IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version Nov 23, 2021.

Digital Object Identifier https://doi.org/10.1109/ACCESS.2021.3130489

What and when to explain? A survey of the impact of explanation on attitudes towards adopting automated vehicles

QIAONING ZHANG[1], X. JESSIE YANG[2], AND LIONEL P. ROBERT JR.[1]

[1]School of Information, University of Michigan, Ann Arbor, USA
[2]Department of Industrial Operations Engineering, University of Michigan, Ann Arbor, USA

Corresponding author: Qiaoning Zhang (e-mail: qiaoning@umich.edu).

ABSTRACT Automated vehicles (AV) have the potential to decrease driving-related accidents and traffic congestion and to reduce fuel consumption and carbon emissions. However, because of a lack of trust and acceptance, their widespread adoption is far from certain. One approach researchers have taken to promote trust and acceptance of AVs is to decrease the uncertainty associated with their actions by providing explanations. AV explanations are the reasons the AV provides to make its actions easier to understand. There is now a nascent but rapidly growing body of research on AV explanations. Yet, answers to basic questions like whether or when AV explanations are effective still elude us. To better understand what has been done and what should be done with regard to AV explanations, we present a review of the literature, discuss the findings and identify several important future research directions.

INDEX TERMS Automated Vehicle, explanation, interaction, transportation.

I. INTRODUCTION

DESPITE technological advancements, the widespread adoption of automated vehicles (AVs) is far from certain. The Society of Automotive Engineers (SAE) classifies driving automation into six levels spanning from no automation (Level 0) to full automation (Level 5) as shown in Table 1 [1]–[11]. At each ascending level, AVs need less human involvement [12]. Delegating most or all vehicle driving responsibilities to the AVs can potentially reduce driving-related accidents [13]–[15] and traffic congestion [14], [16] and decrease fuel use and carbon emissions [14], [17], [18]. However, because of a lack of trust, the public is reluctant to adopt AVs [1], [19]. Therefore, understanding approaches to promoting trust in AVs remains an important challenge.

Explanations can be crucial to the acceptance of AVs. Explanations—reasoning or logic behind actions—provide essential information to the user that often justifies decisions made by the automation, leading to better interactions between the user and the automation [1], [20]. AV explanations allow the AV's actions to become predictable and understandable, helping the driver form accurate mental models [2]. These mental models create an approximate representation of the system's functions and competency needed to assist the driver in understanding the appropriate action needed [3], [21].

There is now a body of research on how AV explanations impact driver-related outcomes. Yet, the factors affecting the effectiveness of AV explanation are unclear. To answer this research question, this paper reflects on and derives insights from the existing literature on what we know and identified what we should seek to find going forward. To accomplish this, we: (1) survey the literature on explanations provided by Levels 2–5 AVs, which control and perform some or all aspects of the driving [22]; (2) present and discuss the findings with regard to the effectiveness of AV explanations on driver outcomes (e.g., trust); (3) identify remaining challenges and present future research suggestions.

This paper represents a deep reflection on the existing AV explanation literature and provides an important starting point for future research on AV explanations. As such, this paper provides several contributions to the literature. Firstly, this paper highlights major thematic research areas in the study of AV explanation and the acceptance of AVs. Secondly, this paper derives and presents major conclusions from the literature on AV explanations. In doing so, this paper identifies what we currently know about how to design more effective AV explanations. Finally, the paper identifies important gaps in the AV explanation literature and offers guidance for future research.

Accordingly, the remainder of the paper is organized as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3130489, IEEE Access

follows: Section II presents a literature review that explores the relationship between explanations and AVs. Section III presents the remaining challenges and future research directions in this field.

II. EXPLANATIONS AND AUTOMATED VEHICLES

As extracted from the *Cambridge Advanced Learner's Dictionary*, explanations are "reasons that someone gives to make something clear or easy to understand" [23, p. 492]. This can be rephrased in the context of AVs as reasons that the AV provides to make its actions clear or easy to understand. The prior literature on AV explanations can be organized into two research areas: explanation content and explanation timing [1]–[6], [8]–[10]. Table 2 includes the descriptions of explanation content and timing and summarizes the corresponding references that used different content and timing categories. To exemplify varied explanation strategies, we will use one specific driving task (i.e., Stopped Traffic Ahead) to introduce how different explanations contents and timings were designed and utilized in prior researches.

A. AV EXPLANATION CONTENT

AV explanation content refers to the information presented to the driver. Previous studies have examined the impact of the AV explanation content on driver reactions. The content of AV explanations can be classified into three groups: (1) "what," (2) "why" or (3) "what" + "why." The "what" content refers to what actions the AV has taken in the past or will take in the future. The "why" content refers to the information on why the vehicle took or will take a particular action. The "what" + "why" provides both information on what the car did/will do and why the vehicle took/will take a particular course of action [4], [8], [9]. Prior research has found that different content can have different impacts on drivers' attitudes and behaviors. Table 3 [1]–[7], [9] shows a summary of literature review by the impacts of explanation on AV-related outcomes.

