

Investigating Explanations in Conditional and Highly Automated
Driving: The Effects of Situation Awareness and Modality

Lilit Avetisyan

Industrial and Manufacturing Systems Engineering, University of Michigan-Dearborn

Jackie Ayoub

Industrial and Manufacturing Systems Engineering, University of Michigan-Dearborn

Feng Zhou

Industrial and Manufacturing Systems Engineering, University of Michigan-Dearborn

Manuscript type: *Research Article*

Running head: *The Effects of Situation Awareness and Modality*

Word count:

Corresponding author: Feng Zhou, 4901 Evergreen Road, Dearborn, MI 48128,

Email: fezhou@umich.edu

ABSTRACT

With the level of automation increases in vehicles, such as conditional and highly automated vehicles (AVs), drivers are becoming increasingly out of the control loop, especially in unexpected driving scenarios. Although it might be not necessary to require the drivers to intervene on most occasions, it is still important to improve drivers' situation awareness (SA) in unexpected driving scenarios to improve their trust in and acceptance of AVs. In this study, we conceptualized SA at the levels of perception (SA L1), comprehension (SA L2), and projection (SA L3), and proposed an SA level-based explanation framework based on explainable AI. Then, we examined the effects of these explanations and their modalities on drivers' situational trust, cognitive workload, as well as explanation satisfaction. A three (SA levels: SA L1, SA L2 and SA L3) by two (explanation modalities: visual, visual + audio) between-subjects experiment was conducted with 340 participants recruited from Amazon Mechanical Turk. The results indicated that by designing the explanations using the proposed SA-based framework, participants could redirect their attention to the important objects in the traffic and understand their meaning for the AV system. This improved their SA and filled the gap of understanding the correspondence of AV's behavior in the particular situations which also increased their situational trust in AV. The results showed that participants reported the highest trust with SA L2 explanations, although the mental workload was assessed higher in this level. The results also provided insights into the relationship between the amount of information in explanations and modalities, showing that participants were more satisfied with visual-only explanations in the SA L1 and SA L2 conditions and were more satisfied with visual and auditory explanations in the SA L3 condition. Finally, we found that the cognitive workload was also higher in SA L2, possibly because the participants were actively interpreting the results, consistent with a higher level of situational trust. These findings demonstrated that properly designed explanations, based on our proposed SA-based framework, had significant implications for explaining AV behavior in conditional and highly automated driving.

Keywords: Explanations, Situation awareness, Modality, Automated driving.

INTRODUCTION

Automated vehicles (AV) have drawn broad interest. During the development of AV technology, artificial intelligence (AI) plays a fundamental role, but people still have difficulties in understanding or trusting the decisions made by AI due to its black-box nature (Shen et al., 2020). In conditional and highly AVs, i.e., SAE (Society of Automotive Engineers) Levels 3 and 4 AVs, (SAE, 2021), the drivers' responsibility as an active operator is switched to a passive passenger for the majority of the time. This reduces driver's SA since the attention mainly is switched to NDRT resulting in less eye-on-the-road time and harms his/her performance when intervention is needed (Endsley, 2019; Frison et al., 2019). Clark et.al. (2017) showed that in unexpected takeover scenarios drivers who successfully took over the control within an acceptable time frame had a higher level of SA and responded faster than drivers who did not.

When drivers are out of the control loop, they will have a low level of SA, making it difficult for them to comprehend AV's behavior in unexpected situations. Moreover, it limits their ability to successfully take over control in critical situations, leading to accidents. For example, by analyzing Uber's AV fatal accident in Arizona (Garcia, 2018), it was revealed that the driver failed to take over control of the AV because she was engaged on her phone and was not aware of the pedestrian crossing the road. Regardless of who was responsible for the accident, such cases overall had negative impacts on trust in and public acceptance of AV. In particular, being unaware of the situation, drivers tend to interpret the AV's unexpected behavior as system malfunction that leads to trust issues in AVs. Hence, when the automated mode is on, the AVs should provide sufficient information to increase drivers' SA up to the "in-the-loop" level for proper understanding of the situation and to ensure that the situation is under control. It is our belief, that improving the SA level will mitigate the unexpectedness and subsequent trust issues.

