information, learning by transforming existing knowledge (e.g., during the inference process used to elaborate explanations), learning by forming new connections between beliefs, and learning by storing and reindexing the explanations and information-gathering plans that it builds (Leake, 1994). Data-Driven explanations can be based on the local-to-global framework (Pedreschi et al., 2018). This framework is based on local and global explanations. Global explanations explain high-level decisions from an initial point to the destination, such as why an autonomous car chose a specific route, why it changed the planned route in the middle of the trip, and so on. On the other hand, local explanations can explain specific predictions. The CXAI system (SQA) tends to be at a much broader scope than most algorithmic XAI systems achieve insofar as they focus on single cases one at a time (Mamun, Baker, et al., 2021). It has also had the characteristics of goal-driven explanations. Within the framework of the CXAI/SQA system, learning occurs through multiple avenues. This encompasses acquiring knowledge through information collection and adapting existing knowledge through processes such as inference to generate explanations (in Figure 4, individuals utilizing an image classifier share their encounters

with the classifier, allowing a solitary user to amass information. Moreover, the comment section employs an inference process to provide detailed explanations).

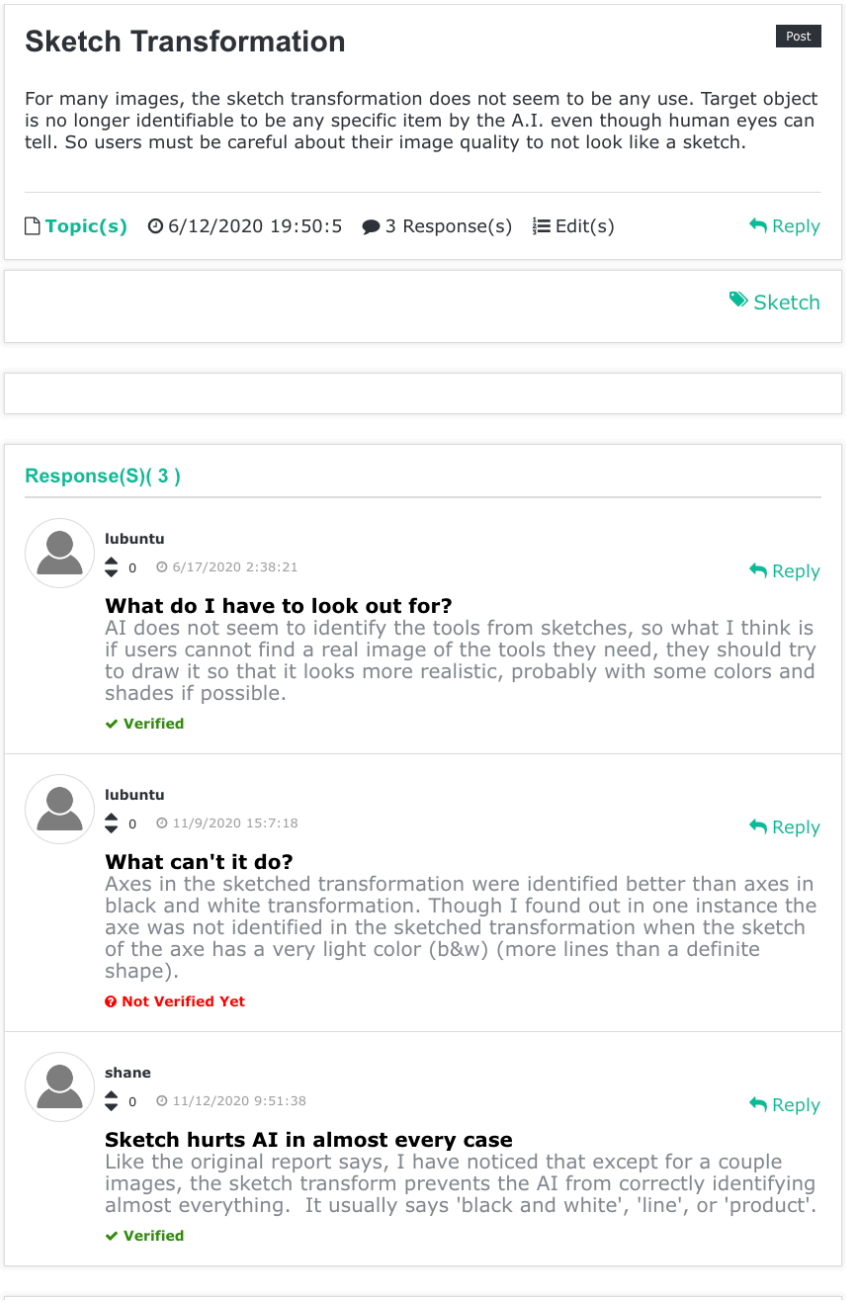


Figure 4. Engagement in the CXAI System regarding an Image Classification

2.3.4 Interactivity

The concept of interactivity comprises the opportunity for users to ask follow-up questions in order to obtain a better understanding of a given explanation. The emphasis of interactivity in traditional algorithmic XAI systems is primarily on the interaction between the system and people. However, in the setting of CXAI, the interaction takes on a new form by focusing on interactions between humans. CXAI, in other words, emphasizes the interaction and participation of various persons striving to collectively improve their understanding of AI-driven explanations.

The ICAP (Interactive>Constructive>Active>Passive) Framework tells us that learning is better when human-human interaction is present through dialoguing (Chi & Wylie, 2014). In SQA sites, the users can ask follow-up questions to their peers to understand an AI's features.

The discussion in the section establishes what dimensions of explanations the CXAI/SQA sites offer. The dimension content type also helps produce reasoning ability in the users for the CXAI/SQA.

2.4 Autonomous Vehicle Takeover and Real-Time Decision-Making

In Level 2 autonomous vehicle systems (Shadrin & Ivanova, 2019), the driver is obliged to take over control when automation fails. The car may provide alerts to prompt the driver to take control of the vehicle, or the driver may be required to intervene when the vehicle is behaving unsafely. In such instances, the driver must have both the time to assess the

situation and a window of opportunity to take control. When drivers need more time to evaluate the driving circumstance or devise an acceptable response, the likelihood of an accident increases (Abe & Richardson, 2004, 2005, 2006; Jamson et al., 2008).

Prior studies focused on take-over warnings via alarm systems rather than analyzing how drivers detect aberrant driving conditions. Damböck et al. (2012) discovered that for most visually distracted drivers, a lead time of 6 seconds was sufficient for take-over. When a 7-second lead time was compared to a 5-second lead time, (Gold et al., 2013) discovered that reaction times were slower but of better quality. According to the same study, when the available time to avoid colliding with a stationary object decreased, drivers responded to take-over requests with more aggressive braking and turning.

As per the findings of Mok et al. (2015), the study revealed that a significant portion of drivers, whether exposed to 5-second or 8-second unstructured transition-to-driving scenarios, managed to effectively navigate road hazards when operating under automated driving conditions while concurrently engaged in secondary tasks. As a result, the timeframe of 5-8 seconds appears as a viable window for making real-time decisions in autonomous driving, based on environmental perception, temporal sequence data processing, and transforming instantaneous observation into appropriate action.

It is critical to effectively relay information acquired from sensors or vision-based systems to drivers in order to support real-time decision-making and assure safe driving based on situational awareness. Lindemann et al. (2018) showed that using explanatory user interfaces that deliver timely and relevant data is beneficial in improving SA.

Endsley (1988) claimed that the intrinsic out-of-loop nature of autonomous driving resulted in drivers having a lower level of situational awareness (SA) in scenarios involving manual

or semi-autonomous driving. Domain expertise could help one to gain better SA (Endsley, 2006).

The strategy of capturing domain expertise, often termed predictive knowledge in the context of reinforcement learning, has traditionally been centered around algorithms. This approach has been predominantly employed within the realm of autonomous systems research to train artificial intelligences (AIs), as evidenced by studies conducted by Comanici et al. (2018), Sutton et al. (2011), and White (2015). The 'algorithms' for humans to gain predictive knowledge for AIs can be social forums.

Linja et al. (2022) conducted research indicating that platforms utilizing SQA (S. Oh, 2018a), possess the capability to accurately compile a substantial volume of resolutions pertaining to AI systems integrated within autonomous vehicles. This reservoir of knowledge holds the potential to provide explanatory insights to the drivers of autonomous vehicles, thereby assisting them in making informed real-time decisions based on prediction. In essence, this approach opens up the avenue for non-algorithmic mediums to contribute significantly to enhancing the decision-making process and user experience in autonomous driving scenarios.

2.5 Communication in Social Forums

The preceding sections highlighted the limitation associated with human inputs inside explainable AI systems, emphasizing the possibility of using SQA, as defined by Oh, (2018a), to alleviate this limitation. This section will look into the communicative aspect of SQA platforms.

Communication has always been an important tool for understanding different circumstances throughout history. Notably, the power of communication extends beyond its sheer utility; it has also been exploited in a variety of circumstances. Marketers, for example, use communication tactics to assess client loyalty, as Ball et al. (2004) discovered. This study shows how the subtleties of communication shared between marketers and customers provide insights into the depth of client loyalty, underlining the diverse importance of communication in understanding and interpreting human behavior and preferences.

Engagement: According to Shah et al. (2009), SQA platforms often use a public, community-driven approach, depending primarily on unstructured, freeform natural-language content rather than standardized formats. These platforms also use simple voting procedures rather than complex algorithms to determine significant, relevant, and precise information.

Even though the posted issues and solutions may potentially contain errors and often originate from anonymous contributors, they carry a degree of credibility. Beside the communal validation process (like voting), the absence of manipulation by the system's developers, rendering the content reliable and representative of collective expertise.

However, in order to motivate users to engage for collective expertise, individuals interacting with a SQA platform must be highly motivated to actively engage within the platform. While a small community or a cohesive team may be inherently oriented toward intrinsic communication, various different SQA systems have incorporated distinct

characteristics like bounties in StackOverflow (Zhou et al., 2020) to encourage participation.

Trust alignment: Koskinen et al. (2019) reported that users of Tesla AP realigned trust in the system after misplacing trust when encountering unexpected situations that differed from their initial expectations. Communication in social groups (e.g. tweets in former Twitter) can be a tool to calibrate trust (Mirnig et al., 2016) and teach the appropriate use of automation, which can lead to fewer safety incidents through communication guidelines (Alambeigi et al., 2021). Researchers can also use social forums to learn more about safety issues that users with specialized knowledge can only discover, unique orientations toward a subject, or experiences in exceptional circumstances (K. Chen & Tomblin, 2021).

Quality of the contents: Although the idea of giving explanations through social forums or SQA sites is still relatively novel, it was previously argued that the contents of a SQA site satisfy many of the 'Goodness Criteria' (Mamun, Baker, et al., 2021) AI explanations. Thus, using social media will likely help generate satisfying user-centric explanations, warnings, workarounds, and the like for users of semi-autonomous/autonomous driving systems.

Collaborative tutoring: Intelligent Tutoring Systems have been used effectively for scientific problem-solving (Friedland et al., 2004), electronic circuit design (Brown & Burton, 1978), propulsion engineering (Stevens & Roberts, 1983), etc. Intelligent Tutoring Systems aim to promote adaptive interaction between the learner and the content (George et al., 2016) through Socratic dialog. Communication in social forum can facilitate collaborative tutoring (Graesser et al., 1995), mainly with such dialog.

In one study, a pair of learners tutored one another. A third person observed collaborative tutoring. The results showed that observing collaborative tutoring can help a learner as much as being directly tutored in a tutoring dialog (Chi et al., 2008). Research has also shown that people can learn from errors if they recognize them (Chi et al., 2001; VanLehn et al., 2003). A joint effort between tutors and students can improve learning (Chi et al., 2001). The ICAP (Interactive>Constructive>Active>Passive) Framework tells us that learning is better when human-human interaction is present through dialoguing (Chi & Wylie, 2014).

Communication elements: Communication elements or factors depend on the field where the communication is taking place. Communication factors will differ for a Fire Fighting Team (Min et al., 2004) to a Nuclear Plant Team (Schraagen & Rasker, 2001). Still, a new scheme for communication can be generated for a new field based on the past literature as they have necessary factors that help to describe communication for a specific situation. Based on past works (Bylund & Makoul, 2005; S. Kim et al., 2010) and several inclusion criteria (for example, excluding inter-personal, off-topic communication) for communication in social media on a shared interest, Linja et al. (2022) coded the communication in a Tesla drivers social forums (more on this will be discussed in later chapters). With other communication factors, the coding scheme has empathic factors to investigate the presence of empathetic elements in this communication, as empathy creates a supportive/confirming atmosphere (Redmond, 1989) in a community of users. The coding scheme will tell us how social forum participants communicate and what drives them to share knowledge with their peers.

The literature review highlighted XAI systems' shortcomings in the autonomous driving sector and how a non-algorithmic approach can solve these problems. The review also tried to justify the use of this approach in autonomous driving.

For this Ph.D. dissertation, the experiments are designed to study the use of SQA as an explanatory and training platform for autonomous driving. The proposal will have one communication analysis, two simulation-based studies, and one survey based on the use of the SQA platform for autonomous driving.

The literature review has influenced the development of several experimental methodologies. This report aims to examine several aspects related to the adoption and usage of the proposed XAI framework. Specifically, it seeks to investigate the time required for individuals to ‘take-over’ based on the training, assess the communication in social forums, explore the potential for collaboration among inexperienced users in understanding autonomous vehicle actions using the social forum option within a non-algorithmic framework, and analyze the persistence of autonomous vehicle operators in utilizing AI technology despite its non-human attributes.

3 Communication Insight: Analyzing User Interactions in Social Forums

The primary objective of this study is to examine the prevalent communication patterns observed inside social forums - SQA platforms (S. Oh, 2018b). The objective is to gain a comprehensive understanding of the potential value of these platforms as instruments for exploration and training for autonomous drivers through a meticulous examination of these communication patterns.

The investigation is centered around an in-depth analysis of the communication patterns seen in the SQA platforms or social forums of Tesla's Full Self-Driving (FSD) system - Bjørner (2017). The interactions, exchanges, and conversations among users of the FSD inside various SQA contexts are meticulously examined in this coding-intensive study. By looking into several communication elements, the study investigates the dynamics of user interaction, collaborative discussions, problem-solving conversations, and knowledge-sharing instances that occur inside this unique ecosystem.

3.1 Research Questions

The study is going to answer the following questions,

R1: What elements (Framing-Reframing, Resolution, Emotion, and Cognitive Empathy) are more prominent in the communication in an SQA platform?

– Definitions for these elements are given in later sections.

R2: Are there any factors that enhance communication within the SQA platform?

3.2 Material

In order to examine the problems and solutions reported by Tesla FSD users on social forums/media, two researchers conducted a study analyzing social media posts from 11 October 2021 to 8 November 2021. This time frame encompasses the initial month prior to and subsequent to the widespread beta launch of the Tesla FSD AI System in October 2021. The selected end date was utilized to prioritize the analysis of preliminary reports and the acquisition of knowledge obtained during the opening weeks of the system implementation. The present study involved an examination of message boards that specifically facilitated threaded conversations. Initially, a total of 1257 posts pertaining to Tesla Full Self-Driving (FSD) were identified. The majority of these online discussion platforms encompassed numerous topics or threads that were unrelated to FSD and could be readily disregarded. From a corpus of 1257 original postings, we have discerned the presence of 101 primary posts and 95 subsequent responses that specifically pertain to the FSD system. This amounts to a cumulative total of 196 posts. Posts were considered eligible if they pertained to an unforeseen reaction or behavior exhibited by the car; a malfunction; a concern regarding safety; an unlawful maneuver; an unfavorable encounter with the decisions made by the vehicle's FSD system; a pertinent remark or proposed resolution in response to a previously mentioned issue. We omitted comments that consisted of jokes, memes, conversations unrelated to the topic, solely positive feedback, talks solely focused on safety scores (the criteria employed by Tesla to evaluate a driver's

eligibility for utilizing the FSD mode), software versions, or remarks pertaining to Tesla employees.

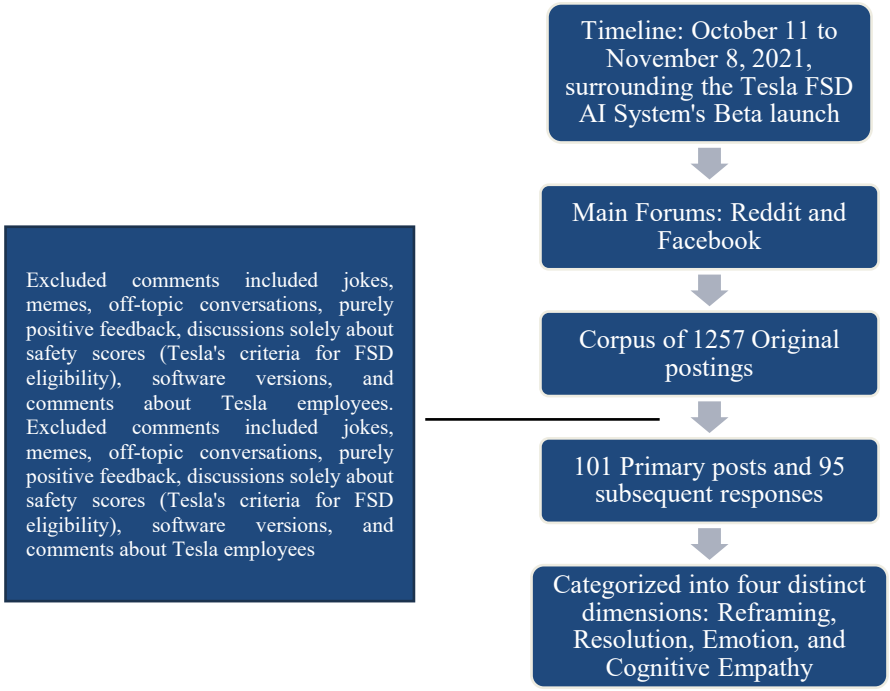


Figure 5. Process for Communication Analysis

The dataset comprised a total of 196 posts, originating from various online platforms. Specifically, there were 46 posts from the Reddit group thread "r/teslamotors" dedicated to Tesla owners and enthusiasts (https://www.reddit.com/r/teslamotors/, accessed on 11 November 2021), 17 posts from the Facebook group "TESLA Owners Worldwide" catering to Tesla owners and enthusiasts (https://www.facebook.com/groups/teslaworldwide, accessed on 11 November 2021), three posts from the Facebook group "Tesla Model 3/Y Owner Technical Support" intended for owners and enthusiasts (https://www.facebook.com/groups/teslamodel3/ownertechnicalsupport/ownertechnicalsupport, accessed on 11 November 2021), three posts from the Facebook group "Tesla Tips & Tricks" serving both owners and non-owners

(https://www.facebook.com/groups/teslatips, accessed on 11 November 2021), and finally, 127 posts from online message boards dedicated to discussions on AI/Autopilot and autonomous Full Self-Driving (FSD) for Tesla owners (https://www.teslaownersonline.com/threads/fsdbeta-megathread-for-all-fsd-beta-discussions.18878/, accessed on 14 January 2022).

The entire corpus is accessible at the provided link (https://osf.io/6jur3/, retrieved on August 9, 2022).

The present study utilized a total of 196 postings for the purpose of conducting a communication analysis, which will be expounded upon in subsequent sections. The aforementioned posts often included threaded discussions consisting of initial inquiries and subsequent responses. We selectively included only those follow-up messages that met the aforementioned criteria. A comment that has numerous sub-comments that are both related to the topic of FSD and distinct from the parent comment is classified as a base post. The process is depicted in Figure 5.

3.3 Coding Scheme

Each response was coded according to four independent dimensions, each with several coding labels. These include,

1. **Framing-Reframing:**

   The performance of a team can be significantly impacted by the quality of communication inside the group. Highly productive teams have proficient communication patterns, such as refraining from engaging in informal conversations during periods of heavy task and actively exchanging information in

advance of its immediate necessity (Schraagen & Rasker, 2001). Several research studies, including those conducted by Smith-Jentsch et al. (1998), Waller (1999), as well as by Shakeri & Khalilzadeh (2020), have provided valuable insights into the significant impact of timely information gathering and exchange on improving team performance. This research emphasizes the significance of teams maintaining a high level of awareness regarding scenario changes, as this understanding plays a crucial role in guaranteeing their efficacy and accomplishment.

One of the key results that have arisen from previous research (S. Kim et al., 2010) is the substantial impact of communication in both fostering comprehension and shaping viewpoints or attitudes on a specific subject or concept.

Research has consistently demonstrated that effective communication involves not just communicating information but also ensuring that the recipients properly comprehend it. Furthermore, communication has the ability to influence how others view or think about a given topic in ways that go beyond simple comprehension. This indicates that when communication is used skillfully, it has the potential to transform people's perspectives or attitudes, causing them to perceive an issue in a new light or shift their preexisting beliefs.

The analysis of this type of communication pattern within an XAI system is crucial as the main goal of such a system is to modify viewpoints and improve understanding of different aspects of an AI system.

So, Framing-Reframing dimension can be defined as the way in which a statement updates, modifies or enhances a user’s thinking about the AI system. This coding scheme was developed based on research on team communication on a Nuclear

Power Plant, Aircraft Cockpit and Command and Control teams (Foushee & Manos, 1981; Min et al., 2004; Schraagen & Rasker, 2001) – see Table 1.

Table 1. Elements/Labels of the dimension: Framing-Reframing and their definitions

Elements/Labels of the Dimension	Description
Evaluation	Evaluative utterances or judgments concerning the activities of the scenario just played out. Analysis of why things went well or wrong.
Clarification	Clarifications serve to clear up misunderstandings from other individuals. Questions and answers that someone either asked or seemed to misunderstand. This includes repetitions for clarification, associations, and explanations.
Observation	A statement that describes the AI's action during use.
Response uncertainty	Statements indicating uncertainty or lack of information with which to respond to a command, inquiry, or observation.
Denial or Disconfirmation	Disconfirming a statement.