1) What-only Explanation

What-only explanations provide descriptions of the AV action (i.e., what will/did the vehicle do?). In the "Stopped Traffic Ahead" driving task, the AV delivers a what-only explanation, "Rerouting", to inform the driver of the AV action of rerouting. Koo et al. (2015) [4] employed a fixed-base driving simulator equipped with a vehicle mock-up at real-world dimensions to explore the effect of AV explanation content on drivers' attitudes (i.e., emotional valence and AV acceptance) and behaviors (i.e., driving performance) [4]. The results indicate that the what-only explanation led to worse performance when compared to the what + why, the why-only and the no-explanation conditions with regard to driving performance and AV acceptance. The authors found that the what-only explanation had the lowest acceptance and led to the most dangerous driving performance [4].

2) Why-only Explanation

Why-only explanation describes the reasoning for the AV actions. For example, the AV provides a why-only explanation, "traffic reported ahead", to explain the reason of rerouting to drivers in the example driving task. Koo et al. (2015) [4] found that the why-only explanation was associated with the least anxiety, highest trust and preference, and the highest driving performance [4]. The why-only explanation enhanced the interaction between the driver and the AV by helping drivers anticipate and coordinate their reaction to upcoming events [4]. Koo et al. (2016) [5] also found that providing the why-only explanation decreased drivers' anxiety levels associated with automated driving, helped drivers maintain internal locus of control, and improved drivers' alertness in automated driving and was the most preferred condition [5]. The why-only explanation was also critical for increasing the drivers' level of situational awareness. To understand drivers' visualization preferences for explanations, Wiegand et al. (2019) [9] conducted a simulator study utilizing a desktop driving simulator with driver seat, steering wheel and pedals. Participants were presented with explanations that consisted of abstract visualizations of different autonomous system components representing AV actions (what-only explanation) and driving contexts (why-only explanation). The components serving as the why-only explanation included object's movement prediction (i.e., where the object on road might move next); context information (i.e., the background information of the situation, abstracting information); sensor symbols (i.e., from which sensor the information was retrieved); sensor range (i.e., the information visualized by a transparent region around the vehicle); environment information (i.e., houses or trees in the scenario); and infrastructure (i.e., a road and traffic lights display) [9]. The what-only explanation included the driver's movement prediction (i.e., where the vehicle might move next) and travel route (i.e., a line on the road visualizing the planned route). The participants were told to choose the explanations they perceived as necessary for their situational understanding. Results show that presenting the why-only explanation including the detected objects and their predicted motion was essential to understanding a situation and increasing situational understanding [9].

In sum, the why-only explanation has benefited drivers by promoting acceptance [4], trust [4], preference [5], perceived understandability [9], alertness [5], and sense of control [5] and by decreasing anxiety [4], and improving safe driving performance [4].

3) What + Why Explanation

What + why explanation describes the vehicle action and outlines the reason for the action. An example of such an explanation could be "Rerouting, traffic reported ahead" which combines both the AV action and reasoning in the driving task. Koo et al. (2015) [4] found that drivers felt more anxious and annoyed when they were told both what and why the vehicle was about to do compared to the why-only

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3130489, IEEE Access

TABLE 1: Summary of literature review by level of automation (SAE).

SAE Level Name	Time Period									
	2015	2016	2017		2018		2019		2020	
Level 0: No Automation										
Level 1: Driver Assistance										
Level 2: Partial Automation	[4]	[5]								
Level 3: Conditional Automation			[3]	[6]	[2]					
Level 4: High Automation							[1]			
Level 5: Full Automation					[7]	[8]	[9]		[10]	[11]

TABLE 2: Summary of literature review by explanation content and timing

Explanation attribute	Category	Description	References
Content	What-only	What-only explanation provides descriptions of the vehicle action itself (i.e., what will/did the vehicle do?)	[4], [9]
	Why-only	Why-only explanation describes the reasoning for action (i.e., why did the vehicle perform that action?)	[4], [5], [9]
	What + Why	What + Why explanation describes the vehicle action itself and the reasoning for the action.	[1], [3], [4], [6], [8] [10]
Timing	Before action	The time to provide explanations is before the AV acts.	[1], [3], [4], [5], [6]
	After action	The time to provide explanations is after the AV acts.	[1], [2], [9], [10]