In complex intelligent systems, the lack of information about system behavior or misunderstanding of automation creates trust issues (Norman, 1990), especially when

the system acts outside of expectations. To foster trust in and acceptance of AV, it is crucial to make the system transparent for drivers and provide appropriate feedback on the system's behavior. One of the concepts proposed to make black-box systems transparent is explainable artificial intelligence (XAI). It contributes to human-AI interaction by providing information about the main factors, which affect AI decisions and its future behavior. The AV, as a complex AI system, also needs to be explained for better human-AV team performance, since it is important to keep an appropriate level of trust in automation and effectively manage uncertainty. Previous studies already confirmed the necessity of feedback in autonomous driving (Seppelt & Lee, 2019; Wiegand et al., 2020; Wintersberger, Janotta, Peintner, Löcken, & Riener, 2021). For example, Wintersberger et al. (2021) found that regardless of the trust in AV, people still preferred to be informed about forthcoming strategies and maneuvers.

Many human factors researchers made use of explanations of AVs' behavior and system feedback and status to help build the driver's mental model of the vehicle (Koo et al., 2016, 2015; Petersen et al., 2019). For example, Koo et al. (2015) found that "why" (describing the reasoning for actions, e.g., "obstacle ahead") information improved participants' understanding, trust, and performance, and "why" and "how" (describing actions, e.g., "the car is breaking") information led to safest driving performance. Du et al. (2021) used explanations about future actions of the vehicle (i.e., "what will" information) and why the vehicle requested the driver to take over (i.e., "why" information) and the combination of the two during SAE Level 3 takeover transition periods. They found that "what will" information and "what will" + "why" information improved drivers' perceived ease of use and perceived usefulness, leading to potentially better takeover performance. These studies emphasized drivers' informational needs about the AV decisions and the driving scenarios during the takeover transition process. However, there is still no direct evidence to support that such information improved drivers' SA and eventually human-AV performance.

The present study

As described above, previous studies addressed different issues in AVs (i.e., trust and takeover performance) through explanations, and provided important implications for designing AV systems. However, these solutions/models did not systematically assess how they improve drivers' trust with a minimal level of cognitive workload. Therefore, it is necessary to frame the explanations theoretically to support human-AV interaction.

In this work, we proposed an SA-based explanation for the AV's black-box system based on Endsley (1995) and Sanneman and Shah (2020). First, we designed the explanations according to Endsley to support three levels of information process, which states that people process information in three hierarchical levels: 1) Level 1 SA: Perception of the elements in the environment, 2) Level 2 SA: Comprehension of the current situation, 3) Level 3 SA: Projection of future status in order to be up-to-date in the dynamic environment. Individuals need three levels of SA in their decision-making process in complex dynamic human-machine interaction in various scenarios. Second, we designed the explanations to understand the decision-making process of the AV's black-box system according to Sanneman and Shah's (2020)'s mixed input/output principles as follows: 1) "what" environmental input AV used to make a decision, 2) "how" the AV understands the input and "how" the input influences AV behavior and 3) "what would happen" if AV did not act in that way.

We hypothesized that explaining AV behaviors to accommodate drivers' informational needs based on the above theories with three levels of SA would result in different levels of understanding and human-AV performance. We expected that our explanation framework would foster trust with a relatively less increase in mental workload compared to the previous approaches due to the mapping of explanations to information processing levels. In order to test the hypothesis, we designed a three by two between-subjects experiment, where three types of explanations were manipulated to three levels of SA with two modalities (visual, visual + auditory) across six scenarios. We examined the effects of explanations in the form of three levels of SA on drivers' situational trust, cognitive workload, and explanation satisfaction."