2. **Resolution:**

The term "resolution" pertains to the act of identifying a solution or offering understanding to a problem, conflict, or state of ambiguity. We included this coding to determine how many of the posts resulted in satisfactory answers to queries, which Mamun, Baker, et al. (2021) determined occurred relatively rarely in CXAI systems. This particular dimension distinguishes itself from the "clarification" element of the Framing-Reframing dimension in communication. In instances where clarification is sought, posts or comments serve the purpose of furnishing an elucidation or explication of a remark, without necessarily proffering a definitive resolution. An example can be found by referring to Record #77-1490 in Appendix C, also see Table 2.

Table 2. Elements/Label of the dimension: Resolution and it's definition

Elements/Labels	Description
Situation Resolved	Combination of some of the other elements of Framing-Reframing—resolution/ workaround/abandonment of a practice conditionally/abandonment of a practice wholly/why it is doing it (not giving a solution but a reason).

3. **Emotion:**

Emotional Intelligence (EI) is defined as a form of social intelligence that involves the ability to perceive and express emotions, understand and utilize them, and manage emotions in a manner that facilitates personal growth (Cobb & Mayer, 2000). The central focus of the definition mostly revolves around individual development. However, it is crucial to acknowledge that EI can also have a favorable impact on team unity, taking into account variables such as influence and orientation towards achievement (Rapisarda, 2002).

In the case of an AI system, this report will primarily investigate the presence of user apprehension or frustration within the social forum related to AI, aiming to determine whether these emotions affect the AI's user base in terms of team cohesion.

For this the report used elements of communication from the article by Foushee & Manos (1981) – see Table 3.

Table 3. Elements/Labels of the dimension: Emotion and their definitions

Elements/Labels	Description
Frustration or anger with AI	During the use of AI.
Frustration or anger on response	During the use of AI.

Appreciation for the AI	During communication.
Appreciation for a Response	During communication.
Embarrassment	Any response apologizing for an incorrect response, etc.

4. **Empathy:**

The literature encompasses a range of opinions on empathy. Empathy has been characterized by certain scholars as possessing a cognitive component, commonly known as "role-taking" (Bellet & Maloney, 1991). Previous studies also have focused on the affective aspect, also referred to as "emotional contagion" in some cases (Stiff et al., 1988) as explored by Spiro (1992) and Zinn (1993). Additionally, Winefield & Chur-Hansen (2000) have examined the behavioral dimension.

In social forums, the presence of experts or leaders in the field can help alleviate misunderstandings related to that particular field (Walter et al., 2021). The leaders can play another vital role beside this. Leader behaviors that focus on fostering interpersonal understanding, demonstrating care, creating a positive atmosphere, and engaging in proactive problem-solving have been shown to have a stronger association with team trust, open communication, personal task involvement, and overall team effectiveness when compared to traditional, task-oriented leader behaviors such as issuing directives and asking questions (Druskat & Pescosolido, 2006).

This report will focus on the phenomenon of "emotional contagion" in the field of XAI. Specifically, it will explore the application of the altruism model, as presented by Stiff et al. (1988), in the content shared on social forums. The objective of the research is to examine the presence and applicability of this paradigm inside the

AI-driven content that is shared and discussed in online communities. For this purpose, empathetic elements such as sharing an experience, understanding someone's feelings, or sharing information have been choose from Bylund & Makoul (2005)'s study – see Table 4.

Table 4. Elements/Labels of the dimension: Empathy and their definitions

Elements/Labels	Description
Agreement/Acknowledgement	'A' conveys to 'B' that the expressed emotion, progress, or challenge is legitimate.
Shared experience	'A' has a similar experience to that of 'B' with progress or a challenge.
Perfunctory recognition	'A' gives an automatic, scripted-type response, or repeats the company's policy/response, giving the empathetic opportunity minimal recognition.
Antagonism	Deflates the other's response, defends or asserts self-response.

A detailed description of the coding criteria is shown in Appendix D.

3.4 Method

Two coders independently coded 196 observations (posts and comments) regarding FSD and Auto Pilot (AP) of Tesla on each of the four dimensions. The coding was dependent on the context, so the coding of a single comment was dependent on the parent post and earlier comments. If a comment was deemed a separate post based on its uniqueness and child comments, the coders separated it as a new observation after coming to a consensus. Following an initial coding round on a subset of items, the coders met to examine disagreements in coding, then completed a second coding round. The coders conducted an initial round of coding, followed by a further evaluation of the posts. During this review, any posts that did not receive unanimous agreement from the coders were deemed to not

pertain to a specific dimension. Following the completion of this coding phase, the inter-rater reliability was calculated.

3.5 Results

The results of this coding are shown in Table 2. The table shows counts of statements for which both coders agreed and counts for statements for which at least one coder specified the category (Elements/Labels). The kappa values in the tables of this section have been computed using the inter-rater agreement between the two coders. The tables presented in this regard provide information regarding the "Agreed by One" metric, which serves to inform readers whether one of the coders has agreed that a particular post corresponds to a specific element or label.

In the aspect of "**Framing-Reframing**," the coding process achieved a high level of agreement, kappa (McHugh, 2012): 0.78 – Table 5. Over half of the comments were categorized as 'observations,' aligning with previous research findings (Mamun, Baker, et al., 2021). It's important to acknowledge that this type of explanation system typically doesn't provide answers to 'why' questions due to the complex nature of these answers, often beyond the general users' knowledge and contextual understanding, which may not be effectively conveyed in an online format. However, the significance of 'what'-style posts lies in offering a type of information that conventional XAI systems tend to overlook.

Table 5. Coding of different subset of Framing-Reframing. The findings indicate that a significant portion of the content pertains to observations regarding the functioning of the AI system.

Elements/Labels	Count - Agreed by Both (Agreed by One)

Evaluation	20 (34)
Clarification	23 (31)
Response uncertainty	16 (19)
Observation	111 (135)
Denial or Disconfirmation	1 (3)
Other	1 (8)

Examples of statements coded in the main categories of Framing-Reframing encompass:

Framing-Reframing: Evaluation: A user summarized various reports about FSD turning behavior, drawing a comparison with the behavior of the Autopilot system (Record #12-1364).

Framing-Reframing: Clarification: A commenter inquired about the specific driving profile (chill, average, or assertive) that was configured (Record #77-1490).

Framing-Reframing: Response uncertainty: A user raised a question regarding the use of only a subset of adaptive cruise control features (Record #14-1371).

Framing-Reframing: Observation: A user observed specific conditions under which FSD was not functioning properly (Record #1-1338).

Moving to the "**Resolution**" dimension - Table 6, approximately 20-30% of the instances pertained to resolution, kappa for the dimension is 0.7. The bulk of posts and comments in the Tesla communication chain revolved around the AI's actions during usage. Although most resolutions centered on responses to other comments, they often marked the conclusion of a discussion thread.

Table 6. Coding of different subset of Resolution. The findings indicate that a significant portion of the content remained unresolved regarding the AI system

Elements/Labels	**Count - Agreed by Both (Agreed by One)**
Situation Resolved	38 (60)

Situation Not Resolved	136 (158)

An illustration of a statement coded as a resolution is as follows:

Situation Resolved: A user suggested turning off sentry mode as a means to prevent conflicts between FSD and Autopilot (Record #2-1343).

Shifting focus to the realm of "**Emotion**," the prevailing sentiment conveyed in Table 7's content primarily consists of frustration and anger directed towards the AI system, with relatively few instances expressing gratitude or satisfaction. The underlying reason for this pattern can be attributed to the AI's imperfections and limitations. The kappa between two coders is 0.78.

This discovery implies that users have had difficulties or limitations during their interactions with the AI, resulting in adverse emotional reactions, leading to the prevalence of displeasure in user feedback. The critical remarks enhance the dynamics inside the social forum, laying the groundwork for cooperative interactions among users of AI system.

Table 7. Coding of different subset of Emotion. The findings indicate that a significant portion of the content are anger with the AI

Elements/Labels	Count - Agreed by Both (Agreed by One)
Frustration or anger with AI	96 (120)
Frustration or anger on response	0 (2)
Appreciation for the AI	27 (36)
Appreciation for a Response	0 (3)
Embarrassment	0 (1)
Non-emotional	47 (60)

Examples of statements coded in each emotion category encompass:

Emotion: Frustration/Anger: A user exclaimed "so ALWAYS be prepared to take over" (Record #8-1355).

Emotion: Appreciation: A user stated ". . .I've enjoyed it so far" (Record #10-1360).

Record #8-1355 presents an instance wherein a participant in a social forum offers a proposal to another participant. This proposal seems to originate from the first user's emotional state of irritation or resentment towards the AI system.

It can be inferred from the context that the first user may have faced a problem or encountered a level of discontent with the AI system, which then evoked an emotional reaction of fury. In light of the aforementioned frustration, the user actively presents a recommendation to another individual, potentially aimed at resolving or alleviating the issue they encountered. This exemplifies the impact of emotions, specifically rage, on user behavior within social forums, resulting in positive behaviors such as offering suggestions or providing comments to other users. The kappa between two coders is 0.81.

With respect to the concept of "**Empathy**", it was noted that individuals had a tendency to assist their counterparts by offering advice, support, or additional perspectives in different scenarios, as outlined in Table 8. The aforementioned discovery suggests the presence of the altruism model in the content published on this specific type of social forum. Users on this platform exhibit unselfish behavior by providing assistance and aid to those who require it, thereby cultivating an atmosphere of cooperation and mutual support within the online community.

Table 8. Coding of different subset of Empathy. The findings indicate that a significant proportion of the content exhibits a lack of empathy towards others. However, there is also a noteworthy observation on the prevalence of shared experiences among users

Elements/Labels	Count - Agreed by Both (Agreed by One)
Agreement/Acknowledgement	23 (54)
Shared experience	21 (33)
Perfunctory recognition	0 (1)
Antagonism	9 (12)
Non-empathy	124 (138)

So, contents reflect empathy, albeit within a relatively small subset of comments (approximately 20%). Some instances of statements include:

Empathy: Agreement: A user expressed agreement by saying "Yes!" in response to a prior post discussing the future of AP navigation (Record #52-1449).

Empathy: Shared experience: A user shared a similar experience of their initial drive using FSD (Record #28-1410).

Empathy: Antagonism: A user noted that a previous user's account of turning behavior did not align with their own experience (Record #31-1415).

The records are in Appendix C.

3.6 Discussion

The analysis showed that user communication focusing on understanding a new system mostly (up to 75%) involves ‘what’ type of observation–explanations (Mamun et al., 2021). Despite this, posts tended to support the resolution (up to 30.6% of statements, rest are non-resolutions). However, many resolutions remained perfunctory, and specific motivations

(Mamun et al., 2021) may be needed to encourage resolutions with more complete reasoning traces.

The coding procedure used to determine emotional content is very intriguing, in part because it was found that about half of the comments had emotional expressions, mostly negative ones. The intriguing aspect of emotional states lies in their capacity to drive motivation. In this context, the prevalent negativity in emotional tone suggests a prominent underlying factor driving individuals to engage by visiting these platforms and sharing problem-related experiences.

This perspective implies that emotions, particularly the potent feelings of frustration and anger, serve as formidable catalysts for individuals to actively participate in these forums. The current findings highlight the necessity to recognize and take use of emotions' power to motivate, while previous research has mostly focused on concrete factors like response quality and gamified rewards. A new aspect of participation dynamics is revealed when it is acknowledged that users' decisions to share experiences and report problems on the SQA platform are greatly influenced by their emotional states.

There are sympathetic elements that show up in instances of shared experiences and agreements within a small but significant portion of the remarks. Intriguingly, these statements show a type of user-to-user empathy collaboration, which may appear strange in the context of a bug-reporting system supported by the system's own vendor.

This finding raises the intriguing notion that users engage with social media platforms because of a shared interest, fostering a sense of cohesive group identity. Users showing empathy and support for one another is a phenomenon that reveals a further intrinsic motive that is not always expected in the context of complaints made to governmental safety

boards or bug-reporting mechanisms supported by the system vendor. The shared experiences may be able to reassure drivers that any unusual driving incidents are more likely the fault of flawed AI systems than of their own driving skills. This needs to be verified in light of additional study.

In a broader sense, this analysis highlights both the positive and negative aspects of the convergence of social media (SQA) systems, which serve as support structures for improving users' knowledge of AI systems. This, in turn, helps to nurture the collaborative cognition that enables semi-autonomous or autonomous vehicle supervision.

This assessment reveals a significant strength in the system's ability to identify possible flaws with the automation process, particularly within the precise situations in which these failures occur. This knowledge is critical because it enables users to foresee such situations and be prepared to either take control or avoid them totally.

In addition, the coding process reveals interesting insights on the synergistic relationship between social media and problem solving. On the one hand, factors inherent in social media platforms help to promote reporting, facilitate conversation, and foster clarification. This behavior is fueled in part by drivers' dissatisfaction and anger, which may lead them away from regular bug reporting channels. The anger was not found as a motivator when participants were asked to participate in a similar type of system for an Image Classifier (Mamun, 2021). The coding process, on the other hand, reveals indicators of cooperative and sympathetic relationships among users. These exchanges demonstrate the platforms' ability to foster a culture of mutual assistance, which is supported by collaborative efforts to assist fellow users in navigating obstacles to overcome.

In essence, our analysis emphasizes that the SQA systems have the ability to give users with useful insights and help as they attempt to comprehend AI systems. This synthesis not only helps with problem identification and contextual knowledge, but it also initiates a cooperative dynamic motivated by user dissatisfaction and empathy. These findings highlight the importance of this hybrid system in developing a collaborative cognitive framework required for supervisory control of vehicles operating in semi-autonomous or autonomous modes.

4 Detection Study

The previous analysis has shown that communication in a SQA platform identifies possible problems with automation, especially the contexts in which the failures happen. The user study will mainly train the users on a corpus of SQA posts and investigate if the users can detect if the drivers can stop abnormal driving before it happens.

4.1 Research Questions

The study will answer the following questions,

R3: Does the existence of autonomous vehicle-oriented social forums lead to an elevated perception of training quality in the autonomous system?

R4: Does exposure to user-generated content on social forums contribute to an improved ability among users to identify potential problems with increased accuracy and speed?

R5: How does training influence the self-reported workload of participants concerning the identification of potential problems through user-created content on social forums?

4.2 Participants

24 participants with an average age of 19.68 with more than 3 years (average) of driving experience participated in the credit-based compensation driving simulator study. The participants are recruited through MTU Sona and have an average driving experience of 6290.4 miles in a year.

4.3 Method

For this study, the researcher of this study identified 33 base posts with comments (linked with abnormal driving situations) from the 196-communication corpus (see section 3.2). Based on the abnormal driving situations, the researcher simulated 15 autonomous abnormal driving situations in a NADS miniSim - https://nads.uiowa.edu/minisim driving simulator. For training purposes, 33 base posts with comments have been put on a web-based platform (see Figure 7 for an example). For base condition, the same amount of news of Tesla based on Google News (see Figure 8 for an example) was collected from January 2022 to September 2022 and put in a separate web-based platform. Two researchers gave

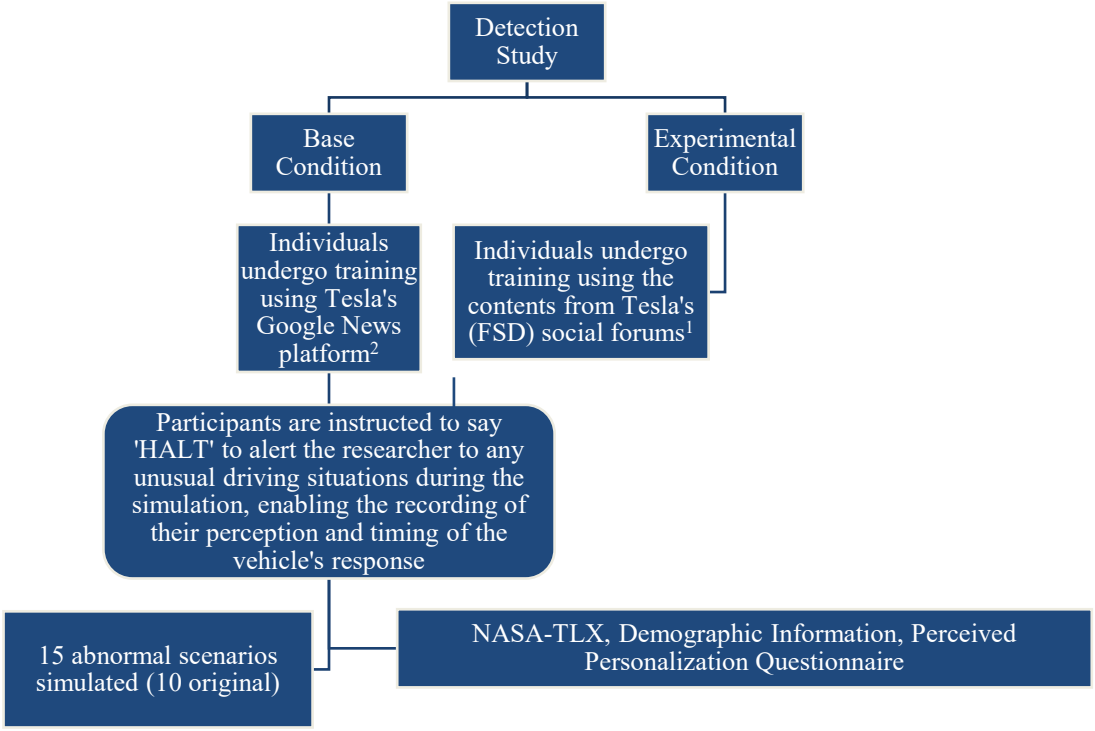


Figure 6. Design for Detection Study

1. https://cxai-auto.netlify.app/
2. https://cxai-auto-c.netlify.app/

- Use any numerical participant id to see the contents.

keywords to each base post based on consensus - see this article (Werner et al., 2019) for consensus coding, in the two conditions. The conditions are depicted in Figure 6.

To understand training quality, the researcher used the perceived personalization questionnaire (Zhang & Curley, 2018)- a scale from Strongly Agree to Disagree Strongly. To assess if the training effects the perceived workload during driving, NASA-TLX (Hart & Staveland, 1988) questionnaire was used.

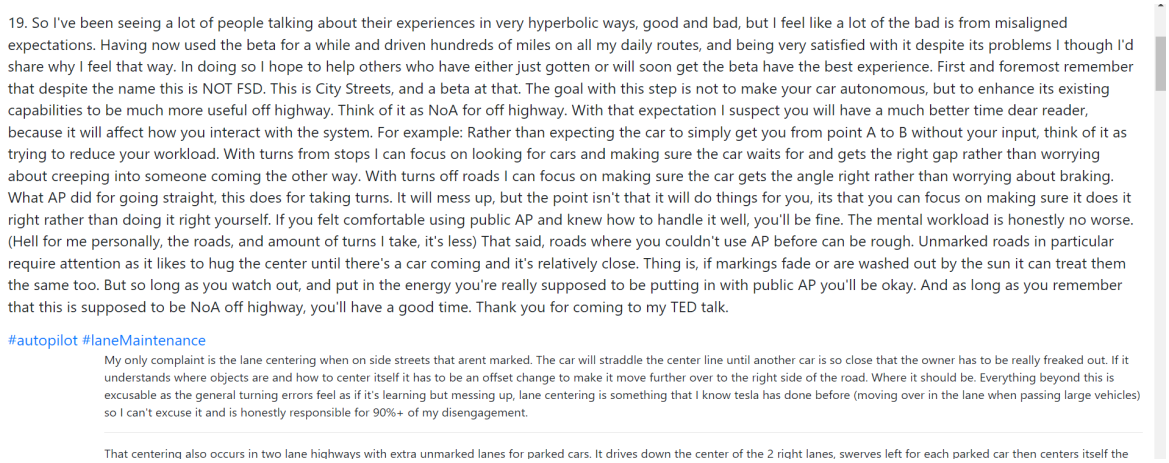


Figure 7. An Example of the Training Material for the Experimental Condition

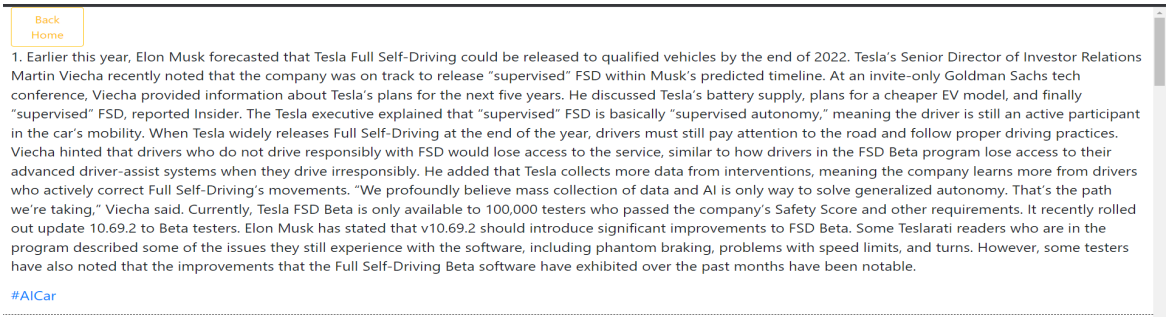


Figure 8. An Example of the Training Material for the Base Condition

An independent researcher who is not linked with this research went over the 15 abnormal driving conditions and determined the correct response based on the corpus from both conditions.

The participants filled up the demographic questionnaire before the study and went through a simulator sickness test in the simulator. The participants were given access to the training website based on the condition. They were given a maximum of 30 minutes to go through the observations on the site. The researcher of the study meticulously recorded how long participants spent on each webpage. Participants were provided with guidance to audibly express "HALT" (in the event of encountering an abnormal driving scenario during the simulation) as a means to notify the researcher during driving. This enabled the researcher to record both the timing and the driver's interpretation of the vehicle's response to the abnormal driving event. After the simulator study, the participants filled up the two questionnaires. Participants were assigned randomly to one of the two conditions.