TABLE 3: Summary of the impacts of explanations on AV outcomes

	Explanation Attributes				
	Content			Timing	
Explanation Outcomes	Why + What	Why-only	What-only	Before action	After action
Trust	(+) [3]	(+) [4], [5]		(+) [1], [3], [7]	(-) [1]
Driving Performance	(+) [4]		(-) [4]		
Anthropomorphism	(+) [3], [7]			(+) [3]	
Acceptance	(+) [3], [4], [6]	(+) [4], [5]		(+) [3], [6]	
Anxiety	(+) [4]	(-) [4], [5];		(-) [5];(‡)1001[1]	
Preference		(+) [5]		(+) [1], [5]	(-) [1]
Understandability		(+) [9]		(+) [9]	(+) [2]
Alertness		(+) [5]		(+) [5]	
Sense of Control		(+) [5]		(+) [5]	
Workload	(-) [6]			(-) [6];(‡)1001[1]	
Annoyance	(+) [5]				
Likeability				(+) [7]	
Usability	(+) [3], [6]			(+) [3], [6]	
Intelligency				(+) [7]	

Notes: "+", "-", and "‡" show the positive, negative, and marginal negative effects of explanation on outcomes compared to **no-explanation** condition, respectively; results of literature investigating moderators (i.e., [8], [10], [11]) are not shown in this table but can be found in Section II; the works of [1] and [4] give more comparisons results among explanation conditions in addition to the comparison with the no-explanation condition and can be found in Section II.

explanation and what-only explanation [4]. Although the what + why explanation led to more anxiety and annoyance, it was also associated with the safest driving performance [4]. According to [4], the what + why explanation assisted in coordinating the actions of the driver with the AV.

The benefits of AV providing a what + why explanation was confirmed by other empirical studies. Forster et al. (2017) [3] explored the potential of adding the what + why explanation to promote trust in AVs using a motion-based driving simulator. The what + why explanation was also found to be superior in promoting trust, anthropomorphism and usability compared to the no-explanation condition [3]. Naujoks et al. (2017) [6] also affirmed the benefit of the what + why explanation. They found that when compared to the no-explanation condition, the what + why condition was more effective at decreasing visual workload by reducing the driver's need to monitor the AV's interface [6]. The what + why explanation made the driving automation more accessible because the drivers did not have to monitor the driving environment to understand the system's intentions and actions. For the human driver, this makes the system easier to understand, learn and use [3], [6].

Prior research showed that the effectiveness of the what + why explanation on trust can be conditional on the driving event and vehicle actions, driving environment, and the point-of-view of explanation. The importance of the driving event and vehicle actions on influencing the relationship between explanation and AV trust was highlighted in Hatfield's (2018) study [8]. This study examined the "Trolley problem" and found that providing no explanation was better in terms of trust than providing a what + why explanation when the AV remained in the original lane where it would crash into five persons. When the AV intervened and directed itself to another lane where it would hit one person, there was no

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3130489, IEEE Access

TABLE 4: Summary of literature review by driving event

Driving Event	References
Parking	[11]
Stop sign	[11]
Vehicle crashed	[7]
Red traffic light	[7], [10]
Unclear lane lines	[1]–[3]
Missing GPS data	[2]
Speed limit ahead	[3], [6]
Road hazard ahead	[5], [11]
Roadway obstruction	[1], [3], [5], [6], [10]
Pedestrian jumping in	[4], [5], [7], [8], [10], [11]
Traffic reported ahead	[1]
Bicycle/motorcycle ahead	[7]
Highway intersection ahead	[3], [6]
Emergency vehicle approaching	[1] [9]
Emergency vehicle on shoulder	[1]
Vehicle with hazard light ahead	[1]
Oversize vehicle blocking roadway	[1]

TABLE 5: Summary of literature review by vehicle action

Vehicle Action	References
Stop	[1], [7], [9], [10]
Yield	[7]
Reroute	[1], [3], [6]
Slow down	[1], [3]–[7], [9], [11]
Change lanes	[1], [3], [6], [8], [11]
Ask driver for takeover	[2], [3]

difference between providing no explanation and the what + why explanation in terms of trust in AVs [8]. Tables 4 and 5 ([1]–[11]) summarize the literature by driving event and vehicle action. Ha et al. (2020) [10] examined the effects of perceived risk and explanation on trust in AVs using a driving simulator with a virtual reality device. Four automated driving environments were designed with different weather (i.e., clear day and snowy night) and driving speed (i.e., fast—faster than 40 km/h and slow—slower than 40 km/h). Three explanation conditions were presented: no explanation, what + why explanation with no subject (e.g., "stopped after identifying the sudden appearance of a pedestrian in the road"), and what + why explanation with a third-person point of view (e.g., "the autonomous vehicle stopped after identifying the sudden appearance of a pedestrian in the road"). Results showed that the perceived risk of driving environment and explanation conditions significantly moderated the effectiveness of the what + why explanation on trust in AVs. Specifically, when drivers perceived low risk, third-person explanations were the most effective on trust. However, as users' perceived risk increased, the effect of third- person explanations decreased, and providing no explanation was the most effective on trust [10]. The summary of literature by explanation point of view is shown in Table 6 [1]–[11].