Related Work

Explanations in AV

In human factors research, explanations about the AV's behavior, system feedback and status, and driving scenarios were designed and provided to improve the transparency of system decisions and driver trust. For instance, Wintersberger et al. (2019) showed that augmented reality by coding traffic objects and future vehicle actions increased automation transparency and improved user trust and acceptance. Koo et al. (2015) designed three different types of information to explain AV behavior about: 1) "how" the car was acting, 2) "why" the car was acting and 3) "how" + "why" the car was acting. Authors investigated AV-driver interaction in a scenario where the AV took control from the driver and suddenly braked to avoid collision with an obstacle. They explained the AV behavior before the AV started acting, and found that "how" + "why" information resulted in the safest AV-driver cooperation, but also produced the greatest cognitive workload than other explanations, which could lead to confusion and anxiety. The "how" only information led to worse driving performance and unsafe cooperation since the drivers tried to take the control back from the AV but did not understand why the AV behaved in that way. Mackay et al.'s (2019) investigation into different amounts of feedback found that "more information does not necessarily lead to more trust and may, in fact, negatively affect cognitive load". Taehyun et al. (2020) stated that type of explanation significantly affects trust in AVs and suggested an explanation format based on the attribution theory (Weiner, 1979). They found that perceived risk moderated the effect of explanations on trust, i.e., attributional explanations led to the highest level of trust in low perceived risk compared to no or simple explanations.

In addition, the timing of the explanations (i.e., before or after particular action) also plays an important role in trust and acceptance in AVs. For example, Körber et al. (2018) provided explanations of the causes of takeover requests after the takeover transitions, which led to no decrease in trust or acceptance, but improved participants' understanding of system behaviors. Koo et al. (2015) argued that explanations should

be provided ahead of an event which also was supported by Haspiel et al. (2018) and Du et al. (2019) studies, who found that explanations provided before the AV's action promoted more trust than those provided afterward. Thus, it is recommended that we should provide explanations before the vehicle takes action.

Other types of factors, such as forms, contents, and modalities of the explanations also play important roles in explanations in AVs. Wang et al. (2020) explored how information modality influenced driver's performance and showed that both visual and auditory modalities had a significant influence, but on different aspects of driver's performance. In particular, visual information boosted performance efficiency and auditory information decreased reaction time. Seppelt and Lee (2019) showed that continuous feedback helped drivers to be involved in the loop of system performance and operations. Consistent with the multiple resource theory (Wickens, 2008a), they found that the combined visual-auditory interface performed the best regarding drivers' confidence and trust.

Situation awareness and the out-of-the-loop problem

Merat et al. (2019) differentiated three kinds of loops in AV systems and described them as follows: 1) A driver was in the control loop when he/she was both in the physical control and monitoring the driving task, 2) a driver was on the control loop when the driver was only monitoring the driving task, and 3) a driver was out of the control loop as long as he/she was not monitoring the driving task. Thus, the out-of-the-loop problem in AVs describes the situation when the driver is not actively monitoring the system or the environment (Radlmayr et al., 2014). This issue is mostly due to driver's overtrust in AVs, since a certain level of "control" is needed to properly respond to situational changes or to reduce uncertainty in automated driving, such as monitoring and takeover control (Du, Ayoub, et al., 2019; Du, Yang, & Zhou, 2020; Du, Zhou, et al., 2020).

Merat et al. (2019) emphasized that a key aspect to be in the control loop was the drivers' attention and cognitive responses to the changes in the system and in the

dynamic environment, which was characterized by the driver's SA. In other words, when the driver is not in the control loop of the AV, the SA of system status and the driving environment may be reduced (Sebok & Wickens, 2017; Zhou, Yang, & de Winter, 2021; Zhou, Yang, & Zhang, 2019). Even if the driver is on the control loop (i.e., not in physical control of the vehicle, but monitoring the driving situation) (Merat et al., 2019), he/she becomes a passive information processor, which would negatively affect the operator's understanding and comprehension (SA Level 2) of dynamic changes in the system even though the driver is aware of low-level information (SA Level 1) (Endsley & Kiris, 1995). This is further aggravated by the black-box decision-making process of the AV and the monotonicity of automated driving, which lead to low vigilance and even drowsiness (Zhou et al., 2020; Zhou, Alsaïd, et al., 2021). However, SAE Levels 3-4 AVs allow drivers to conduct non-driving-related tasks without monitoring the driving task (Ayoub, Zhou, Bao, & Yang, 2019). In order to resolve such conflicts (i.e., conducting NDRTs in AVs vs. requiring a certain level of SA in AVs), explanations are needed to help drivers resume their SA in time when a certain level of "control" or understanding is needed to respond the situational changes, especially during unexpected driving scenarios.