4.4 Results

We determined whether the participant stopped the scenario, whether they correctly identified the situation, and the time (in relation to the event marker), with larger negative values indicating anticipation of the problem and larger positive numbers indicating slower response to the problem. Responses were coded, and each participant received a score based on the number of scenarios properly identified. Although there was no statistically significant difference in the number of times participants paused the playback, participants in the experimental condition provided correct justifications more frequently (5.79; se=0.41) than the control (3.45; se=0.56) - see Figure 9, which was significantly different

($F(1, 20) = 20.81$, $p < 0.001$). Furthermore, the responses arrived significantly earlier for the experimental group ($F(1, 20) = 12.08$, $p < 0.001$) in cases where events were accurately identified.

This suggested that after being exposed to social media posts, participants identified occurrences with greater promptness and accuracy. We also used the PPQ (Perceived Personalization Questionnaire) to assess how successful the training material was for the task.

An independent-samples Welch t-test revealed a significant difference between the conditions ($t(21.81) = -3.18$, $p = 0.004$), as shown in Figure 10, which explained the dimensions of the questionnaire. This discrepancy was reflected by greater ratings for the experimental condition across all three dimensions of the scale.

Finally, we looked into whether the training had an effect on perceived workload (see Figure 10), as measured by NASA-TLX ratings. Despite somewhat greater workload evaluations in the experimental condition (47.33 vs 40.5), this difference was not statistically significant ($t(19.33) = 1.37$, $p > 0.05$). The sub-scale evaluation of the NASA-TLX was not also statistically significant between two conditions except for frustration where the result shows higher frustration (8.92 vs 5.08) for the experimental training ($t(21.1) = 2.5$, $p < 0.03$).

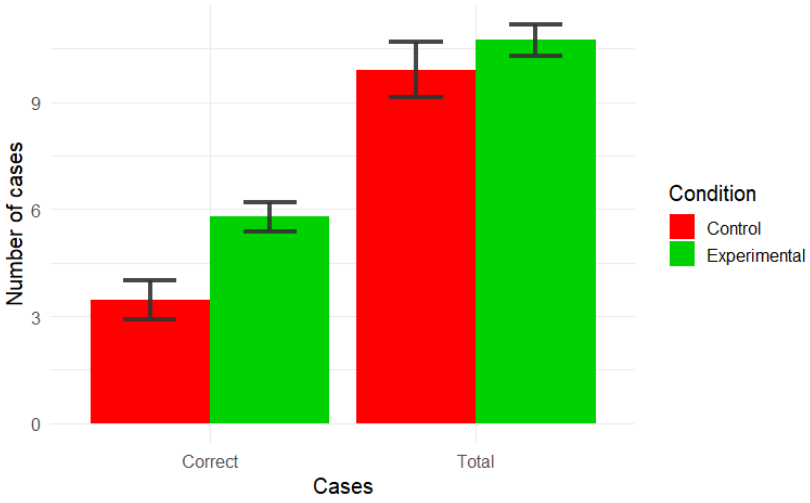


Figure 9. Total and Correct Cases Detected

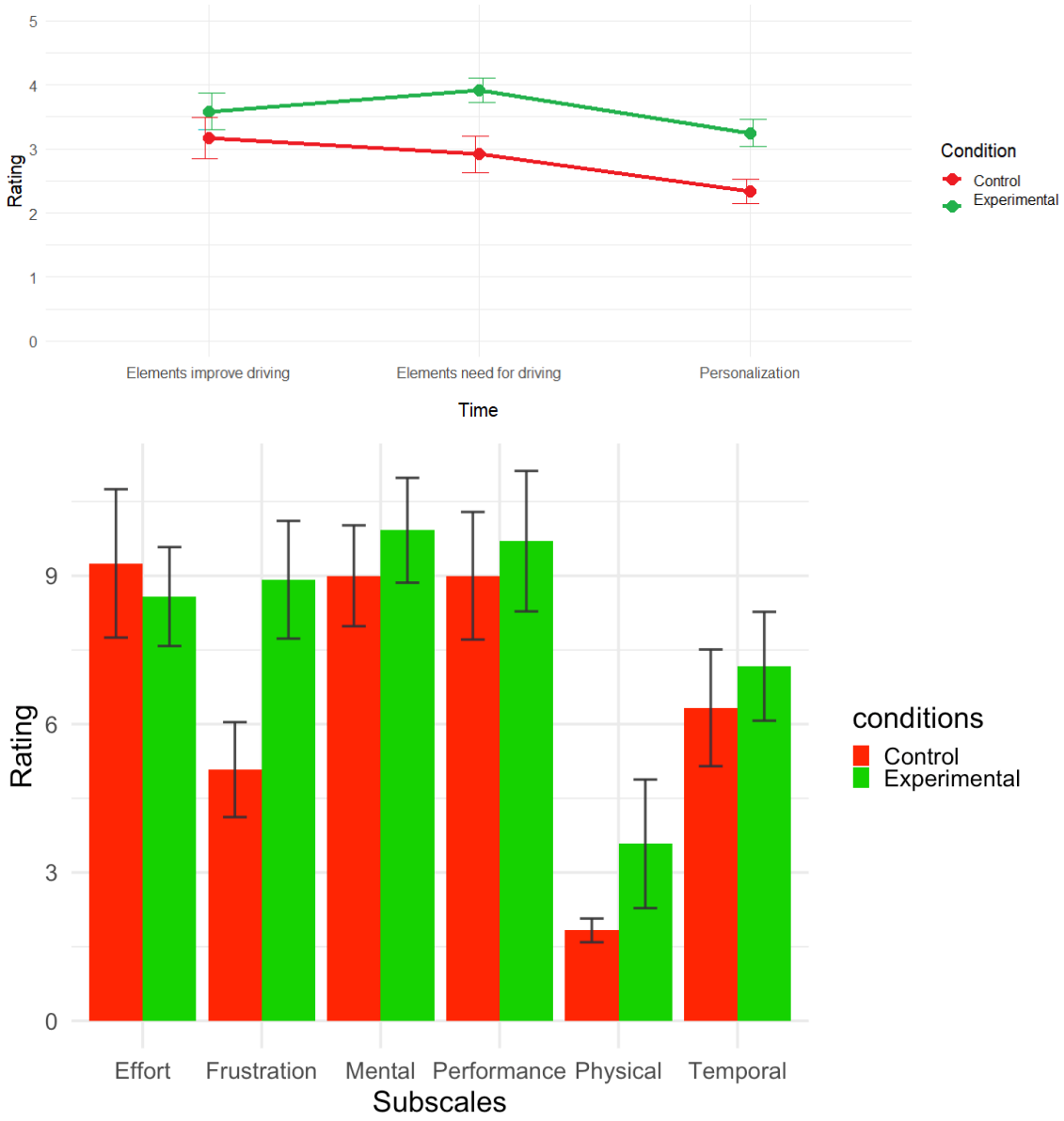


Figure 10. Top: Perceived Quality of the Training based on Three Separate Classifications. Bottom: Result from NASA – TLX (subscales). Effort: $t(21.58) = -0.44$, $p = 0.67$; Frustration: $t(21.1) = 2.5$, $p = 0.03$; Mental: $t(21.98) = 0.62$, $p = 0.54$; Performance: $t(21.77) = 0.09$, $p = 0.93$; Physical: $t(11.76) = 1.32$, $p = 0.21$; Temporal: $t(21.89) = 0.52$, $p = 0.61$

4.5 Discussion

This study found that giving participants the opportunity to peruse social media posts on Tesla's FSD boosted their abilities to recognize and identify potential problems in a simulated takeover scenario. Specifically, when compared to individuals who did not engage with the social media posts, these participants were not only more skilled at detecting scenarios, but they also did it more accurately and with faster response times. Surprisingly, individuals who received this training did not report an elevated workload during the scenario, showing that the extra exposure did not increase their perceived cognitive load during driving, but they were frustrated. The lack of structured training materials may have led to difficulty in recalling information during the drive. Alternatively, in the control condition, participants might have experienced less frustration since there was no meaningful information connected to the driving scenarios to recall. A more structured training platform may be needed to address this frustration as our website was very basic.

The trained participants did, however, exhibit faster reaction times and increased sensitivity to identifying anomalous driving conditions, although their average point of response occurred after the event marker rather than BEFORE it. This raises the question of whether this training actually helps users prevent accidents or if it primarily makes it easier for them to understand the accident's cause more quickly after it happens.

However, multiple reasons point to this finding being mostly an artifact of the simulation system. Participants may have felt safe in the simulated environment, which may have led to an over-reliance on the AI (Garcia et al., 2022). Furthermore, despite instructions to

instantly vocalize 'Halt' when detecting a risky scenario, the simulation's frequent recurrence of such occasions, along with the absence of actual accidents in the video, may have contributed to a greater threshold for participants to halt the simulation.

A following study was undertaken using the same videos and training content, but with a change in the response paradigm, to go deeper into the assessment of users' ability to foresee hazardous circumstances. The simulation was paused PRIOR to the incident in this new setting, and participants were tasked with guessing the nature of the impending crisis. This alternate technique sought to determine whether users could anticipate hazardous scenarios.

5 Prediction Study

Based on the result of the Detection Study, the researchers of this dissertation have decided to run another user study. We saw that in the detection task, the mean reaction time is closer to 0 for the SQA-trained participants. Prediction Study will investigate if the drivers can correctly predict incidents before a 5-second window (see section 2.4 behind this time frame) based on the training.

The challenge of trust calibration presents a significant obstacle in the context of human-automation collaboration, namely in the interaction between drivers and AVs. In order to establish a trust system that is highly accurate and dependable, it is crucial to obtain an accurate and up-to-date evaluation of drivers' levels of trust. The utilization of real-time measurement is crucial in enabling timely interventions or modifications in the operation of automated driving systems. One feasible strategy entails the utilization of machine learning algorithms and physiological data to understand the shifting patterns of trust (Ayoub et al., 2023).

Psychometric measures have also been employed in previous scholarly works to assess trust. Liu et al. (2019) identified three primary markers of acceptability towards AVs. These indications comprise the factors of broad acceptance, readiness to pay, and behavioral intention. Other studies have established robust associations between individuals' trust in AVs and several characteristics, including perceived dangers and safety, perceived utility, perceived emotions, perceived privacy, and knowledge about AVs (Ayoub et al., 2019; Kaur & Rampersad, 2018; Raue et al., 2019). This report will employ a trust questionnaire to examine the dynamic evolution of trust during the automated

driving based on the training sessions facilitated through social forums. The trust questionnaire encompassed six dimensions: System Reliability/Competence, System Understanding/Predictability, Developers' Intentions, Propensity to Trust in Automation, Trust in Automation, and System Familiarity (Körber, 2018).

5.1 Research Questions

The research question that will be addressed is,

R6: Can drivers of an autonomous vehicle correctly predict abnormal driving situations before 5 seconds with the training from SQA?

R7: What type of situations will SQA exposure improve simulated drivers ability to predict 5 seconds before the event?

R8: How does the level of trust and reliance shift from the training phase to the actual driving phase for an autonomous vehicle equipped with a malfunctioning AI?

5.2 Participants

Twenty-four college students (12 Females, 11 Males, 1 Non-Binary, and 1 unknown) have been recruited through MTU Sona based on a credit-based compensation structure. The participants are recruited through MTU Sona and have an average driving experience of 6669 miles in a year.

5.3 Method

Participants were assigned randomly to one of three conditions: the No-training (control) group, the partial training group, and the full training group. The training material employed in Detection Study was partitioned into two distinct sets denoted as Set A and

Set B. In the case of the full training[1] condition, participants underwent training on 10 original driving scenarios (with 4 scenarios from Set A[3] and 6 from Set B[4]) – see Figure 11. Meanwhile, participants in the partial training condition were exclusively trained on the 6 driving scenarios from Set B. Conversely, control participants engaged in reading generic Tesla materials[2]. Following a review period of approximately 10-15 minutes, each participant proceeded to navigate through 15 distinct scenarios (see Table 3) within the simulator. The initial split of the scenarios into two sets was conducted in a random manner. However, a subsequent examination was performed to identify instances where two or more cases were included inside the same social media post. In such circumstances, the scenarios were then segregated into distinct sets.

At the critical juncture of five seconds prior to each anomalous event, the simulation halted, prompting participants to elucidate their anticipation of the forthcoming developments (they were asked 'What will happen next?'). These responses were documented by the experimenter. Additionally, participants were administered a trust in automation questionnaire related to self-driving vehicles at three distinct intervals: prior to the commencement of the study, after undergoing the training, and post the simulated driving experience.

1. https://cxai-auto.netlify.app/
2. https://cxai-auto-c.netlify.app/
3. https://cxai-auto-50-a.netlify.app/
4. https://cxai-auto-50-b.netlify.app/

- Use any numerical participant id to see the contents.

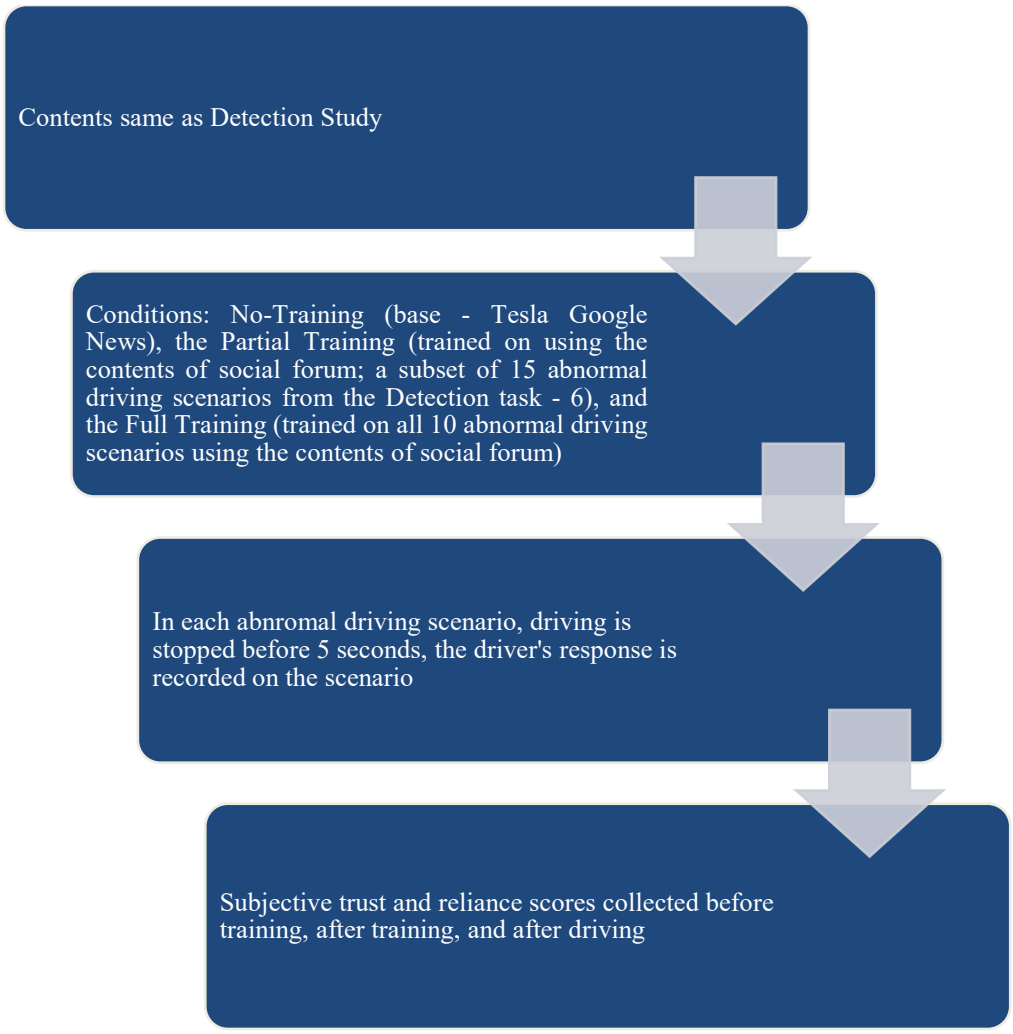
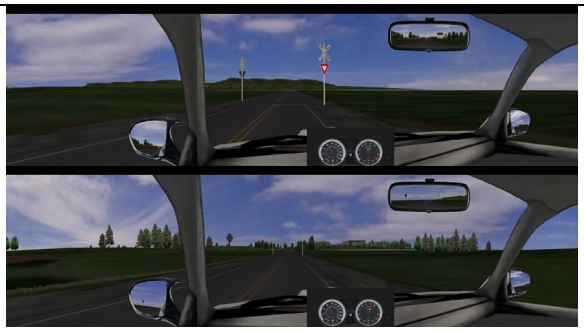


Figure 11. Process for Prediction Task

Table 9. 15 Anomalous Driving Scenarios and Correct Predictions

Anomalous Driving Scenario	Simulation Images	Correct Prediction
Occurred Before a Rail Crossing		Will not slow down before crossing

Occurred When Saw a Yellow Sign/Blinking Yellow		Will not read the yellow sign
Occurred When Saw No Center Lane Marking in the Neighborhood		Will move to middle in the un-marked road
Occurred When Saw a Stop Sign		Will stop early
Did a Rolling Stop		Creep forward
Occurred When Saw Stopped Cars in the Lane		Move to the oncoming traffic thinking the stopped cars as parked cars

Occurred When Saw a Road Work Ahead Sign		Will not understand the sign
Occurred When Saw a Bump		Recognitions is Hit or miss
Occurred When Saw a Fence/Gate		Don't recognize neighborhood gate
The Vehicle Suddenly Pulling Over to the Right		Will pull over to the right thinking there is an emergency vehicle in the lane

5.4 Results

The data were analyzed using a proportional odds logistic regression (polr) model implemented via the MASS library of the R statistical computing language. A Type-II Analysis of Deviance test showed a significant statistical difference between the accuracy of forecasting across the three conditions (X^2 (2, N = 28) = 20.52, p < 0.001) The participants identified significantly more anomalous driving situations in the full training condition (mean 0.62, standard error 0.04) than in the control/no training condition (mean 0.37, standard error 0.04) - t(17.99) = -3.68, p = 0.002 - see Figure 12.

Next, when we considered only Set B, the partial training condition had a mean accuracy of 0.63 which was not significantly different from the full training (t(16) = -0.53, p = 0.61), but was significantly greater than the untrained participants (t(15.9)=4.14, p<.001)

indicating that the social media posts were equally effective in both conditions in comparison to the untrained participants. For Set A, the partial training was significantly worse than the full training (mean 0.32, standard error 0.05; $t(15.5) = -3.07$, $p = 0.007$), but there was no statistically significant difference between the two untrained conditions ($t(15.9)=-1.19$, $p=.25$).

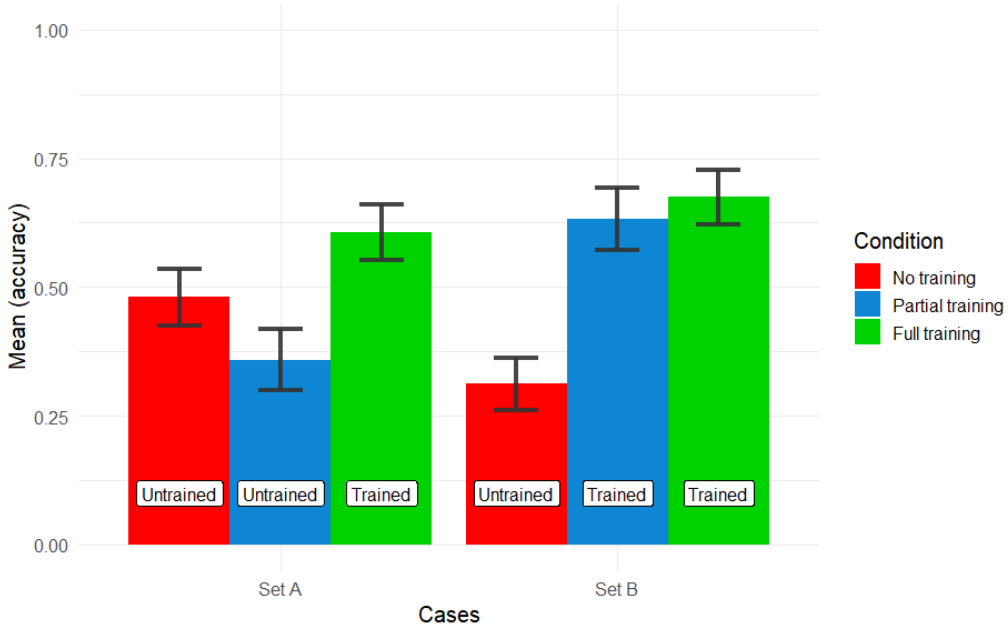


Figure 12. Mean Accuracy for identifying problems when prompted. The no-training condition studied neither Set A or Set B media posts; the partial training studied only Set B, and the full training studied both Set A and Set B

The outcomes of subjective trust and reliance scores (see Figure 13) revealed that, in comparison to the baseline scores collected before training, the average scores across all six dimensions rose by nearly one unit immediately following training across all three conditions. Subsequently, a paired-samples t-test confirmed that the mean ratings post-training significantly exceeded those pre-training ($t(46) = -7.54$, $p < 0.001$). It's noteworthy that although the increase was least pronounced for users exposed to the most negative

social media posts (full training) and greatest for users who encountered no such negative posts (control), these differences were not statistically significant based on a one-way ANOVA ($F(2, 9.9) = 1.19$, $p = 0.343$). Furthermore, there was minimal change after testing, and the slight decrease observed in the control users' ratings from before to after testing (individuals who had not been previously exposed to descriptions of driving errors) was also not statistically significant according to a one-way ANOVA ($F(2, 16) = 2.4$, $p = 0.123$).