In sum, previous research suggests that presenting the what + why explanation is an effective method to promote trust [3], [10], perceived anthropomorphism [3], acceptance of AVs [3], [4], [6], and driving performance [4], but the what + why explanation can also increase anxiety and annoyance when compared to the what-only or why-only explanation [4], [5].

4) Summary across AV Explanation Content Studies

The literature can be organized into three overarching findings. One, the why-only explanation content leads to the best driver outcomes. The why-only explanation has consistently been shown to be associated with promoting positive attitudes including acceptance, trust, preference, understandability, alertness, and a sense of control; decreasing negative feelings like anxiety; and assisting drivers in driving safely [4], [5], [9]. Two, the what-only explanation content is associated with the worst driver outcomes. The what-only explanation led to the most dangerous driving and reluctance to accept the AV [4]. Finally, the why-only explanation content has shown mixed results. Although the what + why explanation produced positive emotional valence and safe driving performance, drivers felt anxious and annoyed when receiving the what + why explanation [3], [4], [6]. Additionally, the effectiveness of the why + what explanation was subject to three conditions: driving event, driving environment and explanation point of view [8], [10].

B. AV EXPLANATION TIMING

The timing of the AV explanation—when the AV provides the explanation—is likewise crucial to the effectiveness of AV explanations. AV explanation research has operationalized the impact of timing as providing the explanation either before or after the AV has acted.

1) AV Explanation before Action

Prior literature investigated the relationship between explanations and AV-related outcomes when the AV explanations were provided before the AV acted. In the example driving task, if the vehicle delivers explanations seconds prior to the intersection event (i.e., reroute), then the AV provides its explanation before its action. Providing an explanation before the AV takes action has been closely associated with higher positive attitudes (i.e., trust, anthropomorphism, acceptance, preference, situational awareness, sense of control, and alertness) and lower negative feelings (i.e., anxiety and workload) [1]–[3], [5]–[7], [11].

The specific time for prior studies to provide the explanation can be organized into three categories: 1 second (s) before the AV action, 7 s before the AV action and undefined time. AV explanations were presented 1 s ahead of the AV's action in the work of Koo et al. (2016) [5] to examine how the explanation accompanying the vehicle's autonomous action affects the driver's attitude and driving behavior [5]. Their results showed that when the AV explained what it was going to do before it acted, it decreased drivers' anxiety, promoted preference and alertness, and increased drivers' sense of control. The sense of control is essential for drivers because it is closely linked to driving performance and perceptions. According to the concept of "locus of control," drivers feel that either they themselves (an internal determinant) or the automated systems (an external determinant) are mainly responsible for the behavior of the vehicle [24]. Providing insufficient explanations might drive an individual to assume

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3130489, IEEE Access

TABLE 6: Summary of literature review by explanation content and point of view

Content	Point of View	Example	References
What-only	First-person	I am giving way to the bicyclist.	[7], [11]
	Third-person	The car is about to slow down.	[4], [9] , [11]
Why-only	No subject	Road hazard ahead.	[4], [5], [7], [9]
What + Why	No subject	Emergency vehicle approaching, stopping.	[1], [6], [10]
	First-person	I will slow down because the traffic light is broken.	[3], [7], [11]
	Third-person	The car will slow down because the traffic light is broken.	[4], [8], [10], [11]

a passive position relative to the automated system. As a result, this passive role might cause the driver to fail to maintain a sense of control, leading to reduced safe-driving performance.

AV explanations were also presented 7 s before the AV's action. Du et al. (2019) [1] conducted an experiment using a fixed- base driving simulator to understand the effects of the explanation timing on drivers' perceptions, including trust, preference, anxiety and mental workload [1]. The authors found evidence that explanations provided before the AVs take action (i.e., explanations presented 7 s before the AV action) prompts the highest trust and preference compared to conditions where the explanation is given after the action (i.e., explanations presented within 1 s after the AV takes action) and where no explanation is presented. Also, the explanation provided before AV acted led to the least anxiety and workload. Likewise, Forster, Naujoks and Neukum (2017) found that an AV with speech-out messages explaining the action the AV was going to take (i.e., explanation presented 7 s before the AV's action) was rated as superior for its trust, anthropomorphism, and usability when compared to the no-explanation condition [3], [6].