METHOD

Participants

In total, 340 participants (151 females and 189 males; Age = 39.0 ± 11.4 years old) in the United States participated in this study. All the participants were recruited from Amazon Mechanical Turk (MTurk) with a valid US driver's license. On average, participants had 15 ± 11.8 years of driving experience and the driving frequency was 5 ± 1 days per week. They were randomly assigned to one of the seven conditions as shown in Table 1, where L1, L2, and L3 conditions were mapped closely to three SA levels proposed by Endsley. More detailed information about the experiment conditions is described in the "Scenario Design" section. This study was approved by the Institutional Review Board at the University of Michigan. Each participant was

compensated with \$2 upon completion of the study. The average completion time of the survey was about 26 minutes across the conditions.

Table 1: *Experimental design with Modality and SA level as independent variables. The modality factor had two levels: 1) Visual, i.e., the explanation was given only in text format, and 2) Visual + Audio, i.e., the explanation was given in text and voice format simultaneously. The SA level factor had three levels: 1) SA L1, i.e., the explanation included only SA level 1 information (i.e., perception), 2) SA L2, i.e., the explanation included SA level 1 + level 2 information (i.e., perception and comprehension), and 3) SA L3, i.e., the explanation included SA level 1 + level 2 + level 3 information (i.e., perception, comprehension, and projection). Table cells represent the treated conditions in the experiment.*

SA Level	Modality	
	Visual	Visual + Audio
SA L1	Text SA L1	Text + audio SA L1
SA L2	Text SA L2	Text + audio SA L2
SA L3	Text SA L3	Text + audio SA L3

* A control condition was included in the experiment where participants did not receive any explanation.

Apparatus

The study was conducted using a survey developed in Qualtrics (Provo, UT) and was published in MTurk. The survey was designed to evaluate the effects of SA and explanation modality on participants' situational trust, explanation satisfaction, and mental workload in uncertain situations while driving an AV. The driving scenarios were presented in videos created in the CarMaker autonomous driving simulation environment (Karlsruhe, DE).

Table 2: *Dependent variables*

Measure	Description	Scale
Trust	Measured at the end of each scenario	STS-AD
Explanation Satisfaction	Measured at the end of each scenario	Explanation satisfaction scale
Mental Workload	Measured once participants watched all the 6 scenarios	DALI

Experimental design

Independent variables. The experiment was a three (SA level: SA L1, SA L2, and SA L3) by two (modality: visual, visual + auditory) between-subjects factorial design with 6 scenarios. Alongside the 6 experimental conditions, a control condition with no explanations was also tested. The independent variables were the three levels of explanations mapped to three SA levels presented to the participants according to Endsley's SA model (Endsley, 1995) and in two types of modalities, i.e., visual and visual + auditory. During the experiment, the participants' SA was measured through the Situation Awareness Global Assessment Technique (SAGAT) (Endsley, 1988). The SAGAT is a freeze-probe technique that requires pausing the simulation and asking a series of questions to assess the participants' awareness of the current situation. For each scenario, three different questions were developed to test the participants' perception of surrounding objects, comprehension of the current situation, and projection of the future state for that uncertain situation. All the questions designed for the SAGAT technique were developed based on a previous study (van den Beukel & van der Voort, 2017). Table 3 shows an example of multiple-choice questions for the training scenario (see Table 4). Regardless of the experiment conditions, for each scenario, three SA questions were included in the survey corresponding to three levels of SA. The participants obtained one point if they answered the question correctly. With three questions for each scenario, the participants could get as many as 18 points, indicating perfect SA.

Table 3: *Example questions for the training scenario to measure SA with a SAGAT Questionnaire.*

Level of SA	Question	Options
Perception	The simulation just “froze”. Which road user was in front of the AV?	1) Bus, 2) <u>Pedestrian</u> , 3) Cyclist, 4) I don’t know, 5) Other
Compre- hension	What caused you to seek your attention in this situation?	1) <u>Pedestrian’s intention to cross the street</u> , 2) Approaching heavy traffic, 3) Approaching closed road, 4) Faulty road lanes, 5) I don’t know, 6) Other
Projection	If the simulation resumes after this “freeze”, what situation would require your extra attention or intervention?	1) Other road user’s violations, 2) <u>AV’s possibility to hit pedestrian</u> , 3) Impeding the traffic by stopping at intersection, 4) I don’t know, 5) Other

* The underlined option indicates the correct answers.