The driving incidents (see Table 3) are categorized into three event groupings. Within these, we have 'Center Lane Absence,' 'STOP Sign,' and 'Yellow Sign,' which are all examples of Road Signs. Additionally, we can classify events like 'Road Bump,' 'Fence,' 'Road Work Ahead,' and 'Rail Crossing' as Obstacles. Lastly, the occurrences of 'Sudden Right Pullover' and 'Rolling STOP' can be linked to Unplanned Vehicle Maneuvers.

Drivers exhibit an exceptional understanding of the actions connected to road signs and barriers, as seen in Figure 14. However, it appears that they have certain knowledge gaps, particularly when it comes to non-exploratory acts like abrupt turns.

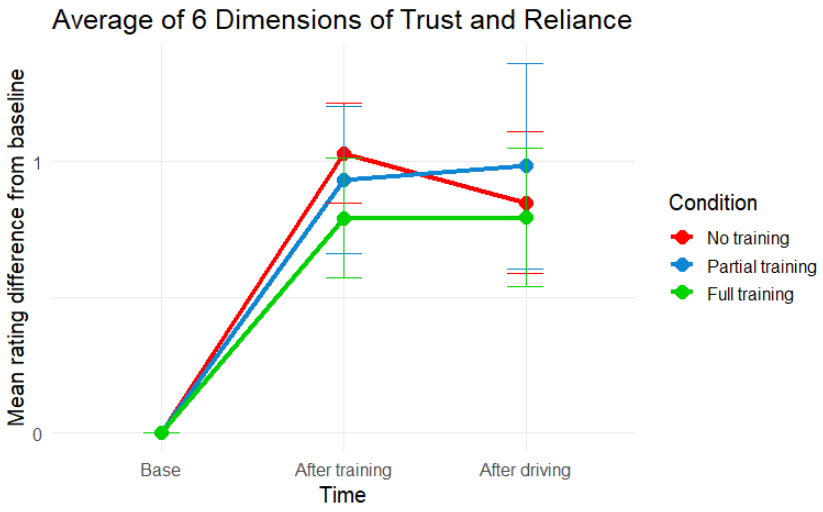


Figure 13. Mean Trust and Reliance ratings in comparison to baseline ratings both after training and after driving. In all conditions, (including partial and no social media training) trust and reliance scores improved equally

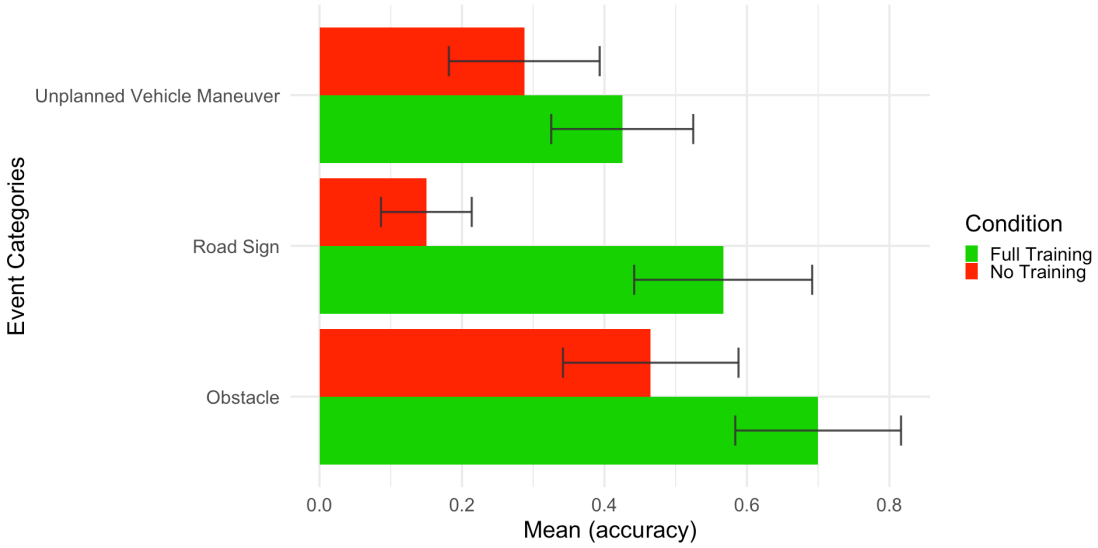


Figure 14. Mean Accuracy: Training vs. No-Training by Event Categories

5.5 Discussion

The findings of the study indicated a significant increase in subjective ratings of trust and reliance for both the control and experimental training conditions. Interestingly, the dip in

trust and reliance ratings after exposure to negative examples was not notably lower than the ratings derived from general Tesla news. But statistically the level of trust in AI remained unchanged despite the presence of negative instances in the full-training condition. This suggests that participants may have exhibited a tendency to place excessive trust in the AI system.

The training wielded a considerable influence on participants' capacity to predict forthcoming events with a five-second window prior to their occurrence. Moreover, participants who underwent training on half of the material performed at a similar level to those who completed the full training regimen, specifically concerning the pertinent aspects under evaluation. The SQA-related training also possesses the ability to furnish information that adapts to diverse road conditions, aiding drivers in retaining this valuable knowledge throughout their autonomous vehicle drive.

6 Exploring the Role of Social Forums in Building a Knowledge Base for AI: An Interview Study

The rapid advancement of semi-autonomous vehicles to near self-driving has triggered a dramatic shift in the automobile industry and transformed our perception of mobility. However, as these intelligent systems become more prevalent, understanding how users perceive and interact with them becomes crucial to ensure seamless integration and acceptance.

One of the essential aspects of studying user behavior and experiences with AI is uncovering their mental models. A mental model refers to an individual's internal representation of how a system or process works. When users interact with AI-driven applications, they form mental models that influence their expectations, decisions, and overall satisfaction. As AI continues to shape human-computer interactions, grasping the nuances of user mental models becomes paramount in designing user-friendly and effective AI applications. In the competitive market of emerging AI technologies, it is natural for these innovations to be less than 100% accurate. The process of comprehending how novice users approach such situations and adapt for long-term use can yield valuable insights for future AI users, potentially reducing the adjustment period significantly. This understanding can prove instrumental in enhancing the user experience, streamlining AI adoption, and maximizing the technology's benefits in the long run.

Interviews have demonstrated their efficacy in understanding the cognitive frameworks of individuals in several industries, notably within the realm of automation. Considering that the proposed XAI framework for autonomous driving is specifically designed for users

with little knowledge, the utilization of interviews might be beneficial in assessing how individuals without technical experience in the system can eventually offer non-technical explanations of machine learning and address any early misunderstandings as shown for intelligent office system by Tullio et al. (2007) and for an image classifier by Mamun (2021). Preusse & Rogers (2016) shown a high level of proficiency in error management for automation systems through the utilization of the interview technique, efficiently identifying and understanding errors. The process involved interpreting errors by leveraging various contextual cues, such as context, measurement comparisons, device mental models, device self-checks, consistency checks, component information, and additional pertinent data. The distinguishing characteristic of this approach lies in its ability to produce numerous interpretations for a singular error, all based on the extensive range of cues available to it. The utilization of a multi-faceted interpretation technique not only serves to improve the identification and resolution of errors, but also serves to demonstrate the extensive error management by users of automation. In a semi-structured interview focused on a Level 2 automated system, Nees et al. (2020) illustrated that mental models can be significantly influenced by the availability or absence of interface feedback, as well as the constraints experienced during interaction with the system. Utilizing the identical approach, the user's mental model within a recommender system comprises four essential steps: data acquisition, user profile inference, user profile or item comparison, and recommendation generation. These steps collectively constitute the foundational framework for all models employed in the recommender system (Ngo et al., 2020).

Within the domain of driving, the interview technique has been extensively employed to gain comprehensive insights into the relationship between drivers and automation systems.

Strand et al. (2018) has shown that the initial interaction with an automated system in the vehicle has significant importance, as an unfavorable initial experience can discourage users from further involvement. This not only impedes the potential good effects on safety but also affects consumers' perception of the system. In order to address this concern and guarantee a favorable initial encounter, it is imperative to provide customers with a meticulously crafted initial engagement with the system. A meticulously designed initial user experience has the potential to assist users in constructing a thorough cognitive framework and discourage the improper utilization of the system. Beggiato et al. (2015) have also emphasized the significance of the initial phase because it encompasses a significant portion of the learning process. For Tesla autopilot users, the acquisition of knowledge regarding the functions of the system is primarily reliant on the utilization of trial-and-error methods. Furthermore, individuals utilize self-regulatory strategies, such as establishing acceptable usage parameters and maintaining a safety buffer, to proactively avoid extremely hazardous circumstances while concurrently engaging in a secondary activity (Lin et al., 2018). The interview study also showed the users have a positive attitude towards the system.

The report's Section 3 discusses how social forums dedicated to an AI technology (in this case, Tesla FSD) have evolved into platforms where people share a wide range of information. These forums have generated a dynamic community of users eager to engage with and assist one another, from trading shared experiences to offering useful tips and tactics for overcoming AI's limits.

This study primarily seeks to gain insight via qualitative interview methods into how Tesla FSD users build their knowledge base and mental models of the system from

different sources, including friends and social media. The secondary objective is to explore the additional benefits these social forums offer to drivers, beyond simply building their knowledge base. The interview study is semi-structured (Crandall et al., 2006). So, the research question that will be addressed is,

R9: How do Tesla FSD users conceptualize the process of training on the technology when official vendor training is scarce or limited?

R10: In what circumstances did users of Tesla's FSD system resort to public forums to seek assistance?

R11: What factors drive individuals to utilize still-developing AI technology to carry on its usage?

6.1 Participants

For this study, 10 Tesla FSD users were recruited based on word-to-month (8) and social forums (2) advertisements. The participants are using the technology on an average 2.5 years and have previous experience on driving semi-autonomous technologies like adaptive cruise control, and lane assistance. The participants drive at least 5 hours per week using FSD.

6.2 Interview Guide

The interview part can be divided into two parts.

a. Training on the AI

b. Perception about the AI

First the participants were asked about how long they have been driving semi-autonomous technology, and in particular FSD. Also, they were asked how they were trained by the technology provider and how they educated themselves using social forums (like reddit, Facebook Tesla groups, twitter, YouTube, etc.).

Subsequently, the participants were presented with 10 anomalous driving situations encountered while using FSD, which had been utilized in previous simulations conducted as part of this report (see Table 11). They were then asked to share their experiences and insights related to these specific driving scenarios. Additionally, the participants were requested to elaborate on two cases by explaining whether they were already aware of those instances and whether they sought help or advice from social forums in dealing with them (see Table 12).

The participants were also questioned about instances where the AI violated their expectations, and these occurrences were assessed on a 4-point hierarchy. The interview process is depicted in Figure 15.

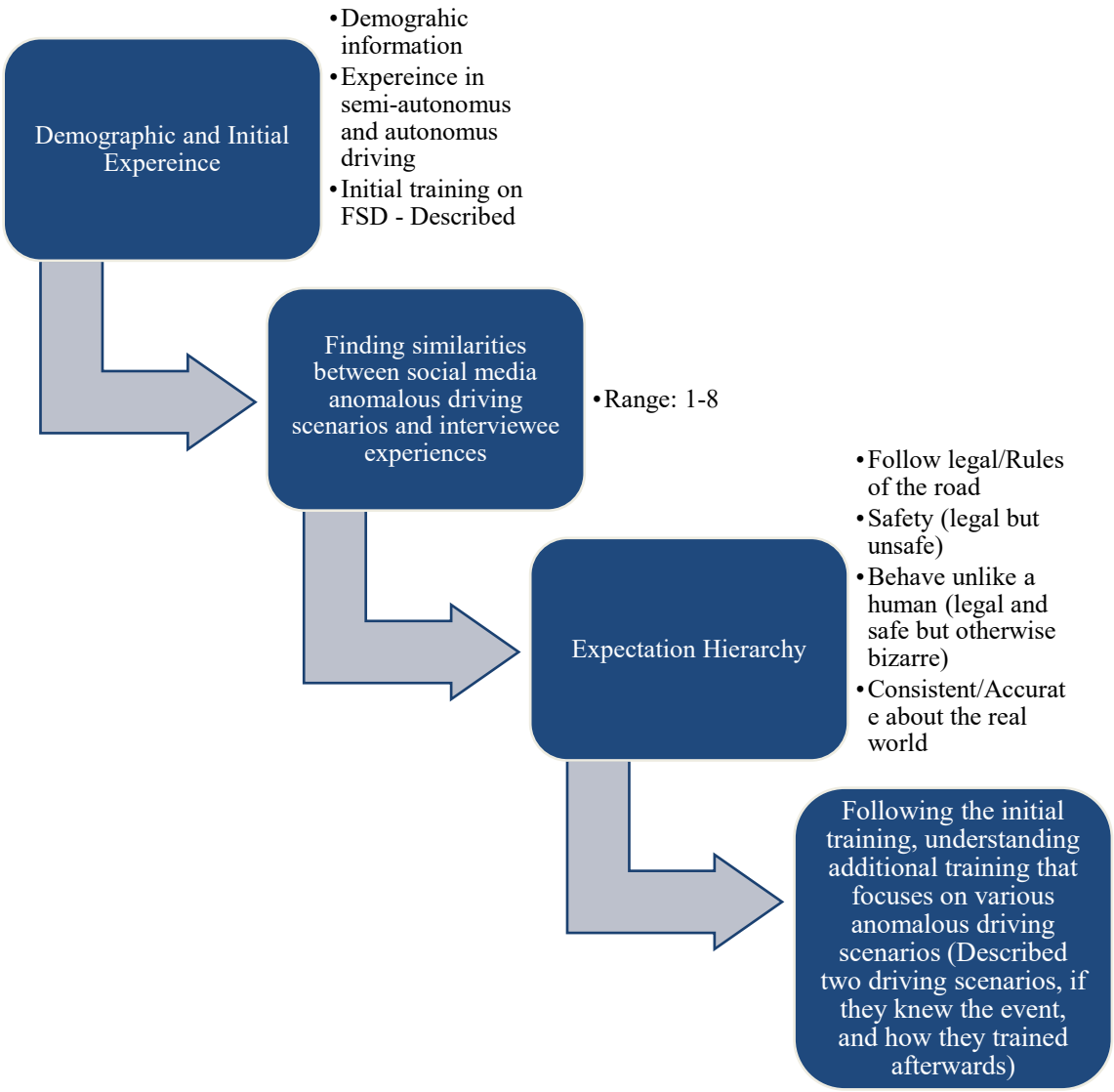


Figure 15. Interview Process of Tesla FSD Users

Table 10. Counts for the 10 Anomalous Driving Situations

Incidents	Count (Range 1-8)
Occurred Before a Rail Crossing	5
Occurred When Saw a Yellow Sign/Blinking Yellow	5
Occurred When Saw No Center Lane Marking in the Neighborhood	6
Occurred When Saw a STOP Sign	6

Did a Rolling Stop	2
Occurred When Saw Stopped Cars in the Lane	5
Occurred Saw a Road Work Ahead Sign/Crew	3
Occurred When Saw a Bump	6
Occurred When Saw a Fence/Neighborhood Gate	0
The Vehicle Suddenly Pulling Over to the Right	4

Table 11. Learning before and after an Event (For full table see Appendix A)

Events	**Learning Examples Prior the Events**	**Learning Examples After the Events**
Yellow Sign/Blinking Yellow	NA	Google Search, Experience Search/Sharing in Social Forums
Lane Merging	NA	Google Search
No Center Lane Marking in the Neighborhood	Self-Driving	Experience Search/Sharing in Social Forums
Suddenly Pulling Over to the Right	NA	Videos, Experience Search/Sharing in Social Forums
Bump	NA	Ask Acquaintances
Phantom Braking	Self-Driving	Videos
Before Rail Crossing	NA	Experience Search/Sharing in Social Forums
STOP Sign	Videos	NA
Stopped Cars in the Lane	Self-Driving	NA
Road Work Ahead Sign/Crew	Self-Driving	Experience Search/Sharing in Social Forums

While not included in the aforementioned list, phantom braking (count 4) and atypical lane changing (count 2) remains a significant concern, and numerous participants discussed their experiences with it.

The 4-point of expectation hierarchy (examples are in the Table 13),

1. Follow legal/Rules of the road
2. Safety (legal but unsafe)
3. Behave unlike a human (legal and safe but otherwise bizarre)
4. Consistent/Accurate about the real world

Table 12. Examples from the Expectation Hierarchy (For full table see Appendix B)

Expectation Hierarchy	Examples
Follow legal/Rules of the road	• Went over speed bump without slowing down • Turns using bike lane
Safety (legal but unsafe)	• Fast turn • Sometime accelerates over speed limit
Behave unlike a human (legal and safe but otherwise bizarre)	• Phantom Braking • Jerky auto-park
Consistent/Accurate about the real world	• Couldn't recognize the speed bump • Blinking Yellow mistaken as Traffic Light

Table 12 showed the engagement in social forums varies for users of the AI. Form of engagement are.

1. Look for solution,
2. Learn others experience (non-solution),

3. Post one's experience and tricks to avoid abnormal driving situations.

This type of engagement is further explored in the next section. In this phase, the coding process was conducted using the training materials and procedures employed by the participants before and after their driving sessions. Moreover, the hierarchical scale serves as a coding manual for effectively categorizing the expectations derived by the artificial intelligence system. The quotations utilized by the participants reflect the broader trend within the context.

6.3 Training on the AI

The participants were asked about their source of first training on a new AI system in this case FSD. Purpose of this section is to understand in what sources the drivers of a new AI go to if the initial training source from the vendor is limited. The FSD is perfect for the situation as all the participants reported they got limited support from the vendor (short walkthrough or release note for new version of FSD). Table 12 has been employed to gain insights into the training situation that emerge when a driver progresses beyond their initial training phase. For a new AI system, the external vendor training is sorted into three distinct groups; the key determinant for the segmentation of the group is the incorporation of collaborative activities (or lack of it),

A. In-person training – training from self-driving, acquaintances;
B. Collaborative online training – training from social forums through ask and answer, looking into others experiences;

C. Non-collaborative online training – training from video platforms where active collaboration is not needed, search using keywords.

As drivers starts as novices, they typically acquaint themselves with the system through the aforementioned approaches. These methods serve to educate them about the AI's features, as well as its strengths and weaknesses. However, when they advance and achieve a fundamental level of technical proficiency, their training style changes to a case-based approach. In order to further hone their skills, they start concentrating on particular real-life circumstances and scenarios experienced by others. This change in training process reflects their growing technical knowledge and comprehension of the nuances of the technology.

A. In-person training

People who are familiar with the technology, frequently close friends or colleagues who are already users, are the main sources of training and guidance for using it for the new users.

'I always like to try new things, but in this case, nobody explained how this thing works, I wasn't comfortable sort of giving away control as I used to be more novice, and then I think I went to a colleague or friend who had a Model X, I went and drove with him few times and saw how he was using it, and then I started to become more comfortable.' (p1)

This training covers a range of topics, including introducing the system, providing guidance on whether to activate or deactivate it depending on specific events. This training assists novice users in grasping tips and techniques for specific situations where they typically used to take control, but now they perform actions that prevent

seizing control. For instance, they might learn to gently push the wheel during certain events to avoid triggering a takeover, instead of immediately assuming control as they used to.

'The main thing I sort of learned from my colleague, see this whole thing about keeping your hands on the steering wheel. And making sure you're applying. A slight pressure on the in the opposite direction. That took me a while to figure out because I used to turn it off.' (p1)

The training is also done in group training sessions. These sessions allow a close-knitted community of new users to interact, share knowledge, and improve their collective comprehension of the functionality and best practices of the technology.

'There was three of us that got the car pretty close together….and three of us bouncing ideas of each other, I only got the FSD, rest of them had autopilot. So, we were talking about autopilot' (p8)

'Sometimes we have discussions (regarding issues about FSD) in our groups as well' (p9)

This training methodology also has the potential to aid users in efficiently managing errors by providing drivers with access to relevant data from various diverse sources.

Self-training is another method new users of the technology apply to become proficient. Curiosity is a common driving force in this process. The next quotation comes from a driver who wants to test the vehicle out of curiosity.

'Beyond the intersection was basically like a core points, you know, concrete divider that so the car found itself heading towards because it deviated from the

lane…. I was out. At about 2:00 in the morning to make sure that (no traffic was there)' (p2)

Drivers actively investigate and evaluate the numerous scenarios in which the AI succeeds or fails as they travel. They are able to gradually identify the AI's advantages and disadvantages thanks to this hands-on approach, empowering them to decide when to exercise control and when to rely on the technology.

B. Collaborative online training

Collaborative online training comes from different online social forums like reddit, Facebook, twitter, etc. Drivers read about an incident they face or post notes if they did not find it in their desired forum.

'Even though I have a microscopic Twitter following, I did post a couple of my experiences like brief notes about those' (p2)

Within the social forum, individuals express their perspectives, leaving each driver to make judgements on the correctness of a perspective based on their experience. Social forums also help drivers to gathers various individuals' accounts of how they dealt with different iterations of an adverse driving condition.

'I want to see how to handle this (a specific problem). And so, I go (to the forum) and see experiences on how it was handled (different versions) of this problem… What it does it gives me a better understanding of the things the car can do and the things the things the car can't do' (p7)

This methodology has the potential to induce a transformation in the initial perception or comprehension of the system. As individuals accumulate further exposure and familiarity with the system, they may undergo a process of revising

and enhancing their initial notions and beliefs pertaining to its functioning and potential accomplishments.

C. Non-collaborative online training

Primarily, this training occurs via diverse video platforms such as YouTube, although interaction among drivers is constrained on these platforms. However, these video platforms serve as the primary online training resource. Here, drivers encounter a variety of online profiles that furnish feature lists and offer insights into tackling various challenges.