Ruijten et al. (2018) [7] designed a simulator experiment to understand the effect of providing an explanation on agency and trust [7]. Explanations were provided before AV actions without specifying the time. Their results suggest that when the AV provides an explanation for its behavior, it is trusted more, is considered to be more intelligent, is seen as more human-like, and is liked more than when the AV does not provide explanation [7].

2) AV Explanation after Action

Researchers have also investigated AV explanations given after the AV acted. One example of AV explanation after action is the explanation that AV provides after rerouting in the example driving task. The specific time to provide explanations after the AV action can be organized into three groups: 1 s after the AV action, 14 s after AV action, and undefined time.

Du et al. (2019) [1] investigated the impact of explanation timing on drivers' perceptions [1]. In one condition, the AV explanation was presented 1 s after the AV took action. Results indicated that presenting the explanation after the AV action led to the lowest AV trust and preference compared to the conditions where the AV explained its action and status before acting and provided no explanation [1].

Korber et al. (2018) [2] conducted a mixed-design study that presented the explanation after the AV action (i.e., explanations presented 14 s after the AV had acted) to examine the effect of AV explanations on AV trust and acceptance [2]. The drivers' trust and acceptance in AVs did not significantly increase when contrasted with the condition where no explanation was provided, despite drivers feeling strongly that they had understood the system, the reason for the AV's action, and takeover request when they were provided the explanation [2].

Prior literature indicates that the necessity of providing an explanation after the AV action correlates with driver types and driving scenarios, but these studies did not specify an ideal time for AV explanation after actions. Shen et al. (2020) [11] conducted a study using automated vehicle driving videos to investigate in which driving scenarios people need explanations and how the critical degree of explanation shifts with situations and driver types [11]. Participants were instructed to watch short driving video clips without an explanation, and after each video they rated how necessary an explanation was for the clip. Results indicate that driver types and driving scenarios were correlated with explanation necessity. Specifically, the more aggressive drivers were, the less they needed an explanation after watching the videos. Also, an explanation was found to be highly necessary for near-crash situations.

3) Summary across AV Explanation Timing Studies

This literature can be organized into several overarching findings. One, providing AV explanations before AV actions is the most preferred timing because it can prompt positive emotional valence (e.g., trust and preference) and decrease negative feelings (e.g., anxiety and workload) [1]–[3], [5]–[7], [11]. Two, providing an AV explanation after AV actions has mixed results. On one hand, providing AV explanation after AV actions did not provide any benefits with regard to trust and preference for AVs [1], [2]. On the other hand, providing an AV explanation after AV actions did increase the driver's understanding of what just occurred [2], [11]. This was especially true for less aggressive drivers and after accidents.

III. DISCUSSION AND OPPORTUNITY

The existing literature has advanced our understanding of the effectiveness of AV explanations by investigating the effects of explanation content and timing. Nevertheless, there are several major research gaps. In this section, we present research opportunities related to driving simulation,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3130489, IEEE Access

AV explanation modality, moderating factors and mediating factors.

A. DRIVING SIMULATION

One area in need of additional research relates to the field's over-reliance on driving simulators. The literature on AV explanations has exclusively relied on driving simulators with varying levels of fidelity. Although driving simulators make it possible to conduct research safely, issues of external validity cannot be ignored. Much of what we think we know about AV explanations could be undermined if human emotions and behavior in driving simulators does not correspond to real road driving.

First, drivers in a simulator may feel differently. Previous literature showed that drivers are emotionally more relaxed driving in a simulator when compared to driving in the real world, where they maintain higher levels of vigilance [25]. Based on the literature review on AV explanation, we conclude that the why-only explanation and the before-action timing are preferable because they are more effective at promoting positive emotional attitudes. However, it is not clear whether this would be true in a real-world setting, where drivers are often more stressed. For example, prior research found that the what + why explanation was associated with the safest driving performance but induced more anxiety and annoyance compared to the why-only and what-only explanations [4]. In a real driving environment, drivers might prefer the why + what explanation over the why-only because they might be more concerned about driving safety and less concerned about being annoyed.

Second, drivers in a simulator might behave differently. For example, evidence suggests that people drive faster in simulators than in real road driving environments [26]–[28]. Also, unlike driving in a real-world setting, most people have no experience with driving simulators and might simply behave differently in a new environment [29]. From the prior research in AV explanation, we understand that the effectiveness of explanation is susceptible to other factors, such as the driving event and environment [10]. Given the discrepancy in driving behaviors, future research should investigate whether people would prefer the same types of explanations in real road conditions as they favored in the driving simulators.

Taken together, future research is needed to investigate AV explanations under real-world conditions with high external validity, for example real-road studies employing either real automated vehicles or "fake" automated vehicles under real-world conditions. The fake automated vehicle using the Wizard of Oz method hides human operators inside the vehicle or gives them remote access to the car [30]–[34]. Although there are challenges associated with this, such as AV accessibility and safety concerns, using the real-road studies to investigate the impact of AV explanations could reduce the potential for biased research outcomes.