'Used YouTube mostly. To mostly figure out things that I can't figure out like how to open the glove compartment or how the phantom breaking works or how the most efficient way to drive' (p9)

While this medium is valuable, it occasionally falls short in addressing tailored issues or get a second opinion on a specific issue regarding the AI. To address this gap, drivers turn to social forums for more customized problems.

'Yeah, certainly. I mean, YouTube certainly has a lot of information about the functionality of FSD. There are some relatively long-time testers, some of whom have, you know, relatively consistent over the weekend. I guess look to it for a reasonably coherent snapshot of the current state of the system… And the ones that I would primarily participate in are Tesla Motors Club forum. It's a good place to go for sort of a counter reaction.' (p2)

Nonetheless, these video platforms excel in presenting an array of typical problems and corresponding strategies for circumventing them.

The drivers might just search via google keywords to learn generally or for specific situations.

'I've done a lot of Google searches and stuff since then about it (FSD)…. Search most of the time, my search was just Tesla FSD latest version or something like that, I was not looking something specific.' (p3)

Anxiety prompts individuals to seek information or concentrate on specific situations, such as anticipating how the vehicle will respond to an event.

'(Google) searches on how do you turn it on or turn it off and what does it do if I break or don't break, like just looking at specific situations where I worry about something, then I go do a search and see' (p1)

6.4 Perception about the Technology

To get the best system performance, it is crucial to gain understanding of how AI works, know its capabilities, and maintain an acceptable level of confidence (Garcia et al., 2022). While AI holds the promise of widespread integration, a noteworthy challenge to its broad adoption lies in how users and customers perceive emerging technologies and innovations. User acceptance of a product may be correlated with the perception of trust (Reynolds & Ruiz De Maya, 2013). In the context of autonomous vehicles, placing substantial trust in automation hinges largely on the vehicle executing accurate actions at the appropriate junctures. In-depth research has extensively explored the acceptance of technology, analyzing attributes such as trustworthiness, usability, and utility. Previous encounters with AI have also been observed to influence the willingness to adopt it (Hengstler et al., 2016; Hoff & Bashir, 2015; C. Oh et al., 2017). But trust in automation can be lost quicker than

it is regained, as demonstrated by Wiegmann et al. (2001). Researchers also looked into AI's ability to categorize different images during driving and human's perception on AI's capabilities regarding these images . In addition to human perception of AI's environment classification, my perception analysis will encompass three supplementary categories pertaining to safety and human-like behavior. This perception analysis is conducted through the assessment of anticipated violations performed by the vehicle. The uniqueness of this analysis lies in its focus on experienced drivers who have already embraced AI, as opposed to novice drivers who still have the option to reject it. By concentrating on experienced drivers, the study delves into the perspectives of those who have gained familiarity with AI integration and have chosen to accept it as part of their driving experience. This approach offers valuable insights into the factors that contribute to successful AI acceptance and utilization, providing a deeper understanding of the dynamics at play among individuals who have chosen to adopt this technology. Table 13 provides insight into the specific instances that serve as illustrations of the Expectation Hierarchy, which in turn contributes to the analysis of the perception of the AI system.

A. Follow legal/Rules of the road

According to participant replies, the FSD technology has been meticulously developed to provide a high degree of precision and sophistication in upholding road rules.

'It's a Boy Scout. It (FSD) really follows the rules' (p7)

However, a small number of individuals made a significant observation. An example was provided in which the lane-changing behaviors of the FSD system were described as somewhat erratic or sudden. This observation was particularly evident when the system encountered a wide bicycle lane in a rural setting (which may be due to the more seldom

occurrence of such generous bicycle lanes in comparison to conventional road lanes) and when it needed to modify its velocity to align with the prescribed speed limit.

B. Safety (legal but unsafe)

The occurrence of irregular turns, which have the potential to cause confusion among following vehicles, has garnered the attention of drivers as a singular concern. This problem pertains to scenarios in which the autonomous vehicle's navigation during turns has the potential to either raise or reduce its speed, hence causing confusion among other vehicles on the road. In addition to this concern, the vehicle adheres to established standards of safe driving.

'the right turn, it just staying there and then suddenly taking a right, some guy coming behind me, he might think, I'm going straight and he might go to the right' (p1)

C. Behave unlike a human (legal and safe but otherwise bizarre)

The implementation of FSD technology has encountered numerous challenges due to this particular attribute. One of the concerns expressed by a study participant in various driving settings was the possible risk associated with this specific behavior, as it has the ability to fool other human drivers. In densely populated urban regions, it is atypical for a vehicle to change lanes later than expected due to congestion. However, vehicles equipped with FSD capabilities occasionally exhibit this behavior. Under rule of the road, the activation of the turn signal is employed as a means to communicate an intention to merge. Nevertheless, the successful execution of this maneuver necessitates the gradual deceleration of the vehicle in preparation for the turn, accompanied by a cautious forward movement and diligent observation for any oncoming vehicles. Regrettably, the smooth and uninterrupted execution that characterizes skilled human driving is lacking in FSD's behaviors.

FSD system demonstrates a notable escalation in its vigilance when confronted with scenarios whereby a trailing vehicle possesses the ability to pass the FSD-enabled vehicle. The heightened level of awareness exemplifies the system's recognition of potential hazards and its prioritization of safety in such situations. In another scenario, the vehicle may increase its speed in order to overtake a truck expeditiously. Nevertheless, in the scenario the participant took measures to ensure that they keep a prudent distance and prevents the situation from escalating to a level where there is a potential for jeopardizing the safety of individuals involved.

Even the operators of FSD systems occasionally encounter instances where the system's actions are perplexing to them. One participant noted that there are occurrences in which the FSD system appears to adhere to a predetermined navigation route before suddenly indicating a desire to make a right turn, deviating from its original trajectory. The following statement can summarize the findings,

'It could be confusing to other drivers and therefore unsafe, it does things like I said, like maybe shifting lanes in the middle of the turn, being too cautious at an intersection, like inching forward too slowly and taking too long' (p7)

D. Consistent/Accurate about the real world

The visual representation exhibited on the FSD panel often provides a realistic depiction of the immediate surroundings.

'In general, it's really does a pretty good job of seeing the world and like it knows the difference between a person and a bicycle' (p8)

However, the system deviates from this typical behavior under many instances. The occurrence of "phantom braking" is a prominent illustration. This phenomenon occurs

when the Forward Collision Warning (FCW) system unexpectedly initiates braking due to its misperception of obstacles ahead, despite the absence of any actual barriers, or misinterprets a blinking yellow light as a signal to come to a complete stop. While infrequent, users have observed this occurrence and it may sporadically induce discomfort.

'so the blinking yellow, there have been a lot of trouble with it, it sees them as almost a red light, so it starts to slow down for them' (p8)

In a different scenario, a driver discovered that the FSD system relies on map data, rather than sensors, to determine the speed limit.

'I think sometimes though, it really needs to be like humans and be able to read the speed limit signs because sometimes the map data isn't right, and it'll be doing the wrong speed limit. You know, and so it thinks it's doing the right thing, but the posted speed limit is different than what the map data has' (p8)

These types of explanations align with the finding of Tullio et al. (2007), which highlight the ability of non-technical individuals to offer explanations pertaining to machine learning.

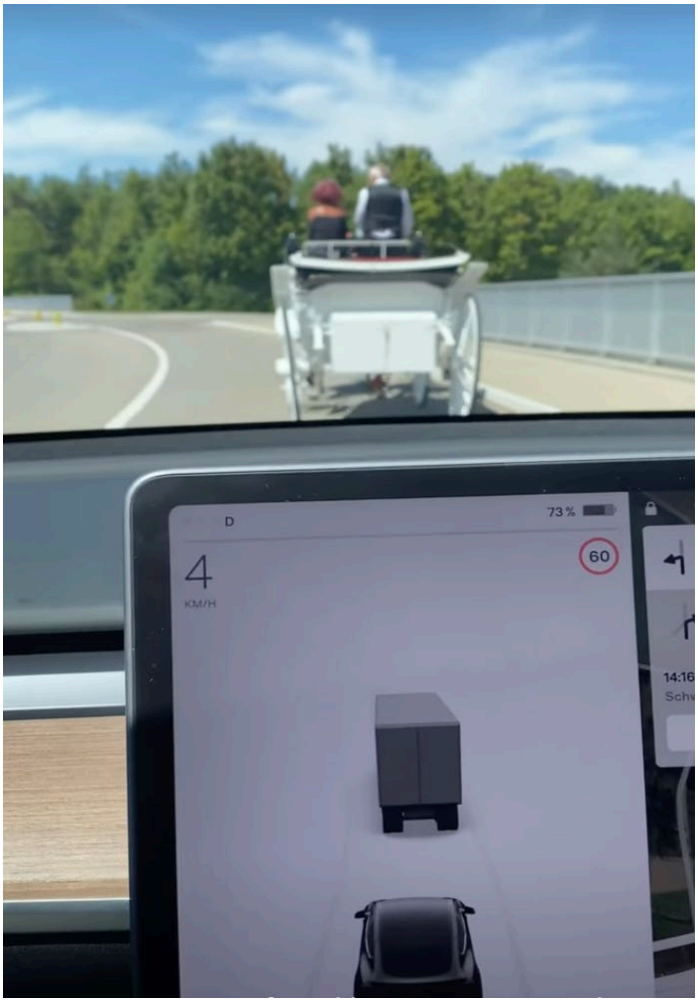


Figure 16. An illustration: AI exhibits inaccuracies in its understanding of the world

Furthermore, unusual environments like garages provide difficulties for the FSD's perceptive abilities. In such situations, the system can wrongly classify everyday items like washers and dryers as larger automobiles like trucks – see the example in Figure 16 for an instance where Tesla's AI transformed a carriage into a truck/lorry. This error results from the difficulties of effectively identifying items in constrained or unfamiliar locations, and it reveals a shortcoming in the system's capacity to discriminate objects of different sizes and functions in such situations. Even though the accuracy of the FSD system is generally

outstanding, these situations show how much further development is needed to improve its ability to perceive and make decisions in a variety of real-world situations.

6.5 Discussion

The interviews revealed a common pattern where vendors frequently omit discussions about the vulnerabilities of new AI systems. As a result, inexperienced drivers often turn to those they know well, who have already possessed the system for a period, as their initial source of insight into the AI system. The involvement of close acquaintances serves as an essential starting point for novice drivers to acquire knowledge about the AI technology. Collaboration plays a pivotal role in this learning process, as novices form small groups to exchange insights, techniques, and strategies related to the AI system, thereby bypassing the necessity for expert guidance. This significant discovery provides compelling evidence supporting the utilization of the CXAI system (Mamun, Hoffman, et al., 2021), particularly in scenarios where inexperienced users commonly engage in collaboration to familiarize themselves with AI technology.

The online video platform, primarily YouTube, holds a crucial role in educating novice users about AI, covering a broad range of common topics. However, AI systems can sometimes present intricate challenges that are not universally covered in general tutorials. In such cases, drivers encounter custom or unique issues that necessitate tailored solutions. This is where social forums come into play. These online communities provide a space for users to discuss and troubleshoot AI-related matters that might be specific to certain configurations, settings, or scenarios. By tapping into the collective knowledge of these

forums, drivers can find more specialized advice, insights, and workarounds to address their particular AI challenges.

The level of acceptability of AI technology is primarily determined by the perceived safety of its activities when in use. Even when the AI system exhibits behavior that is distinct from human-like responses, users are still inclined to utilize it. This shows that, while users may notice non-human elements of AI behavior, such distinct decision-making patterns, they are ready to overlook these differences provided the technology continuously demonstrates to be safe and abiding by established laws for the sector. This highlights the importance of having a solid foundation of safety in AI systems in order to foster user confidence and general acceptance.

7 General Discussion

The dissertation lays the foundation for establishing a social forum or CXAI as an XAI platform within the autonomous vehicle domain. The incorporation of the literature review and Detection Study has significantly influenced the formulation of diverse experimental approaches during the course of the evolution process. The discussion is divided into several sections, with the aim of structuring the discussion in a manner that promotes greater organization and focus.

7.1 Key Insights Derived from the Research

The research serves as the foundation for adopting a non-algorithmic approach as proposed by S. Mueller et al. (2021) in the field of autonomous driving. So, this report aims to examine several aspects related to the adoption and usage of the proposed XAI framework. Specifically, it seeks to investigate the time required for individuals to 'take-over' based on the training (based on social forum), assess the communication in social forums, explore the potential for collaboration among inexperienced users in understanding autonomous vehicle actions using the social forum option within a non-algorithmic framework, and analyze the persistence of autonomous vehicle operators in utilizing AI technology despite its non-human attributes.

The report begins by analyzing content related to explanations, validating whether these explanations assist in identifying anomalous driving situations. Additionally, it explores the current state of perception and training processes for AI in cases where vendor training is absent and evaluates the inclusiveness of social forums in such training.

The primary and noteworthy finding pertains to the level of engagement in collaborative problem-solving endeavors, such as the clarification of inquiries or the evaluation of actions undertaken in prior situations, within social platforms. The phenomenon described is examined within the context of the Framing-Reframing dimension in the field of **Communication Analysis**. The present discovery addresses the matter of initiating interaction, a topic that was highlighted in Mamun (2021)'s study on collaborative endeavors in the field of non-algorithmic CXAI.

The results of the **two Simulation Experiments** revealed another significant discovery within the realm of automated vehicle situations. They showed that simplified off-line explanations can be utilized to accurately identify atypical driving incidents, without the need for intricate real-time algorithmic XAI systems.

Based on the training derived from the communication corpus for several abnormal driving scenarios, participants demonstrated improved reaction times and heightened sensitivity in detecting unusual driving situations. However, it's noteworthy that their responses were generally reactive rather than proactive. This may have stemmed from a sense of security in the simulated environment. This led us to develop the **Prediction Study**, which demonstrated that the training had a significant impact on the participants' ability to anticipate problems five-seconds before they occurred. For the training itself, the users perceived it was successful in training them on the automation system. But the dip in trust and reliance ratings after exposure to negative examples was not notably lower than the ratings derived from the control condition (general Tesla news). This discovery somewhat reinforced perspectives held by actual FSD users in real-world settings in the final **Interview Study**, as users often made a distinctions between trust and reliance for actions

of the AI. For example, many drivers have indicated a significant dependence on Tesla's FSD system, and one referred to it as a "boy scout" because of its rigorous compliance with traffic laws and regulations. Nevertheless, the level of confidence in the FSD system's ability to execute all driving tasks with the same level of proficiency as a human driver (e.g., not slowing down for speed bumps, moving to the central lane when there is no mark) appeared to fluctuate.

This observation suggests that FSD users often have confidence in the system's ability to follow established rules, implying that they rely on it to handle routine tasks effectively. However, their trust may diminish when it comes to the system's capacity to mimic the nuanced decision-making and adaptability of a human driver, highlighting a distinction between reliance on FSD for rule-based actions and trust in its human-like driving capabilities.

The **Interview Study** also suggests AI's non-human traits may not hamper the use of the autonomous vehicle. The absence of vendor-provided training for the AI system underscores the notable role of social forums in obtaining information about the AI system. This is particularly relevant during the initial stages of user training and as user knowledge continues to advance.

As part of the dissertation, an examination is conducted on how a novice driver's mental model evolves during training for an AI system in the autonomous vehicle domain. Collaboration is observed throughout the early phase of training, wherein assistance is sought from acquaintances or through discussions within groups sharing similar experiences level with the AI. This is consistent with the principle of CXAI, which emphasizes the need for collaborative explanation.

7.2 A Concise Overview of Development and Usage of CXAI

Through a thorough examination of the communication factors, the dissertation sought to get a deeper understanding of the efficacy of the social forum as a medium for facilitating engagement and cooperation among its participants as an XAI system. The analysis of communication revealed a notable presence of problem resolutions and shared experiences within the communication corpus. These findings also highlighted the existence of motivating factors for engagement, particularly related to negative emotions regarding AI interactions. Notably, the shared experiences represented a substantial portion of the communication corpus. As indicated in Table 12, it was intriguing to observe that many expert drivers actively sought out the experiences of others to learn from different perspectives. This could also foster a sense of reassurance, as these shared experiences highlighted that they were not to blame for the AI's actions. This emphasizes the importance of communal experiences in discussions concerning AI.

As per Mamun, Baker, et al. (2021), a noteworthy discovery arising from their study regarding explanation in CXAI system (see Figure 17 for the CXAI System) pertains to the investigation of several categories of explanations. Upon their examination utilizing the framework proposed by Lim et al. (2009), it becomes apparent that a significant proportion of the explanations provided can be classified into the 'what' category. Upon doing a more thorough examination of the extent of these explanations, it was noted that these explanations, characterized by the use of 'what'-style questions, predominantly contained overarching patterns that spanned across numerous photographs for an image classifier. In

contrast to the majority of XAI algorithms, which are commonly developed to handle specific 'why'-oriented queries by offering reasons for individual decisions. As a result, a CXAI system may not provide the same information as traditional systems, but it has the potential to enable different modes of understanding.

In the realm of autonomous driving, we find a similar trait in the social forums, where resolutions from social forums are typically straightforward and lack deep introspection, in contrast to the more elaborate 'Why'-type explanations created by XAI algorithms.

But the resolutions in social forums might be associated with specific discussion topics or triggers (S. T. Mueller et al., 2019) as depicted in Figure 17. The provided explanations serve as a practical guide for users to navigate atypical driving situations, outlining the recommended course of action.

While the resolutions offered in social forums may appear somewhat arbitrary, they nonetheless give rise to tailored issues or queries. Based on the analysis of the interviews and a comprehensive body of observations in the area of communication analysis, it can be inferred that individuals frequently express a need for guidance on appropriate responses when faced with unfamiliar or unexpected situations while operating autonomous vehicles. This may encompass a range of issues that can be solved to guarantee safety or minimize potential hazards. Furthermore, it may be of importance to users to have a comprehensive grasp of the limitations and constraints that define the capabilities of the autonomous system. For instance, individuals may express interest in determining the vehicle's capability to execute specific maneuvers or inquire about circumstances that may prompt the disengagement or necessitate human intervention inside the system. The inclusion of these specific details is of utmost importance for consumers seeking to make well-informed

judgments regarding autonomous driving technology, as well as for those seeking to anticipate the performance of said technology in diverse scenarios.

The findings from both user and simulator experiments indicate that persons who use CXAI or a social forum are able to understand the system and in the simulator studies, effectively utilize its knowledge to predict aberrant driving scenarios ahead of time.

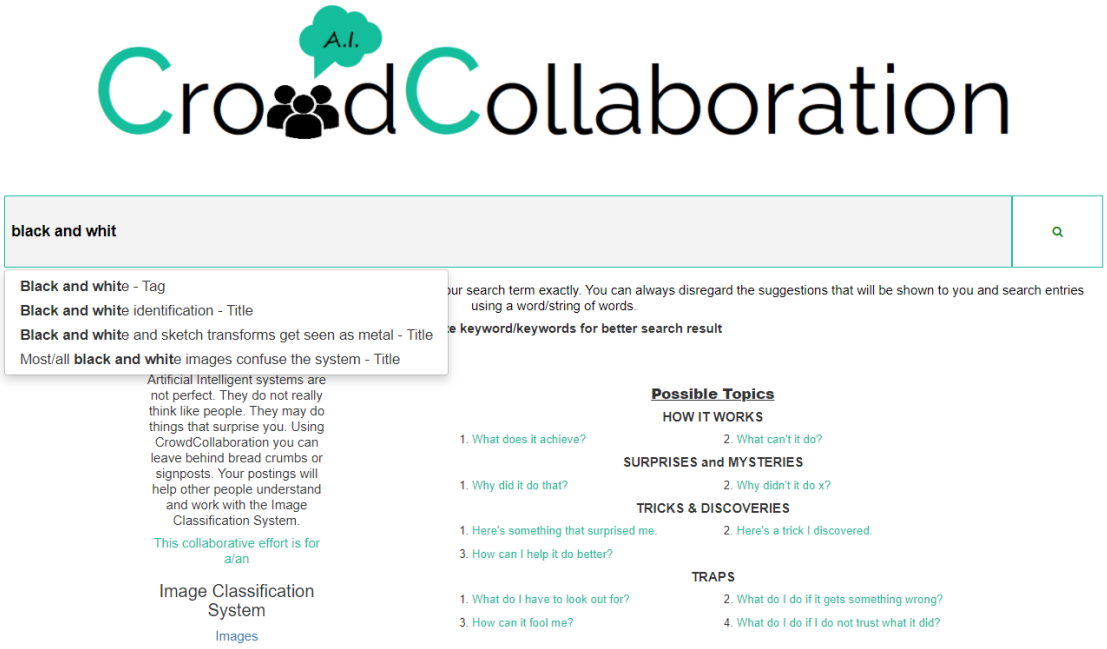


Figure 17. A CXAI system (Mamun, Hoffman, et al., 2021)

Figure 18 illustrates an orderly representation of the development and utilization of the CXAI.

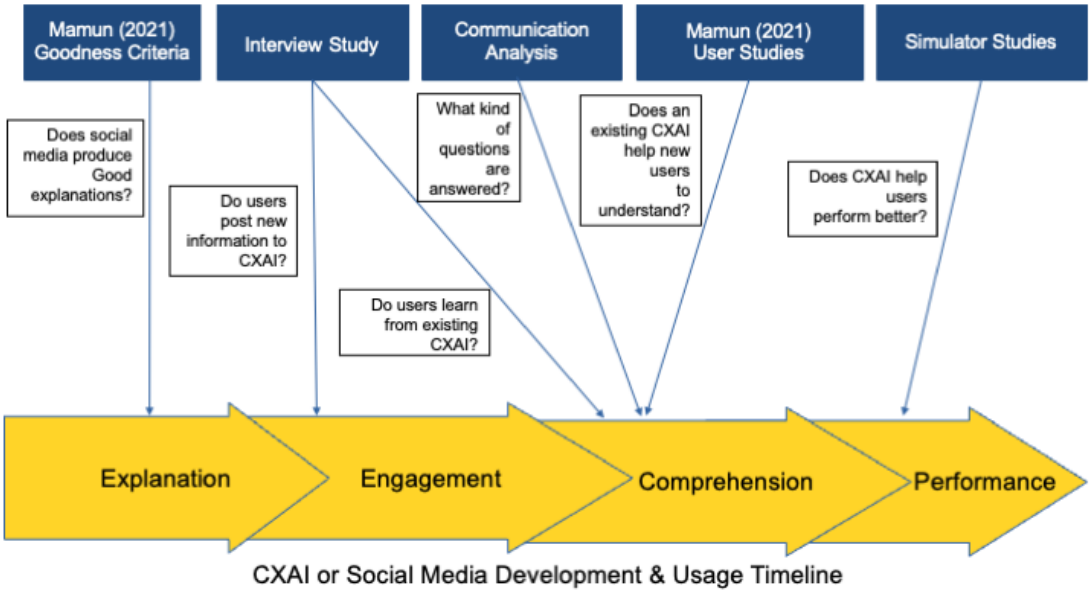


Figure 18. Timeline and Objectives of Research Relating to CXAI

7.3 Individual Learning Sources Integrated into a Social Forum

Upon analyzing Appendix A, it becomes evident that users not only make use of social forums but also depend on three other sources for AI training: acquaintances, video platforms, self-driving, and web searches. When we view the social forum as the ultimate platform for knowledge sharing and recognize these other sources as the primary avenues for gaining insights about the AI, it becomes clear that these sources can indeed play a crucial role in spreading knowledge within the social forum. The paradigm given in Figure 19 illustrates a social interaction, integrating many sources of knowledge inside a social forum. It is part of e-Learning that can be characterized as the amalgamation of various learning methods that leverage computer technology to facilitate the learning process (Tavangarian, 2004).

The learning process under the model mainly depends on individual learning, wherein knowledge, skills, or information are acquired by individuals through self-directed efforts. Though the prominent characteristic within the CXAI concept is the significant emphasis on collaboration, particularly seen in online social forums where users of AI engage in extended discussions, sharing their perspectives and observations regarding the AI system. Nevertheless, it is crucial to acknowledge that the efficacy of these relationships is significantly dependent on the initial self-directed endeavors undertaken by the individuals involved.

Gregorc & Ward (1977) categorized individual learners into four categories; **Abstract Sequential** - Individuals with an abstract sequential learning preference exhibit a high level

of proficiency in decoding many forms of symbols, including written, verbal, and visual representations. Individuals who exhibit this preference demonstrate a vast cognitive repository of conceptual pictures, which they utilize to evaluate and synchronize information they come across via reading, hearing, or observing visual and graphical representations, **Abstract Random** - notable for their strong emphasis on the study of human or system behavior and their outstanding ability to perceive and grasp tiny cues or stimuli. Individuals in this group exhibit an elevated level of perceptiveness towards the subtleties present in their surroundings and the dominant emotional atmosphere, **Concrete Random** - individual with this learning technique possess this tendency demonstrate a rapid comprehension of fundamental concepts and exhibit exceptional aptitude in effectively navigating unstructured problem-solving scenarios, hence facilitating creative breakthroughs, and **Concrete Sequential** - characterized by a proficient ability to acquire knowledge through direct engagement in practical activities and firsthand experiences.

The Socio-Technological Learning Model, which incorporates a Social Forum, caters to diverse learning preferences as outlined in the aforementioned categories. Self-directed learning promotes active participation, akin to the act of operating a vehicle, wherein learners have the opportunity to acquire cues and insights through their activities. This process of self-learning can be a topic of discussion within a peer group, where peers can demonstrate exceptional skills in effectively addressing unstructured problem-solving scenarios based on fundamental concepts.

While there is currently a lack of scientific evidence definitively confirming the efficacy of a particular learning style (Bishka, 2010; Deng et al., 2022; Glenn, 2009; Kirschner, 2017; Nancekivell et al., 2021; Rohrer & Pashler, 2012), the analysis shows a pattern in

individuals' learning preferences, future research might provide valuable insights into the factors influencing their propensity towards one strategy over another to learn about an AI system in autonomous vehicle.

By incorporating a combination of social and technology-based learning approaches in this model, a wide range of learning resources can be effectively incorporated into the social forum. Peers are frequently exposed to a diverse range of media formats, thereby augmenting their ability to comprehend and evaluate different forms of information.

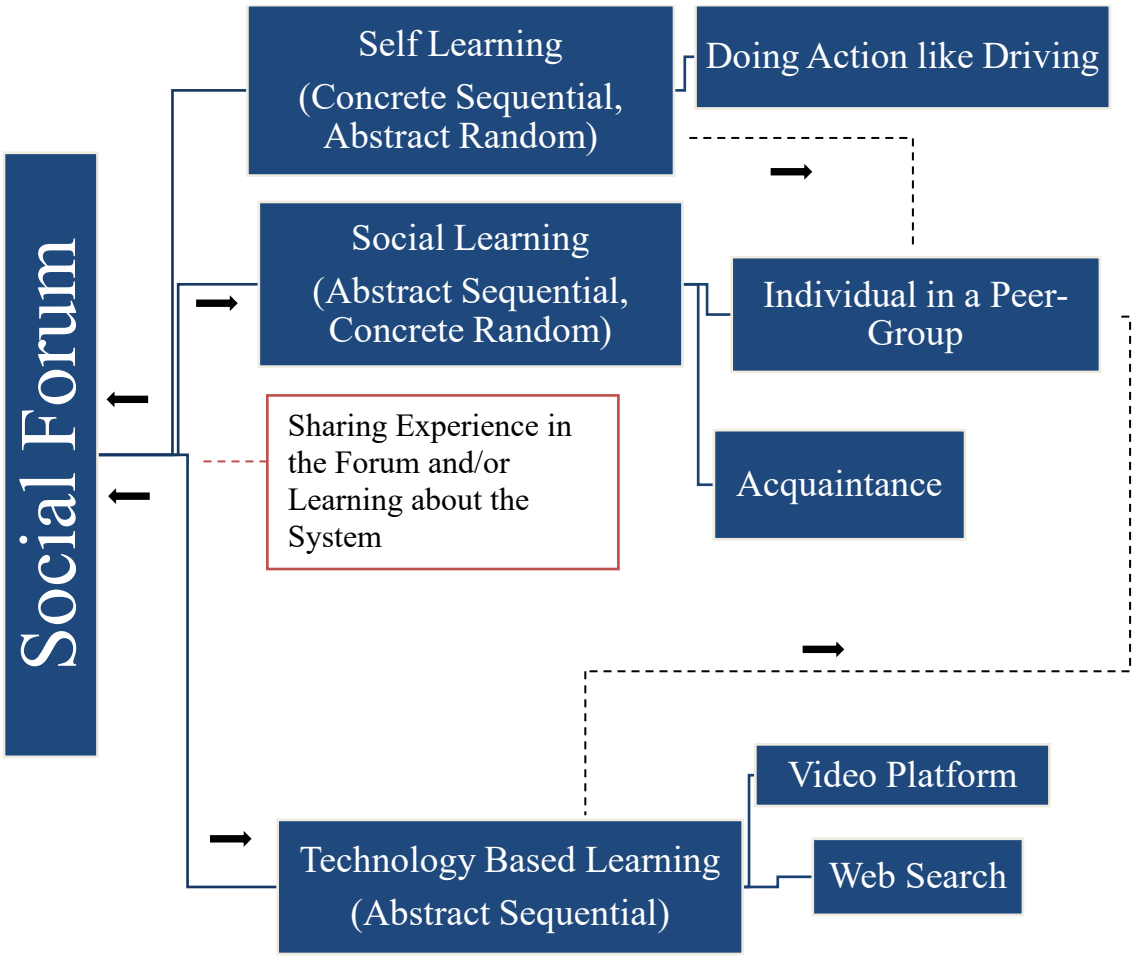


Figure 19. The Socio - Technological Learning Model inside a Social Forum - the model shows a bi-directional communication between different categories of individual learners

7.4 Potential Implementation of CXAI

The research conducted in Collaborative Explainable AI, as detailed in both this report and the report by Mamun (2021), demonstrates the potential of using CXAI or social forums to generate effective explanations about AI systems. These explanations serve the dual purpose of enhancing understanding of the AI system and facilitating user learning about AI. Since the lay users are already using this approach for different AI systems (most notably for Tesla FSD), the companies can adopt this approach by integrating CXAI practices into their AI systems and user interfaces. Given the architecture-agnostic nature of this approach, there is no necessity to consistently update the XAI system with each iteration of the AI, hence alleviating any supplementary cost burden. Incorporating the insights from the two reports mentioned earlier in this paragraph, this report will propose a framework for the successful implementation of a CXAI system, as illustrated in Figure 17. The components of the Framework can be classified into two distinct categories, as seen in Figure 20.

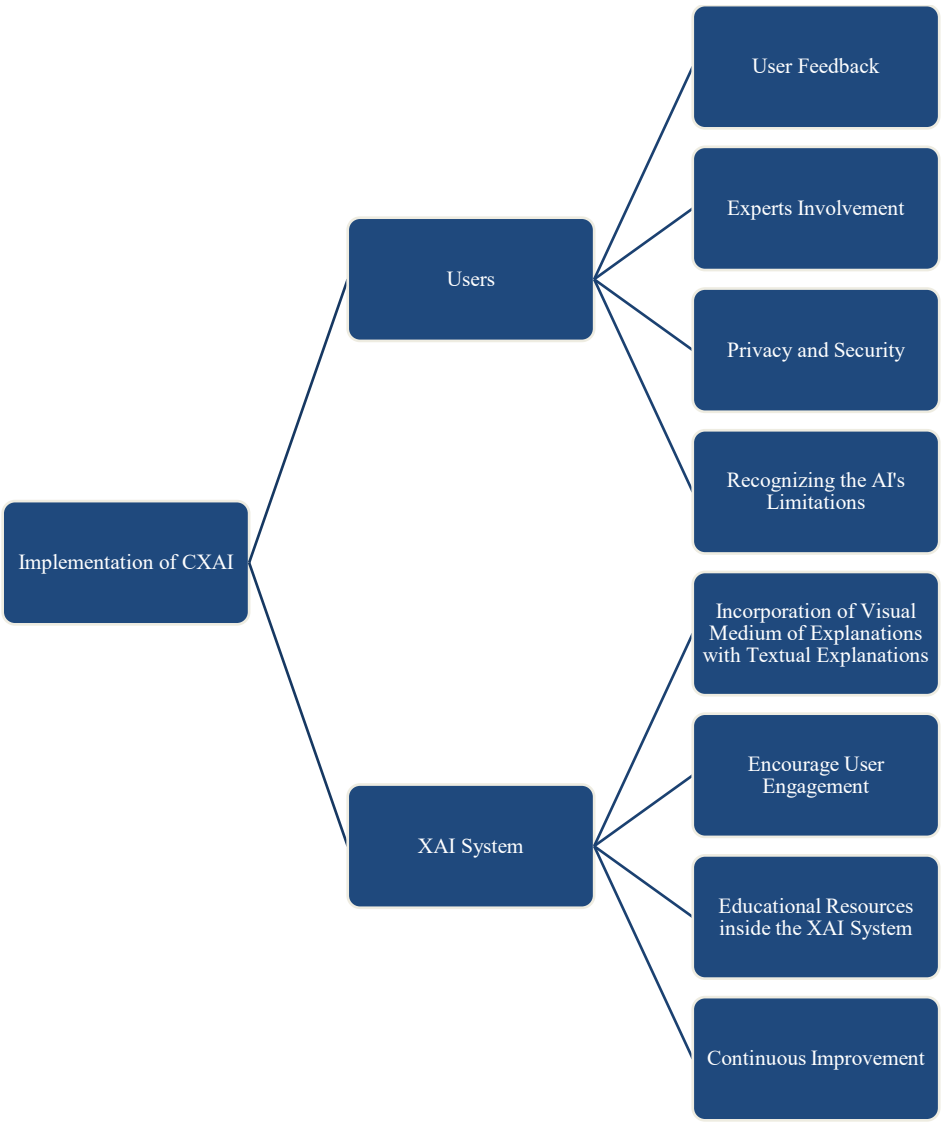


Figure 20. CXAI System Implementation Framework for an AI system - The implementation can be divided into two primary categories: one pertaining to human involvement, specifically lay-users, and the other pertaining to the XAI system

A. Incorporation of Visual Medium of Explanations with Textual Explanations:

The study conducted by Mamun, Baker, et al. (2021) emphasizes that explanations in CXAI systems and social forums are commonly communicated through natural language and are designed to be easily understandable for users. Nevertheless, the

results obtained from the interview study conducted in this report provide valuable insights into a noteworthy discovery: The utilization of video platforms, particularly YouTube, is a prevalent practice among FSD users, as they largely depend on these channels for both their initial and continuous training requirements. This highlights the crucial significance of visual media in the process of explanation.

Experienced drivers frequently utilize these instructional videos as a platform to recount their personal experiences with the technology, thereby offering practical demonstrations of their interactions with it. These demonstrations provide significant value for inexperienced users due to their provision of real-life scenarios and user perspectives, which contribute to the establishment of confidence and dependence on the material.

Moreover, the integration of a reference system within the CXAI system can be advantageous. This approach facilitates the sharing of connections to diverse websites and information by non-expert users, thereby reinforcing their individual ideas and beliefs. This practice not only cultivates a sense of community and cooperation but also empowers users to make valuable contributions to the shared repository of knowledge.

The utilization of visual media beside textual explanations, such as instructional movies, is of great significance in the elucidation of AI systems, specifically in fostering user confidence and comprehension. Furthermore, the incorporation of reference systems facilitates the dissemination of information and the integration

of user experiences, so fostering a more comprehensive and knowledgeable user community.

B. Encourage User Engagement:

If AI users fail to participate in social forums or a CXAI system, the XAI system may suffer from a dearth of significant and important content. To address this challenge, it is essential for the company to proactively stimulate user engagement across a diverse spectrum of experience levels, considering that different users may possess distinct perspectives based on their level of interaction with the AI.

Nevertheless, it is important to acknowledge that the conventional incentives commonly observed in social forums, such as expressions of anger or frustration towards the AI, may not be as widespread in the context of a corporate forum. Individuals may exhibit reluctance in openly articulating unfavorable opinions regarding the AI system employed by the company. To overcome this barrier and encourage active participation, an organization or company can use a point-based incentive structure, similar to the approach utilized by StackOverflow as expounded upon by Berger et al. (2016). This system would reward users for actively contributing by posting relevant topics related to the AI and engaging in discussions by commenting on other users' posts.

By implementing the proposed point incentive system, the organization can establish a conducive and motivating atmosphere wherein users are encouraged to actively contribute their perspectives and apprehensions regarding the AI system. This methodology not only facilitates user involvement but also contributes to the

accumulation of useful information and feedback for the purpose of improving the AI system and its explanations progressively.

C. Experts Involvement:

Experts possess the capability to assume a crucial function in rectifying erroneous beliefs and resolving misinterpretations inside a social forum. Their intervention is notably more effective than that of non-expert (Walter et al., 2021). They can also be tasked to review the analysis made by non-experts (Bowman et al., 2022). In the early stages of a forum, experts can play a significant role by offering informed solutions to complex inquiries pertaining to the machine-learning components of the AI. However, the employment of experts can be costly, and a continuous reliance on their expertise can become burdensome over time.

However, social forums provide a clear benefit in this aspect. Novice users of the AI system are empowered to transition into experts by leveraging the extensive knowledge shared by existing experts through the material they upload. As humans interact with the AI system over a period of time, they amass experience and insights, gradually evolving into well-informed members of the community.

In addition, the establishment of user groups with similar interests and a common goal of exchanging ideas, as evidenced in the interview-based study, can accelerate the development of expertise. In such groups, the collaborative atmosphere fosters a rapid exchange of insights and a collective learning experience, effectively reducing the financial burden on the company. In addition, the act of exchanging a range of viewpoints regarding a particular matter can be beneficial for users as it allows them to compare and contrast different ideas. This process can contribute to

their understanding of whether the topic at hand differs across different stages of utilizing the artificial intelligence system. Therefore, the company should consider hiring external experts during the initial phase, and as users develop their expertise, gradually reduce reliance on external experts.

D. User Feedback:

As previously stated, user feedback serves as a powerful tool for acquiring insights and improving the AI system. Feedback from users is a valuable resource for undertaking usability studies, as highlighted by Nikiforova & McBride (2021). Nevertheless, its functionality goes beyond just enhancing the system.

User feedback can also be harnessed to craft preliminary training materials tailored for novice users. This approach serves the purpose of easing novice users into the AI ecosystem, transforming them into semi-experts with a foundational understanding of the AI's functionalities. By using user feedback to shape training materials, the company not only bolsters the user's initial experience but also fosters an empowered user base capable of making the most of the AI system.

E. Educational Resources inside the XAI System:

In conjunction with the primary instructional materials obtained from user feedback, it is crucial to provide additional educational assets that can facilitate the effective utilization of the XAI system. Supplementary materials may manifest as instructional tutorials, compilations of frequently asked questions (FAQs), or comprehensive guides. These resources play a crucial role in facilitating users' acquisition of a deeper comprehension of the operations of the artificial intelligence system, encompassing its functionality and the processes involved in its decision-

making. The lack of comprehensive guides led Tesla FSD users to seek assistance from external sources for creating a knowledge base.

Tutorials provide step-by-step instructions on how to navigate and utilize the XAI system effectively, ensuring that users can make the most of its features. Frequently Asked Questions (FAQs) are designed to proactively address typical inquiries from users and offer concise responses, hence minimizing the necessity for users to seek external support (Gehrke & Turban, 1999). The comprehensive guidelines, which can be an expansion of the primary instructional materials created by the developers of the AI, aim to provide users with a detailed understanding of the decision-making procedures utilized by the AI system. This enables users to make well-informed decisions and effectively address any potential challenges that may arise during their engagements with the AI system.

F. Privacy and Security:

The preservation of privacy and security within the social forum is of paramount significance. The CXAI system could offer significant benefits when customized for a limited user population, such as an internal team within an organization or a closely interconnected community of interest. Within a corporate setting, it is plausible that this system may function as a substitute for a conventional bug-reporting mechanism. In doing so, it would not only provide valuable information regarding workarounds and limitations of tools, but also offer valuable insights into the experiences and requirements of users. In an alternative situation, a collective community of interest, such as radiologists employing a particular algorithm for diagnosing specific illnesses, might potentially derive significant advantages from

the implementation of a tailored CXAI system aimed at augmenting their collective comprehension and utilization of the technology. This approach acknowledges the significance of providing focused explainability and support specifically tailored to niche user groups (Ibne Mamun et al., 2022).

This system ensures the collection of honest viewpoints from AI users and protects them from any reprisals by hostile individuals affiliated with the institution. In order to attain this objective, it is recommended that the corporation implement an anonymous posting feature, which would enable users to express their opinions without the requirement of relinquishing rewards for engaging in the forum.

Additionally, it is imperative for users to possess a robust level of confidence in the security procedures implemented by the forum. Users should be provided with the assurance that their interactions and discussions take place inside a safe environment, ensuring the protection of their data against unwanted access or misuse. By offering a combination of anonymous features and strong security measures, the company has the ability to establish a secure environment that promotes open and sincere interaction, while also ensuring the protection of user trust and the integrity of their data.

G. Continuous Improvement:

While the CXAI system is intended for broad applicability, individual companies may still have distinct requirements stemming from their specific work environment and operational structure. It is crucial for organizations to regularly assess the effectiveness of the explanations generated using a CXAI system or social forums. Fitzgerald & Stol (2014) added continuous innovation with the

continuous improvement when a system goes to operation. Continuous Improvement emphasizes the use of lean thinking principles (Thangarajoo & Smith, 2015), which promote evidence-based decision-making and the removal of inefficiencies. It centers on implementing gradual, incremental enhancements to quality. These enhancements have the potential to generate significant advantages and pose difficulties for competitors attempting to duplicate them (X. Chen et al., 2007; Fowler & Beck, 1997; Jarvinen et al., 1999; Krasner, 1992), whereas Continuous Innovation is a sustainable process that can effectively respond to evolving market dynamics necessitates the use of suitable metrics at every stage of the cycle, including planning, development, and runtime operations (Cole, 2001; Holmström Olsson et al., 2012; Ries, 2011).

Through the consistent collection of feedback and implementation of user surveys, organizations are able to assess the efficacy of their explanatory endeavors. This feedback loop facilitates the identification of potential limitations, locations where users may face challenges or topics that necessitate more extensive inclusion. Consequently, these observations can contribute to iterative enhancements in the explanations, ultimately augmenting the user experience and the transparency and usability of the AI system.

H. Recognizing the AI's Limitations:

When adopting a novel AI system, it is critical to acknowledge that perfect performance cannot be reasonably expected in the early stages of implementation. However, if an organization portrays the AI as flawless or fails to appropriately communicate its limitations to users, the initial degree of trust in the system may

be significantly raised. However, the trust in AI is diminished when its flaws become evident during the practical applications, as demonstrated by the results of the Prediction Study (see Figure 13 – after the driving, in the study, though not significant, the trust in the system declines for no training and remains at the same level, without any augmentation in trust towards the system, for the rest of the trainings).

The training materials for the AI with shortcomings might focus more on providing tips and strategies for dealing with unusual situations rather than emphasizing the AI's positive aspects. Therefore, it's essential to inform AI users that the training materials derived from social forums are intended not to promote trust in the AI but to aid them in comprehending how the AI works.

In essence, it is imperative to ensure transparency on the limitations of an AI system in order to effectively manage user expectations. The act of concealing or downplaying these limits has the potential to create a deceptive perception of trustworthiness at first, however ultimately may lead to disillusionment and diminished trust once the defects of the AI system become evident.