B. AV EXPLANATION MODALITY

Another area in need of research relates to the effectiveness of AV explanation modality. A modality is the classification of a single independent channel of sensory input/output between automation and a human [35]. Previous research on AV explanations mainly employed two types of modality to provide explanations: auditory and visual. The auditory explanation was typically presented by a simulator in the form of a standard American accent with a neutral tone in a male or female voice [1], [2], [4], [5], [7], [36]. The visual explanations were presented to drivers in the form of text [8], [9].

None of the studies examined the effectiveness of a particular modality over another modality. Previous literature on vehicle display design has found differences between the effectiveness of displaying signals visually versus auditorily. For example, the auditory modality, in general, is a better option than the visual modality for providing hazard signals and for rapidly conveying the magnitude of the potential hazard [37], [38]. Unfortunately, the auditory modality has also been associated with increases in annoyance when compared to the visual modality [39]. On the other hand, the visual modality is superior to the auditory in supporting continuous awareness of the surrounding traffic and is associated with shorter warning recognition times [40]. That being said, explanations contain more complex information for drivers to comprehend than simple alerts. Therefore, future research is needed to investigate what explanation modality is best at promoting drivers' trust and safe driving. In doing so, these studies might provide insights in understanding how to best present explanations.

C. ADDITIONAL MODERATING FACTORS

Research is needed to identify the conditions that determine when AV explanations are likely to be effective. Moderators or contextual factors are essential in helping us both theoretically and practically comprehend the influence of AV explanations. For example, the moderating effects of the driving situation on the relationship between AV explanation and trust in AVs were examined [10]. The results indicated that the what + why explanation would be more effective in improving trust in AVs when drivers had higher perceptions of risk about the driving situation. Aside from the driving situations, future researchers could focus on investigating the circumstances under which the what + why explanation is advantageous versus problematic. The what + why explanation performs well in improving driving safety and promoting positive attitudes toward AVs. However, it also leads to more anxiety and annoyance [4]. This could be a result of information overload, where the what + why explanation is simply too much information. Future studies should investigate the cognitive tradeoffs between the benefits of why + what explanation and when it becomes too much information.

The study of AV explanation timing also needs to be further explored separately and jointly with AV explanation

content. Based on the prior research, we concluded that AV explanations before action are the preferred approach. However, it is not clear just how far ahead the explanation should be presented. Prior literature has investigated the impact between -explanation timings, including AV explanations before the action and after the action on AV-related outcomes. The results showed that providing AV explanation before the AV acts (i.e., 1 s prior to the AV actions, 7 s before the AV action and an undefined time) is superior to giving explanations after the AV acts (i.e., 1 s after the AV action, 14 s after AV action and an undefined time) to promote positive driver outcomes (e.g., trust). Although it seems clear that the best option is to inform the driver before the AV acts, it is less clear whether this should be 1 s or 7 s before the action, or if it even matters. For example, is an AV explanation given 7 s before the action significantly better at promoting trust than 1 s? There is also the possibility that the timing might interact with the explanation content. For example, maybe what + why is the preferred content when the explanation is given 7 s ahead rather than 1 s ahead, or perhaps the why-only explanation might be preferred when the explanation is given 1 s ahead rather than 7 s ahead. It is also not clear whether an explanation that is given too far ahead or not far enough ahead might change the driver's preference for before versus after explanations. For example, providing the driver with the explanation 1 s before the AV action might not be any different from providing the explanation 1 s after the action, or maybe it is. In all, additional research is needed to answer these important questions.

Research is needed to determine the influence of individual differences among drivers. Shen et al. (2020) [11] demonstrated that driver types (i.e., aggressive or cautious) and explanation scenarios are closely correlated with the need for an explanation [11]. However, this study only presented the correlation among these factors, which lacked information about, for instance, the extent to which the type of driver and explanation scenarios might influence the need for an explanation. It should be noted that different explanation conditions were applied to investigate the effectiveness of explanations across the previous literature, as shown in Tables IV and V. It remains unknown whether these varied situations moderate the relationship between AV explanations and outcomes.

Thus, more factors should be considered when examining the moderating mechanism between AV explanation and AV outcomes. Theoretically, understanding the moderating factors that impact AV explanation effectiveness could help us understand under what conditions the AV explanation is positive or negative. Practically, an understanding of moderators could assist AV designers in making AVs that promote trust, acceptance and safe driving.