The CXAI or social forums represent a significant leap forward in the field XAI, where its implementation aligns with a straightforward framework. These innovations not only enhance the transparency and interpretability of AI systems but also promote collaboration between humans and machines. Importantly, this research highlights the often-overlooked aspect of human-human collaboration within the XAI domain. The results of experiments demonstrate that with a higher quality of training content, CXAI systems offer a feasible solution within the autonomous vehicle sector. Unlike complex and challenging-to-

understand XAI systems, CXAI systems can be relatively easy to implement, particularly for a diverse user population. This approach ensures that AI systems in autonomous vehicles remain accessible and comprehensible to users, fostering a safer and more user-friendly experience.

8 Limitations

This dissertation covers the preliminary stage of implementing CXAI, which involves employing social forums as a non-algorithmic XAI approach in the self-driving vehicle sector. Although the framework has not been subjected to practical testing with active Tesla drivers, a research study conducted through interviews is the initial stage in engaging real Tesla users for the evaluation and enhancement of the framework. The subsequent phase of user studies may be undertaken using a Tesla vehicle that is equipped with FSD capabilities, and these studies would be carried out on real-world roads in a closed-loop approach and in the presence of a secondary task. This XAI method can be embedded during the drive, to start a new thread with an abnormal driving in a dedicated system. But learning during the drive from the system may be distracting and unsafe. But it can be used in driver training where risk is minimized.

9 Conclusion

The integration of CXAI, also known as social forums, into the realm of autonomous driving as a XAI system holds significant potential and offers promising prospects. This expedition has emphasized that CXAI provides a distinct avenue for bridging the divide between complex machine learning algorithms and human understanding. In contrast to model-based XAI systems, this framework is not constrained by a particular version of AI, hence obviating the necessity for repeated adaptations when newer iterations of AI are introduced. Integrating social forums/CXAI with AI technology will play a crucial role in the proactive identification of atypical driving behaviors, thereby contributing to establishing a safe road environment that offers advantages to drivers and bystanders.

The notion of CXAI not only provides us with valuable understanding of the decision-making mechanisms utilized by autonomous vehicles, but also enables drivers to make informed and safe decisions when faced with intricate driving situations. Furthermore, the inquiry has underscored the crucial significance of integrating tangible user input and feedback from real-world scenarios to improve XAI systems designed specifically for autonomous driving. This endeavor aims to bridge an existing void of not including user input in present XAI systems. This methodology guarantees the creation of customized solutions to fulfill the distinct requirements and anticipations of those who significantly depend on these vehicles.

Nevertheless, it is imperative to recognize the inherent difficulties linked to the implementation of CXAI. The issues at hand comprise the imperative to ensure the protection of data privacy, address any biases, and lack of experts. The aforementioned

concerns emphasize the importance of continuous research and development endeavors focused on enhancing CXAI systems specifically tailored for autonomous driving. Furthermore, it is worth considering the testing of CXAI systems in various domains, particularly in the healthcare sector, given that the existing XAI systems are not well-suited for this specific domain (Ghassemi et al., 2021).

10 Reference List

Abe, G., & Richardson, J. (2004). The effect of alarm timing on driver behaviour: An investigation of differences in driver trust and response to alarms according to alarm timing. *Transportation Research Part F: Traffic Psychology and Behaviour*, *7*(4–5), 307–322.

Abe, G., & Richardson, J. (2005). The influence of alarm timing on braking response and driver trust in low speed driving. *Safety Science*, *43*(9), 639–654.

Abe, G., & Richardson, J. (2006). Alarm timing, trust and driver expectation for forward collision warning systems. *Applied Ergonomics*, *37*(5), 577–586.

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alam, L. (2020). *Investigating the Impact of Explanation on Repairing Trust in Ai Diagnostic Systems for Re-Diagnosis*.

Alam, M. I., Halder, R., & Pinto, J. S. (2021). A deductive reasoning approach for database applications using verification conditions. *Journal of Systems and Software*, *175*, 110903.

Alambeigi, H., Smith, A., Wei, R., McDonald, A., Arachie, C., & Huang, B. (2021). A novel approach to social media guideline design and its application to automated vehicle events. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *65*(1), 1510–1514.

Albrecht, S. V., Brewitt, C., Wilhelm, J., Gyevnar, B., Eiras, F., Dobre, M., & Ramamoorthy, S. (2021). Interpretable goal-based prediction and planning for autonomous driving. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 1043–1049.

Atakishiyev, S., Salameh, M., Yao, H., & Goebel, R. (2021). Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions. *arXiv Preprint arXiv:2112.11561*.

Ayoub, J., Avetisian, L., Yang, X. J., & Zhou, F. (2023). Real-Time Trust Prediction in Conditionally Automated Driving Using Physiological Measures. *IEEE Transactions on Intelligent Transportation Systems*.

Ayoub, J., Zhou, F., Bao, S., & Yang, X. J. (2019). From manual driving to automated driving: A review of 10 years of autoui. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 70–90.

Ball, D., Coelho, P. S., & Machás, A. (2004). The role of communication and trust in explaining customer loyalty: An extension to the ECSI model. *European Journal of Marketing*.

Beggiato, M., Pereira, M., Petzoldt, T., & Krems, J. (2015). Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *35*, 75–84.

Bellet, P. S., & Maloney, M. J. (1991). The importance of empathy as an interviewing skill in medicine. *Jama*, *266*(13), 1831–1832.

Berger, P., Hennig, P., Bocklisch, T., Herold, T., & Meinel, C. (2016). A journey of bounty hunters: Analyzing the influence of reward systems on stackoverflow question response times. *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 644–649.

Bishka, A. (2010). Learning styles fray: Brilliant or batty?: Learning Styles Fray: Brilliant or Batty? *Performance Improvement*, *49*(10), 9–13. https://doi.org/10.1002/pfi.20181

Bjørner, T. (2017). Driving pleasure and perceptions of the transition from no automation to full self-driving automation. *Applied Mobilities*.

Bojarski, M., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., Muller, U., & Zieba, K. (2016). Visualbackprop: Visualizing cnns for autonomous driving. *arXiv Preprint arXiv:1611.05418*, *2*.

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Olah, C., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., … Kaplan, J. (2022). *Measuring Progress on Scalable Oversight for Large Language Models* (arXiv:2211.03540). arXiv. http://arxiv.org/abs/2211.03540

Brewitt, C., Gyevnar, B., Garcin, S., & Albrecht, S. V. (2021). GRIT: Fast, Interpretable, and Verifiable Goal Recognition with Learned Decision Trees for Autonomous Driving. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1023–1030.

Brown, J. S., & Burton, R. R. (1978). A paradigmatic example of an artificially intelligent instructional system. *International Journal of Man-Machine Studies*, *10*(3), 323–339.

Bylund, C. L., & Makoul, G. (2005). Examining empathy in medical encounters: An observational study using the empathic communication coding system. *Health Communication*, *18*(2), 123–140.

Chen, K., & Tomblin, D. (2021). Using data from reddit, public deliberation, and surveys to measure public opinion about autonomous vehicles. *Public Opinion Quarterly*, *85*(S1), 289–322.

Chen, X., Sorenson, P., & Willson, J. (2007). Continuous SPA: Continuous assessing and monitoring software process. *2007 IEEE Congress on Services (Services 2007)*, 153–158. https://ieeexplore.ieee.org/abstract/document/4278791/

Chi, M. T., Roy, M., & Hausmann, R. G. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science*, *32*(2), 301–341.

Chi, M. T., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, *25*(4), 471–533.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243.

Choi, J. K., & Ji, Y. G. (2015). Investigating the importance of trust on adopting an autonomous vehicle. *International Journal of Human-Computer Interaction*, *31*(10), 692–702.

Chromik, M., & Butz, A. (2021). Human-XAI interaction: A review and design principles for explanation user interfaces. *IFIP Conference on Human-Computer Interaction*, 619–640.

Cobb, C. D., & Mayer, J. D. (2000). Emotional Intelligence: What the Research Says. *Educational Leadership*, *58*(3), 14–18.

Cole, R. E. (2001). From Continuous Improvement to Continuous Innovation. *Quality Management Journal*, *8*(4), 7–21. https://doi.org/10.1080/10686967.2001.11918977

Comanici, G., Precup, D., Barreto, A., Toyama, D. K., Aygün, E., Hamel, P., Vezhnevets, S., Hou, S., & Mourad, S. (2018). *Knowledge representation for reinforcement learning using general value functions*.

Corso, A., & Kochenderfer, M. J. (2020). Interpretable safety validation for autonomous vehicles. *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 1–6.

Crandall, B., Klein, G. A., & Hoffman, R. R. (2006). *Working minds: A practitioner's guide to cognitive task analysis*. Mit Press.

Cunningham, P., Bonzano, A., & Smyth, B. (1995). *An incremental case retrieval mechanism for diagnosis*. Technical Report TCD-CS-95-01, Trinity College, Dublin, Ireland.

Cunningham, P., Smyth, B., & Bonzano, A. (1998). An incremental retrieval mechanism for case-based electronic fault diagnosis. *Knowledge-Based Systems*, *11*(3–4), 239–248.

Damböck, D., Farid, M., Tönert, L., & Bengler, K. (2012). Übernahmezeiten beim hochautomatisierten Autofahren. 5. Tagung Fahrerassistenz 2012. *München. Germany*.

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv Preprint arXiv:2006.11371*.

Deng, R., Benckendorff, P., & Gao, Y. (2022). Limited usefulness of learning style instruments in advancing teaching and learning. *The International Journal of Management Education*, *20*(3), 100686.

Dixit, V. V., Chand, S., & Nair, D. J. (2016). Autonomous vehicles: Disengagements, accidents and reaction times. *PLoS One*, *11*(12), e0168054.

Doyle, D., Cunningham, P., Bridge, D., & Rahman, Y. (2004). Explanation oriented retrieval. *European Conference on Case-Based Reasoning*, 157–168.

Doyle, D., Tsymbal, A., & Cunningham, P. (2003). *A review of explanation and explanation in case-based reasoning*. Trinity College Dublin, Department of Computer Science. http://ai2-s2-pdfs.s3.amazonaws.com/f3aa/7f2e9c820527cff2dca84227dd5a37011c35.pdf

Druskat, V. U., & Pescosolido, A. T. (2006). Chapter 2 The impact of emergent leader's emotionally competent behavior on team trust, communication, engagement, and effectiveness. In *Research on Emotion in Organizations* (Vol. 2, pp. 25–55). Emerald (MCB UP). https://doi.org/10.1016/S1746-9791(06)02002-5

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in ai systems. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–19.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. *Proceedings of the Human Factors Society Annual Meeting*, *32*, 97–101.

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors*, *37*(1), 65–84.

Endsley, M. R. (2006). Expertise and situation awareness. *The Cambridge Handbook of Expertise and Expert Performance*, 633–651.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1625–1634. http://openaccess.thecvf.com/content_cvpr_2018/html/Eykholt_Robust_Physical-World_Attacks_CVPR_2018_paper

Fitzgerald, B., & Stol, K.-J. (2014). Continuous software engineering and beyond: Trends and challenges. *Proceedings of the 1st International Workshop on Rapid Continuous Software Engineering*, 1–9. https://doi.org/10.1145/2593812.2593813

Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020). *Fostering human agency: A process for the design of user-centric XAI systems*.

Foushee, H. C., & Manos, K. L. (1981). Information transfer within the cockpit: Problems in intracockpit communications (No. NASA TP-1875). *Moffett Field, CA: NASA-Ames Research Center*.

Fowler, M., & Beck, K. (1997). Refactoring: Improving the design of existing code. *11th European Conference. Jyväskylä, Finland*. http://jaoo.dk/jaoo1999/schedule/MartinFowlerRefractoring.pdf

Friedland, N. S., Allen, P. G., Matthews, G., Witbrock, M., Baxter, D., Curtis, J., Shepard, B., Miraglia, P., Angele, J., & Staab, S. (2004). Project halo: Towards a digital aristotle. *AI Magazine*, *25*(4), 29–29.

Fujiyoshi, H., Hirakawa, T., & Yamashita, T. (2019). Deep learning-based image recognition for autonomous driving. *IATSS Research*, *43*(4), 244–252.

Gang, N., Sibi, S., Michon, R., Mok, B., Chafe, C., & Ju, W. (2018). Don't Be alarmed: Sonifying autonomous vehicle perception to increase situation awareness. *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 237–246.

Garcia, K. R., Mishler, S., Xiao, Y., Wang, C., Hu, B., Still, J. D., & Chen, J. (2022). Drivers' Understanding of Artificial Intelligence in Automated Driving Systems: A Study of a Malicious Stop Sign. *Journal of Cognitive Engineering and Decision Making*, 15553434221117001.

Gehrke, D., & Turban, E. (1999). Determinants of successful website design: Relative importance and recommendations for effectiveness. *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers*, 8-pp. https://ieeexplore.ieee.org/abstract/document/772943/

George, S., Michel, C., & Ollagnier-Beldame, M. (2016). Favouring reflexivity in technology-enhanced learning systems: Towards smart uses of traces. *Interactive Learning Environments*, *24*(7), 1389–1407. https://doi.org/10.1080/10494820.2015.1016532

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745–e750.

Glenn, D. (2009). Matching teaching style to learning style may not help students. *The Chronicle of Higher Education*, 1–3.

Gold, C., Damböck, D., Lorenz, L., & Bengler, K. (2013). "Take over!" How long does it take to get the driver back into the loop? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 1938–1942.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, *9*(6), 495–522. https://doi.org/10.1002/acp.2350090604

Gregorc, A. F., & Ward, H. B. (1977). A New Definition for Individual. *NASSP Bulletin*, *61*(406), 20–26. https://doi.org/10.1177/019263657706140604

Harper, F. M., Moy, D., & Konstan, J. A. (2009). Facts or friends? Distinguishing informational and conversational questions in social Q&A sites. *Proceedings of the Sigchi Conference on Human Factors in Computing Systems*, 759–768.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier.

Hayes, B. K., Heit, E., & Swendsen, H. (2010). Inductive reasoning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(2), 278–292.

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, *105*, 105–120.

Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hoffman, R. R., Klein, G., & Mueller, S. T. (2018). Explaining Explanation For "Explainable AI." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 197–201.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv Preprint arXiv:1812.04608*.

Hofmarcher, M., Unterthiner, T., Arjona-Medina, J., Klambauer, G., Hochreiter, S., & Nessler, B. (2019). Visual scene understanding for autonomous driving using semantic segmentation. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 285–296). Springer.

Holliday, D., Wilson, S., & Stumpf, S. (2016). User trust in intelligent systems: A journey over time. *Proceedings of the 21st International Conference on Intelligent User Interfaces*, 164–168.

Holmström Olsson, H., Alahyari, H., & Bosch, J. (2012). Climbing the" Stairway to Heaven": A Multiple-Case Study Exploring Barriers in the Transition from Agile Development towards Continuous Deployment of Software. *Proc. 38th*

Euromicro Conference on Software Engineering and Advanced Applications, 392–399.

Ibne Mamun, T., Alam, L., Hoffman, R. R., & Mueller, S. T. (2022). Assessing Satisfaction in and Understanding of a Collaborative Explainable AI (Cxai) System through User Studies. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 1270–1274.

Israelsen, B. W., & Ahmed, N. R. (2019). "Dave... I can assure you... That it's going to be all right..." A definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships. *ACM Computing Surveys (CSUR)*, *51*(6), 1–37.

Jamson, A. H., Lai, F. C., & Carsten, O. M. (2008). Potential benefits of an adaptive forward collision warning system. *Transportation Research Part C: Emerging Technologies*, *16*(4), 471–484.

Jarvinen, J., Hamann, D., & Van Solingen, R. (1999). On integrating assessment and measurement: Towards continuous assessment of software engineering processes. *Proceedings Sixth International Software Metrics Symposium (Cat. No. PR00403)*, 22–30. https://ieeexplore.ieee.org/abstract/document/809722/?casa_token=70U0w7Xz4AoAAAAA:s6gLNcIIC4o9u6AhkLTojfzcttk36kEFHPyI0NfEMwwop3xCPg2zDmmADF4ubl7SLqOFsR0

Johnson-Laird, P. N. (1999). Deductive reasoning. *Annual Review of Psychology*, *50*(1), 109–135.

Kaur, K., & Rampersad, G. (2018). Trust in driverless cars: Investigating key factors influencing the adoption of driverless cars. *Journal of Engineering and Technology Management*, *48*, 87–96.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20.

Kim, J., & Canny, J. (2017). Interpretable learning for self-driving cars by visualizing causal attention. *Proceedings of the IEEE International Conference on Computer Vision*, 2942–2950.

Kim, J., Rohrbach, A., Akata, Z., Moon, S., Misu, T., Chen, Y.-T., Darrell, T., & Canny, J. (2021). Toward explainable and advisable model for self-driving cars. *Applied AI Letters*, *2*(4), e56.

Kim, S., Park, J., Han, S., & Kim, H. (2010). Development of extended speech act coding scheme to observe communication characteristics of human operators of nuclear power plants under abnormal conditions. *Journal of Loss Prevention in the Process Industries*, *23*(4), 539–548.

Kim, T., & Song, H. (2021). How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics*, *61*, 101595.

Kirschner, P. A. (2017). Stop propagating the learning styles myth. *Computers & Education*, *106*, 166–171. https://doi.org/10.1016/j.compedu.2016.12.006

Klauer, K. J., & Phye, G. D. (2008). Inductive reasoning: A training approach. *Review of Educational Research*, *78*(1), 85–123.

Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., & Nass, C. (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*(4), 269–275.

Körber, M. (2018). Theoretical considerations and development of a questionnaire to measure trust in automation. *Congress of the International Ergonomics Association*, 13–30.

Koskinen, K. M., Lyyra, A., Mallat, N., & Tuunainen, V. (2019). Trust and risky technologies: Aligning and coping with Tesla Autopilot. *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Kothawade, S., Khandelwal, V., Basu, K., Wang, H., & Gupta, G. (2021). AUTO-DISCERN: Autonomous Driving Using Common Sense Reasoning. *arXiv Preprint arXiv:2110.13606*.

Krasner, H. (1992). The ASPIRE approach to continuous software process improvement. *Proceedings of the Second International Conference on Systems Integration*, 193–194. https://www.computer.org/csdl/proceedings-article/icsi/1992/00217302/12OmNrF2DNw

Leake, D. B. (1994). Issues in goal-driven explanation. *Proceedings of the AAAI Spring Symposium on Goal-Driven Learning*, 72–79.

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15.

Lim, B. Y., & Dey, A. K. (2009). Assessing demand for intelligibility in context-aware applications. *Proceedings of the 11th International Conference on Ubiquitous Computing*, 195–204.

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

Lin, R., Ma, L., & Zhang, W. (2018). An interview study exploring Tesla drivers' behavioural adaptation. *Applied Ergonomics*, *72*, 37–47. https://doi.org/10.1016/j.apergo.2018.04.006

Lindemann, P., Lee, T.-Y., & Rigoll, G. (2018). Catch my drift: Elevating situation awareness for highly automated driving with an explanatory windshield display user interface. *Multimodal Technologies and Interaction*, *2*(4), 71.

Linja, A., Mamun, T. I., & Mueller, S. T. (2022). When Self-Driving Fails: Evaluating Social Media Posts Regarding Problems and Misconceptions about Tesla's FSD Mode. *Multimodal Technologies and Interaction*, *6*(10), 86.

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv Preprint arXiv:1606.03490*.

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A., & Tang, J. (2023). Trustworthy AI: A Computational Perspective. *ACM Transactions on Intelligent Systems and Technology*, *14*(1), 1–59. https://doi.org/10.1145/3546872

Liu, P., Yang, R., & Xu, Z. (2019). Public acceptance of fully automated driving: Effects of social trust and risk/benefit perceptions. *Risk Analysis*, *39*(2), 326–341.

Ma, Y., Wang, Z., Yang, H., & Yang, L. (2020). Artificial intelligence applications in the development of autonomous vehicles: A survey. *IEEE/CAA Journal of Automatica Sinica*, *7*(2), 315–329.

Mamun, T. I. (2021). *INVESTIGATING THE IMPACT OF ONLINE HUMAN COLLABORATION IN EXPLANATION OF AI SYSTEMS*.

Mamun, T. I., Baker, K., Malinowski, H., Hoffman, R. R., & Mueller, S. T. (2021). Assessing Collaborative Explanations of AI using Explanation Goodness Criteria. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *65*(1), 988–993.

Mamun, T. I., Hoffman, R. R., & Mueller, S. T. (2021). Collaborative Explainable AI: A non-algorithmic approach to generating explanations of AI. *International Conference on Human-Computer Interaction*, 144–150.

Martinho, A., Herber, N., Kroesen, M., & Chorus, C. (2021). Ethical issues in focus by the autonomous vehicles industry. *Transport Reviews*, *41*(5), 556–577.

McFarland, M. (2018). Uber shuts down self-driving operations in Arizona. *CNNMoney. Version*, *809*, 3.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Min, D. H., Chung, Y. H., & Yoon, W. C. (2004). Comparative analysis of communication at main control rooms of nuclear power plants. *Proceedings of IFAC/IFIP/IFORS/IEA Symposium*.

Mirnig, A. G., Wintersberger, P., Sutter, C., & Ziegler, J. (2016). A Framework for Analyzing and Calibrating Trust in Automated Vehicles. *Adjunct Proceedings of*

the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 33–38. https://doi.org/10.1145/3004323.3004326

Mok, B., Johns, M., Lee, K. J., Miller, D., Sirkin, D., Ive, P., & Ju, W. (2015). Emergency, automation off: Unstructured transition timing for distracted drivers of automated vehicles. *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, 2458–2464.

Mueller, S., Hoffman, R., Klein, G., Mamun, T., & Jalaeian, M. (2021). Non-algorithms for Explainable Artificial Intelligence. *Applied AI Letters*.

Mueller, S. T. (2009). A Bayesian recognitional decision model. *Journal of Cognitive Engineering and Decision Making*, *3*(2), 111–130.

Mueller, S. T. (2020). Cognitive anthropomorphism of AI: How Humans and Computers Classify images. *Ergonomics in Design*, *28*(3), 12–19.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv Preprint arXiv:1902.01876*.

Mueller, S. T., Mamun, T. I., & Hoffman, R. R. (n.d.). *Development and Investigation on a Collaborative XAI System (CXAI)*.

Mueller, S. T., Veinott, E. S., Hoffman, R. R., Klein, G., Alam, L., Mamun, T., & Clancey, W. J. (2021). Principles of Explanation in Human-AI Systems. *arXiv Preprint arXiv:2102.04972*.

Nancekivell, S. E., Sun, X., Gelman, S. A., & Shah, P. (2021). A Slippery Myth: How Learning Style Beliefs Shape Reasoning about Multimodal Instruction and

Related Scientific Evidence. *Cognitive Science*, *45*(10), e13047. https://doi.org/10.1111/cogs.13047

Nees, M. A., Sharma, N., & Herwig, K. (2020). Some Characteristics of Mental Models of Advanced Driver Assistance Systems: A Semi-structured Interviews Approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *64*(1), 1313–1317.

Ngo, T., Kunkel, J., & Ziegler, J. (2020). Exploring Mental Models for Transparent and Controllable Recommender Systems: A Qualitative Study. *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 183–191. https://doi.org/10.1145/3340631.3394841

Nikiforova, A., & McBride, K. (2021). Open government data portal usability: A user-centred usability analysis of 41 open government data portals. *Telematics and Informatics*, *58*, 101539.

Oh, C., Lee, T., Kim, Y., Park, S., Kwon, S., & Suh, B. (2017). Us vs. Them: Understanding Artificial Intelligence Technophobia over the Google DeepMind Challenge Match. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2523–2534. https://doi.org/10.1145/3025453.3025539

Oh, S. (2018a). Social q&a. *Social Information Access*, 75–107.

Oh, S. (2018b). Social Q&A. In P. Brusilovsky & D. He (Eds.), *Social Information Access: Systems and Technologies* (pp. 75–107). Springer International Publishing. https://doi.org/10.1007/978-3-319-90092-6_3

Omeiza, D., Webb, H., Jirotka, M., & Kunze, L. (2021). Explanations in autonomous driving: A survey. *arXiv Preprint arXiv:2103.05154*.

Ong, L. S., Shepherd, B., Tong, L. C., Seow-Choen, F., Ho, Y. H., Tang, C. L., Ho, Y. S., & Tan, K. (1997). The colorectal cancer recurrence support (CARES) system. *Artificial Intelligence in Medicine*, *11*(3), 175–188.

Owens, C. C. (1992). *Indexing and retrieving abstract planning knowledge.*

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Pappalardo, L., Ruggieri, S., & Turini, F. (2018). Open the black box data-driven explanation of black box decision systems. *arXiv Preprint arXiv:1806.09936*.

Preusse, K. C., & Rogers, W. A. (2016). Error interpretation during everyday automation use. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 805–809.

Rapisarda, B. A. (2002). The impact of emotional intelligence on work team cohesiveness and performance. *The International Journal of Organizational Analysis*, *10*(4), 363–379.

Raue, M., D'Ambrosio, L. A., Ward, C., Lee, C., Jacquillat, C., & Coughlin, J. F. (2019). The influence of feelings while driving regular cars on the perception and acceptance of self-driving cars. *Risk Analysis*, *39*(2), 358–374.

Redmond, M. V. (1989). The functions of empathy (decentering) in human relations. *Human Relations*, *42*(7), 593–605.

Reynolds, N., & Ruiz De Maya, S. (2013). The impact of complexity and perceived difficulty on consumer revisit intentions. *Journal of Marketing Management*, *29*(5–6), 625–645. https://doi.org/10.1080/0267257X.2013.774290

Ries, E. (2011). *The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses*. Currency.

https://books.google.com/books?hl=en&lr=&id=tvfyz-4JILwC&oi=fnd&pg=PA1&dq=TheLeanStartup:HowToday%E2%80%99sEntrepreneurs+UseContinuousInnovationtoCreateRadically&ots=8J8cC15pr-&sig=J9CFKGDBuADphJoKFGZ91Ri1MnE

Riesbeck, C. (1988). An interface for case-based knowledge acquisition. *Proceedings of the DARPA Case-Based Reasoning Workshop*, 312–326.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Mit Press.

Rohrer, D., & Pashler, H. (2012). Learning Styles: Where's the Evidence?. *Online Submission*, *46*(7), 634–635.

Schraagen, J. M., & Rasker, P. C. (2001). Communication in command and control teams. *TNO Human Factors, The Netherlands*.

Sevastjanova, R., Becker, F., Ell, B., Turkay, C., Henkin, R., Butt, M., Keim, D., & Mennatallah, E. A. (2018). *Going beyond Visualization. Verbalization as Complementary Medium to Explain Machine Learning Models*.

Shadrin, S. S., & Ivanova, A. A. (2019). Analytical review of standard Sae J3016 «taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles» with latest updates. *Avtomobil'. Doroga. Infrastruktura.*, *3 (21)*, 10.

Shah, C., Oh, S., & Oh, J. S. (2009). Research agenda for social Q&A. *Library & Information Science Research*, *31*(4), 205–209.

Shakeri, H., & Khalilzadeh, M. (2020). Analysis of factors affecting project communications with a hybrid DEMA^{TEL}-ISM approach (A case study in Iran). *Heliyon*, *6*(8). https://www.cell.com/heliyon/pdf/S2405-8440(20)31274-3.pdf

Smith-Jentsch, K. A., Johnston, J. H., & Payne, S. C. (1998). *Measuring team-related expertise in complex environments.* https://psycnet.apa.org/record/1998-06532-003

Spiro, H. (1992). What Is Empathy and Can It Be Taught? *Annals of Internal Medicine*, *116*(10), 843–846. https://doi.org/10.7326/0003-4819-116-10-843

Sternberg, R. J., & Sternberg, K. (2018). *The new psychology of love*. Cambridge University Press.

Stevens, A., & Roberts, B. (1983). Quantitative and qualitative simulation in computer based training. *Journal of Computer-Based Instruction*, *10*(1), 16–19.

Stiff, J. B., Dillard, J. P., Somera, L., Kim, H., & Sleight, C. (1988). Empathy, communication, and prosocial behavior. *Communication Monographs*, *55*(2), 198–213. https://doi.org/10.1080/03637758809376166

Strand, N., Stave, C., & Ihlström, J. (2018). A case-study on drivers' mental model of partial driving automation. *25th ITS World Congress, Copenhagen, Denmark, 17-21 September 2018*.

Suchan, J., Bhatt, M., & Varadarajan, S. (2019). Out of sight but not out of mind: An answer set programming based online abduction framework for visual sensemaking in autonomous driving. *arXiv Preprint arXiv:1906.00107*.

Sutton, R. S., Modayil, J., Delp, M., Degris, T., Pilarski, P. M., White, A., & Precup, D. (2011). Horde: A scalable real-time architecture for learning knowledge from

unsupervised sensorimotor interaction. *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, 761–768.

Tavangarian, D. (2004). Is e-Learning the Solution for Individual Learning? *Electronic Journal of E-Learning*, *2*(2), pp265-272.

Thangarajoo, Y., & Smith, A. (2015). Lean thinking: An overview. *Industrial Engineering & Management*, *4*(2), 2169–0316.

Tullio, J., Dey, A. K., Chalecki, J., & Fogarty, J. (2007). How it works: A field study of non-technical users interacting with an intelligent system. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 31–40. http://dl.acm.org/citation.cfm?id=1240630

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*, 103404.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Human tutoring: Why do only some events cause learning. *Cognition and Instruction*, *21*(3), 209–249.

Walch, M., Lange, K., Baumann, M., & Weber, M. (2015). Autonomous driving: Investigating the feasibility of car-driver handover assistance. *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 11–18.

Waller, M. J. (1999). THE TIMING OF ADAPTIVE GROUP RESPONSES TO NONROUTINE EVENTS. *Academy of Management Journal*, *42*(2), 127–137. https://doi.org/10.2307/257088

Walter, N., Brooks, J. J., Saucier, C. J., & Suresh, S. (2021). Evaluating the Impact of Attempts to Correct Health Misinformation on Social Media: A Meta-Analysis. *Health Communication*, *36*(13), 1776–1784. https://doi.org/10.1080/10410236.2020.1794553

Wang, C., Weisswange, T. H., Krueger, M., & Wiebel-Herboth, C. B. (2021). Human-vehicle cooperation on prediction-level: Enhancing automated driving with human foresight. *2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*, 25–30.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15.

Wang, J., Antonenko, P., Celepkolu, M., Jimenez, Y., Fieldman, E., & Fieldman, A. (2019). Exploring relationships between eye tracking and traditional usability testing data. *International Journal of Human–Computer Interaction*, *35*(6), 483–494.

Wang, W., & Siau, K. (2018). Living with Artificial Intelligence–Developing a Theory on Trust in Health Chatbots. *Proceedings of the Sixteenth Annual Pre-ICIS Workshop on HCI Research in MIS*, 1–5.

Watson, I. (1998). *Applying case-based reasoning: Techniques for enterprise systems*. Morgan Kaufmann Publishers Inc.

Werner, N. E., Tong, M., Borkenhagen, A., & Holden, R. J. (2019). Performance-shaping factors affecting older adults' hospital-to-home transition success: A systems approach. *The Gerontologist*, *59*(2), 303–314.

White, A. (2015). *Developing a predictive approach to knowledge*.

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science*, *2*(4), 352–367. https://doi.org/10.1080/14639220110110306

Winefield, H. R., & Chur-Hansen, A. (2000). Evaluating the outcome of communication skill teaching for entry-level medical students: Does knowledge of empathy increase? *Medical Education*, *34*(2), 90–94. https://doi.org/10.1046/j.1365-2923.2000.00463.x

Zablocki, É., Ben-Younes, H., Pérez, P., & Cord, M. (2022). Explainability of deep vision-based autonomous driving systems: Review and challenges. *International Journal of Computer Vision*, *130*(10), 2425–2452.

Zhang, J., & Curley, S. P. (2018). Exploring explanation effects on consumers' trust in online recommender agents. *International Journal of Human–Computer Interaction*, *34*(5), 421–432.

Zhou, J., Wang, S., Bezemer, C.-P., & Hassan, A. E. (2020). Bounties on technical Q&A sites: A case study of Stack Overflow bounties. *Empirical Software Engineering*, *25*(1), 139–177.

Zinn, W. (1993). The empathic physician. *Archives of Internal Medicine*, *153*(3), 306–312.

A Results for Training Pre and Post Anomalous Driving Events

Participant No.	Events	Prior the Events	After the Events
1	a. Yellow Sign/Blinking Yellow b. Lane Merging	NA	Search using Google using Keyword after the occurrences.
2	a. No Center Lane Marking in the Neighborhood b. Suddenly Pulling Over to the Right	Learned from self-driving	a. Discussed the occurrence on social forums. He was curious about other people's experiences. b. Looked for videos in snowy conditions
3	a. Bump b. Phantom Braking	NA	a. Asked friends (only instance where help was asked from acquaintances in a non-novice stage of driver) b. YouTube
4	a. Before rail crossing (passive) b. Suddenly Pulling Over to the Right	NA	a. Social forum, learned can't depend on AI all the time b. Social forum, learned a trick to avoid it

5	a. Before rail crossing (active) b. Yellow Sign/Blinking Yellow	NA	a. Shared experience either in reddit or twitter, found other experiences b. Posted on a social forum found other experiences
6	a. Before rail crossing (passive) b. STOP Sign (stop way before the sign)	NA	a. Googled about it b. Didn't look as it did not hamper safety of the road
7	a. Stopped Cars in the Lane b. Road Work Ahead Sign/Crew	Learned from self-driving	a. Common incident for FSD, did not post or follow-up in social forums b. Looked into social forums, learned tricks how to better handle the situation from shared experiences
8	a. Phantom Braking b. STOP sign (stop way before the sign)	a. Learned from self-driving b. YouTube	a. YouTube b. NA
9	a. Suddenly Pulling Over to the Right b. Lane Changing	NA	Learned from self-driving, takeover in the future, teaches other

10	a. Yellow Sign/Blinking Yellow b. No Center Lane Marking in the Neighborhood	NA	a. Send feedback to Tesla, but didn't heard back b. Self-Learned

B Results for Expectation Hierarchy

Participant No.	Follow legal/Rules of the road	Safety (legal but unsafe)	Behave unlike a human (legal and safe but otherwise bizarre)	Consistent/Accurate about the real world
1	√	Bizarre turns	Unsafe lane crossing	Garage washer/dryer displayed as truck, but in the road it is consistent
2	√	Unsafe snowy driving	Unsafe lane selection	√
3	√	Fast turn	Slow turn	√
4	√	Unsafe lane selection	Unsafe lane crossing	√
5	Went over speed bump without slowing down	Sometime accelerates over speed limit	Happened, but shared no experience	Couldn't recognize the speed bump
6	Shifted lane suddenly	Happened, but shared no experience	Phantom Braking	√
7	√	Speedy turn (go to the maximum speed limit soon)	Most unsafe activities: shifting lane mid-turn, overcautious in intersection, not letting other drivers merge, merging too	Gives false positive sometimes, getting better; e.g., slowed down for a vehicle close to its lane marker.

			late, can't sustain traffic speed if average speed is over speed limit	
8	√	Mistaken turn as STOP sign, creeps forward	Phantom Braking	√
9	Turns using bike lane	Slow turn	√	√
10	√	Stops FSD from entering hazardous situations	Bizarre turns (e.g., slow turns), jerky auto-park	Blinking Yellow mistaken as Traffic Light

C Communication Records

Record #12-1364

So I've been seeing a lot of people talking about their experiences in very hyperbolic ways, good and bad, but I feel like a lot of the bad is from misaligned expectations. Having now used the beta for a while and driven hundreds of miles on all my daily routes, and being very satisfied with it despite its problems I though I'd share why I feel that way. In doing so I hope to help others who have either just gotten or will soon get the beta have the best experience. First and foremost remember that despite the name this is NOT FSD. This is City Streets, and a beta at that. The goal with this step is not to make your car autonomous, but to enhance its existing capabilities to be much more useful off highway. Think of it as NoA for off highway. With that expectation I suspect you will have a much better time dear reader, because it will affect how you interact with the system. For example: Rather than expecting the car to simply get you from point A to B without your input, think of it as trying to reduce your workload. With turns from stops I can focus on looking for cars and making sure the car waits for and gets the right gap rather than worrying about creeping into someone coming the other way. With turns off roads I can focus on making sure the car gets the angle right rather than worrying about braking. What AP did for going straight, this does for taking turns. It will mess up, but the point isn't that it will do things for you, its that you can focus on making sure it does it right rather than doing it right yourself. If you felt comfortable using public AP and knew how to handle it well, you'll be fine. The mental workload is honestly no worse. (Hell for me personally, the roads, and amount of turns I take, it's less) That said, roads where you couldn't use AP before can be rough.

Unmarked roads in particular require attention as it likes to hug the center until there's a car coming and it's relatively close. Thing is, if markings fade or are washed out by the sun it can treat them the same too. But so long as you watch out, and put in the energy you're really supposed to be putting in with public AP you'll be okay. And as long as you remember that this is supposed to be NoA off highway, you'll have a good time. Thank you for coming to my TED talk.

Record #77-1490

Stuck with the default of average.

(*Parent Post:* Another anecdote; stop sign reactions are A LOT better. Maybe a little "too good"! Couple times it'd essentially do a rolling stop (if an officer wanted to be picky, it'd be a ticket-able offense). And another time I came to a stop at a 5-way intersection. There were two cars already stopped and a third that just came to a stop before I did. FSD handled it beautifully. It waited, first two cars started proceeding around the same time I stopped and my vehicle appropriately waited for that 3rd vehicle. Issue is that 3rd vehicle was taking a LITTLE long to start going and my 3 jumped at the chance and proceeded (as did that 3rd driver). I panicked and took over. In all honesty, it was creepy as it behaved exactly like a human would. In the video I didn't override FSD until the very end when the 3rd car started to move after FSD began entering the intersection)

Record #14-1371

In the FSD can you use only adaptive cruise control, keeping in the middle (or standard auto pilot) and automatic lane changes when you give turn signals? In other words can you uncheck other options?

Record #1-1338

FSD 10.3 Warning Our regular autopilot is now not working. Conditions are: 1. FSD disabled. 2. At or below speed limit. 3. Hands on wheel. Seemingly randomly we'll get "takeover immediately" message and sounder. FSD 10.3 has been working only intermittently. We also get lots of clearly false forward collision warnings while driving with 10.3 enabled but not turned on. In the image you can see that the visualization is present but autosteer and cruise is unavailable.

Record #2-1343

TEMP FIX found, Turn off sentry mode before you go into drive. That worked for me. So I believe that it works for most people in the morning because they have sentry mode off at home location. And their 2nd drive it starts being buggy because sentry is on before going into drive mode. Is sentry mode set to off at home for you?

Record #8-1355

Read that 98 score people will be getting FSD soon. So here is a tip. The FSD does not understand drives ways into large parking or communities yet. It will think its a road and will hit the sloped paving pretty fast. So ALWAYS be prepared to take over. Of course, always be prepared to take over regardless. This is a beta you know.

Record #10-1360

I have a relatively uneventful work commute on a 2-lane highway in semi-rural North Carolina, and I've enjoyed it so far. My main issues are how much it hugs the yellow line (not very safe or comfortable on a 2-lane road where there are frequent semi trucks), and for some reason it brakes almost every time an oncoming car approaches, even when they're fully in their lane and off toward the shoulder.

Record #52-1449

Yes! NoA is the second stack. Once we have single stack, that won't be needed anymore.

Record #28-1410

My first drive this morning was interesting as well. Mostly around the neighborhood to start. It's challenging here, no center stripe on even the widest residential roads. The car tended to hog the center (and would freak out a bit when it saw an oncoming vehicle). Stop signs are interesting and it's doing kinda what I expect, stops for the stop sign and then creeps forward until it can "see" that the coast is clear and then it goes. A bit jerky but it gets the job done. I will comment that driving with the beta running requires MORE than your full attention when cruising around neighborhoods, so I'm hyper alert when driving with it running. Los Alamos is probably a pretty good place to be doing a beta like this. I'm having fun Pretty impressive progress from EAP from almost 4 years ago when I first tried to use it on Max! But still a lot of work to do IMHO.

Record #31-1415

This was NOT my experience. I had a relatively short drive and lots of disengagement. I'm still blown away at what it can do, but this release did not match my expectations of what I've seen on other videos. Granted some of my scenarios could be more challenging, some things it did wrong seemed elementary. It wanted to go around a car to the left that was stopped at a stop light. Although I was taking a right at the stop light which was only 2 car lengthens in front of me. Also for an upcoming left-hand turn it got moved over to the right one lane before wanting to go back to the left. There are some things that I thought it wouldn't perform well on because its hard for human drivers, but sometimes surprised me. Still very hopefully and blown away that it can make turns. Just not as polished as I would

have expected on the short trip I went on. I sent a lot of feedback so maybe I'll see some changes in the next update which appears to be every 2-3 weeks right now. Still happy to be beta testing but it definitely requires way more effort than being on AP.

D Coding Schema for Communication about AI Systems

Dimensions	Elements/Labels	Description
Framing-Reframing	a. Evaluation b. Clarification c. Observation d. Response uncertainty e. Denial or Disconfirmation	a. Evaluative utterances or judgments concerning the activities of the scenario just played out. Analysis of why things went well or wrong. b. Clarifications serve to clear up misunderstandings from other individuals. c. Questions and answers that someone either asked or seemed to misunderstand. This includes repetitions for clarification, associations, and explanations. d. A statement that describes the AI's action during use. e. Statements indicating uncertainty or lack of information with which to

		respond to a command, inquiry, or observation. f. Disconfirming a statement.
Resolution	Situation Resolved	Combination of some of the other elements of Framing-Reframing—resolution/ workaround/abandonment of a practice conditionally/abandonment of a practice wholly/why it is doing it (not giving a solution but a reason).
Emotion	a. Frustration or anger with AI b. Frustration or anger on response c. Appreciation for the AI d. Appreciation for a Response e. Embarrassment	a. During the use of AI. b. During the use of AI. c. During communication. d. During communication. e. Any response apologizing for an incorrect response, etc.
Empathy	a. Agreement /Acknowledgement b. Shared experience c. Perfunctory recognition d. Antagonism	a. 'A' conveys to 'B' that the expressed emotion, progress, or challenge is legitimate. b. 'A' has a similar experience to

	that of 'B' with progress or a challenge. c. 'A' gives an automatic, scripted-type response, or repeats the company's policy/response, giving the empathetic opportunity minimal recognition. d. Deflates the other's response, defends or asserts self-response.

E Results - Communication Analysis

Dimensions	Elements/Labels	Result	kappa (McHugh, 2012)
Framing-Reframing	a. Evaluation b. Clarification c. Response uncertainty d. Observation e. Denial or Disconfirmation f. Other	a. 20 (34) b. 23 (31) c. 16 (19) d. 111 (135) e. 1 (3) f. 1 (8)	0.78
Resolution	a. Situation Resolved b. Not Resolved	a. 38 (60) b. 136 (158)	0.7
Emotion	a. Frustration or anger with AI b. Frustration or anger on response c. Appreciation for the AI d. Appreciation for a Response e. Embarrassment f. Non-emotional	a. 96 (120) b. 0 (2) c. 27 (36) d. 0 (3) e. 0 (1) f. 47 (60)	0.78
Empathy	a. Agreement /Acknowledgement b. Shared experience c. Perfunctory recognition d. Antagonism e. Non-empathy	b. 23 (54) c. 21 (33) d. 0 (1) e. 9 (12) f. 124 (138)	0.81