D. MEDIATING FACTORS

Prior work investigated the relationship between AV explanations and outcomes, including attitudes and driving performance. However, research is needed to investigate the mediating mechanisms that underlie the impact of AV explanations on those outcomes. Take trust as an example; theoretical trust models have been developed to explore the potential mediating variables that explain the relationship between the explanation provided by a computer and AV-related outcomes [41]. Results indicated that personal attachment, faith, perceived understandability, perceived technical competence and perceived reliability of the system all mediate the effect of explanation on trust [41]. In the area of AVs, work is needed to identify and empirically examine the mediating variables that link AV explanations to driver outcomes such as trust. This research would allow us to better understand why AV explanations are likely to be effective or ineffective.

IV. CONCLUSION

In this paper we reviewed, organized and discussed the impact of AV explanation on driver-related outcomes in two sub-areas: explanation content and explanation timing. AV explanation content and timing are crucial factors in understanding the effectiveness of AV explanations. Theoretically, these findings contribute to the literature by highlighting the impact of AV explanations on driver-related outcomes. Practically, these findings can help in designing AVs that consistently and effectively promote positive attitudes and safe driving. Moreover, in this review we recognized and discussed several significant research gaps and future research opportunities.

REFERENCES

[1] N. Du, J. Haspiel, Q. Zhang, D. Tilbury, A. K. Pradhan, X. J. Yang, and L. P. Robert Jr, "Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload," *Transportation Research Part C: Emerging Technologies*, vol. 104, pp. 428–442, 2019.

[2] M. Körber, L. Prasch, and K. Bengler, "Why do I have to drive now? Post hoc explanations of takeover requests," *Human Factors*, vol. 60, no. 3, pp. 305–323, 2018.

[3] Y. Forster, F. Naujoks, and A. Neukum, "Increasing anthropomorphism and trust in automated driving functions by adding speech output," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 365–372.

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

[5] J. Koo, D. Shin, M. Steinert, and L. Leifer, "Understanding driver responses to voice alerts of autonomous car operations," *International Journal of Vehicle Design*, vol. 70, no. 4, pp. 377–392, 2016.

[6] F. Naujoks, Y. Forster, K. Wiedemann, and A. Neukum, "Improving usefulness of automated driving by lowering primary task interference through HMI design," *Journal of Advanced Transportation*, vol. 2017, 2017.

[7] P. A. Ruijten, J. Terken, and S. N. Chandramouli, "Enhancing trust in autonomous vehicles through intelligent user interfaces that mimic human behavior," *Multimodal Technologies and Interaction*, vol. 2, no. 4, p. 62, 2018.

[8] N. A. Hatfield, "The effects of automation transparency and ethical outcomes on user trust and blame towards fully autonomous vehicles," *(Master's thesis.) Old Dominion University, Norfolk, VA.*, 2018.

[9] G. Wiegand, M. Schmidmaier, T. Weber, Y. Liu, and H. Hussmann, "I drive-you trust: Explaining driving behavior of autonomous cars," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–6.

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

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

[12] A. Daniel and H. Doughty, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles J3016," *SAE International Standards*, 2016.

[13] M. König and L. Neumayr, "Users' resistance towards radical innovations: The case of the self-driving car," *Transportation research part F: traffic psychology and behaviour*, vol. 44, pp. 42–52, 2017.

[14] D. Howard and D. Dai, "Public perceptions of self-driving cars: The case of berkeley, california," in *Transportation research board 93rd annual meeting*, vol. 14, no. 4502, 2014, pp. 1–16.

[15] M. Raue, L. A. D'Ambrosio, C. Ward, C. Lee, C. Jacquillat, and J. F. Coughlin, "The influence of feelings while driving regular cars on the perception and acceptance of self-driving cars," *Risk analysis*, vol. 39, no. 2, pp. 358–374, 2019.

[16] C. Lee, C. Ward, M. Raue, L. D'Ambrosio, and J. F. Coughlin, "Age differences in acceptance of self-driving cars: A survey of perceptions and attitudes," in *International Conference on Human Aspects of IT for the Aged Population*. Springer, 2017, pp. 3–13.

[17] S. Chen, H. Wang, and Q. Meng, "Designing autonomous vehicle incentive program with uncertain vehicle purchase price," *Transportation Research Part C: Emerging Technologies*, vol. 103, pp. 226–245, 2019.

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

[19] L. P. Robert, "Are automated vehicles safer than manually driven cars?" *AI & SOCIETY*, vol. 34, no. 3, pp. 687–688, 2019.

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

[21] A. Toffetti, E. S. Wilschut, M. H. Martens, A. Schieben, A. Rambaldini, N. Merat, and F. Flemisch, "Citymobil: Human factor issues regarding highly automated vehicles on elane," *Transportation Research Record*, vol. 2110, no. 1, pp. 1–8, 2009.

[22] S. O.-R. A. V. S. Committee *et al.*, "Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. sae standard j3016," *SAE International. doi*, vol. 10, p. J3016_201401, 2014.

[23] E. Walter, *Cambridge Advanced Learner's Dictionary*. Cambridge University Press, 2008.

[24] N. A. Stanton and M. S. Young, "Vehicle automation and driving performance," *Ergonomics*, vol. 41, no. 7, pp. 1014–1028, 1998.

[25] H. B. Ekanayake, P. Backlund, T. Ziemke, R. Ramberg, K. P. Hewagamage, and M. Lebram, "Comparing expert driving behavior in real world and simulator contexts," *International Journal of Computer Games Technology*, vol. 2013, 2013.

[26] I. Milleville-Pennel and C. Charron, "Driving for real or on a fixed-base simulator: is it so different? an explorative study," *Presence: Teleoperators and Virtual Environments*, vol. 24, no. 1, pp. 74–91, 2015.

[27] F. Bella, "Driving simulator for speed research on two-lane rural roads," *Accident Analysis & Prevention*, vol. 40, no. 3, pp. 1078–1087, 2008.

[28] A. C. Bittner Jr, O. Simsek, W. H. Levison, and J. L. Campbell, "On-road versus simulator data in driver model development driver performance model experience," *Transportation research record*, vol. 1803, no. 1, pp. 38–44, 2002.

[29] A. Knapper, M. Christoph, M. Hagenzieker, and K. Brookhuis, "Comparing a driving simulator to the real road regarding distracted driving speed," *European journal of transport and infrastructure research*, vol. 15, no. 2, 2015.

[30] M. A. Gerber, R. Schroeter, and J. Vehns, "A video-based automated driving simulator for automotive ui prototyping, ux and behaviour research," in *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2019, pp. 14–23.

[31] S. Baltodano, S. Sibi, N. Martelaro, N. Gowda, and W. Ju, "The rrads platform: a real road autonomous driving simulator," in *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2015, pp. 281–288.

[32] N. Martelaro and W. Ju, "Woz way: Enabling real-time remote interaction prototyping & observation in on-road vehicles," in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 2017, pp. 169–182.

[33] D. Rothenbücher, J. Li, D. Sirkin, B. Mok, and W. Ju, "Ghost driver: a platform for investigating interactions between pedestrians and driverless vehicles," in *Adjunct Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2015, pp. 44–49.

[34] P. Wang, S. Sibi, B. Mok, and W. Ju, "Marionette: Enabling on-road wizard-of-oz autonomous driving studies," in *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, 2017, pp. 234–243.

[35] F. Karray, M. Alemzadeh, J. Abou Saleh, and M. N. Arab, "Human-computer interaction: Overview on state of the art," 2008.

[36] F. Naujoks, Y. Forster, K. Wiedemann, and A. Neukum, "A human-machine interface for cooperative highly automated driving," in *Advances in Human Aspects of Transportation*. Springer, 2017, pp. 585–595.

[37] Y. Cao, A. Mahr, S. Castronovo, M. Theune, C. Stahl, and C. A. Müller, "Local danger warnings for drivers: The effect of modality and level of assistance on driver reaction," in *Proceedings of the 15th international conference on Intelligent user interfaces*, 2010, pp. 239–248.

[38] W. Davenport, "Vigilance for simultaneous auditory and vibrotactile signals," *Australian Journal of Psychology*, vol. 21, no. 2, pp. 159–165, 1969.

[39] C. Nass, I.-M. Jonsson, H. Harris, B. Reaves, J. Endo, S. Brave, and L. Takayama, "Improving automotive safety by pairing driver emotion and car voice emotion," in *CHI'05 extended abstracts on Human factors in computing systems*, 2005, pp. 1973–1976.

[40] D. J. Wheatley and J. B. Hurwitz, "The use of a multi-modal interface to integrate in-vehicle information presentation," 2001.

[41] M. Madsen and S. Gregor, "Measuring human-computer trust," in *11th Australasian Conference on Information Systems*, vol. 53. Citeseer, 2000, pp. 6–8.

Qiaoning Zhang is a Ph.D student at the University of Michigan's School of Information and is a member of the Michigan Autonomous Vehicle Research Intergroup Collaboration (MAVRIC). She obtained her Master degree in Industrial and Operations Engineering from University of Michigan.

X. Jessie Yang is an Assistant Professor at the Department of Industrial and Operations Engineering, University of Michigan Ann Arbor. She obtained her PhD in Mechanical and Aerospace Engineering (Human Factors) from Nanyang Technological University, Singapore in 2014.

Lionel P. Robert Jr. is an Associate Professor in the School of Information and core faculty member of the Robotics Institute both at the University of Michigan. He is the director of the Michigan Autonomous Vehicle Research Intergroup Collaboration (MAVRIC) and an Association for Information Systems Distinguished Member.