Dependent measures

The dependent variables in this study were situational trust, mental workload, and subjective satisfaction with explanations. Situational trust was measured by the self-reported Situational Trust Scale for Automated Driving (STS-AD) (Holthausen, Wintersberger, Walker, & Riener, 2020). The model evaluates situational trust in six categories: trust, performance, non-driving related task (NDRT), risk, judgment, and reaction, by asking the following questions: 1) I trusted the automation in this situation, 2) I would have performed better than the AV in this situation, 3) In this situation, the AV performed well enough for me to engage in other activities, 4) The situation was risky, 5) The AV made a safe judgment in this situation, and 6) The AV reacted appropriately to the environment. All the six STS-AD scales were measured with a 7-point Likert scale. Situational trust was measured right after the participant watched one video that depicted a specific driving scenario. Thus, it was measured six times for six scenarios.

To understand the subjective satisfaction of the given explanations, the explanation satisfaction scale developed by Hoffman et al. (2018) was used. In this study, it was presented to the participants with five items and was measured with a 7-point Likert scale. The following items were included: This explanation of how the AV behavior was 1) satisfying, 2) had sufficient details, 3) contained irrelevant details,

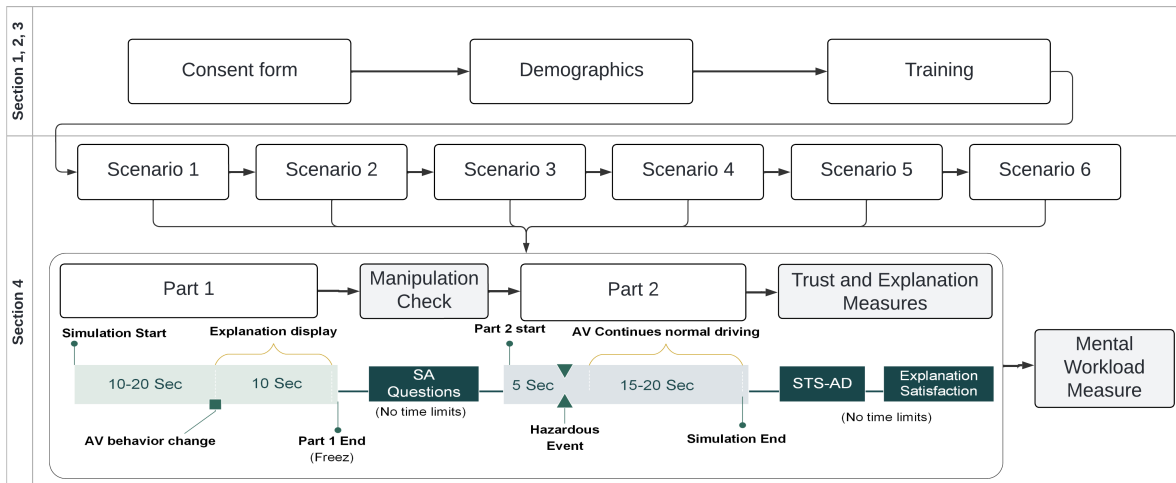


Figure 1. Survey procedure.

4) was helpful, 5) let me judge when I should trust and not trust the AV. Explanation satisfaction was also measured once right after the participant watched one specific driving scenario. Thus, it was measured six times for six scenarios.

The mental workload was measured using the driving activity load index (DALI) (Pauzié, 2008), which is a revised version of the NASA-TLX and specifically adapted to the driving tasks. DALI includes six factors: attention, visual, auditory, temporal, interference, and stress. In order to reduce the time of taking the survey, the cognitive workload was only measured once at the end of the survey using a 7-point Likert scale when the participants watched all the six scenarios. In the control and text-only scenarios, the auditory demand was removed.

Survey Design and Procedure

The survey consisted of four sections as illustrated in Figure 1. The first section included a consent form. In the second section, the participants filled in a set of demographic questions. The third section was a training session, where the participants were given one simulation video example not used in the test session with three SA questions. Since the SA questions were designed based on the SAGAT technique, the freeze-probe technique was imitated for each scenario by dividing the simulation into two parts representing before and after the freeze situations. The fourth test section included six AV driving scenarios as shown in Table 4. The participants watched the

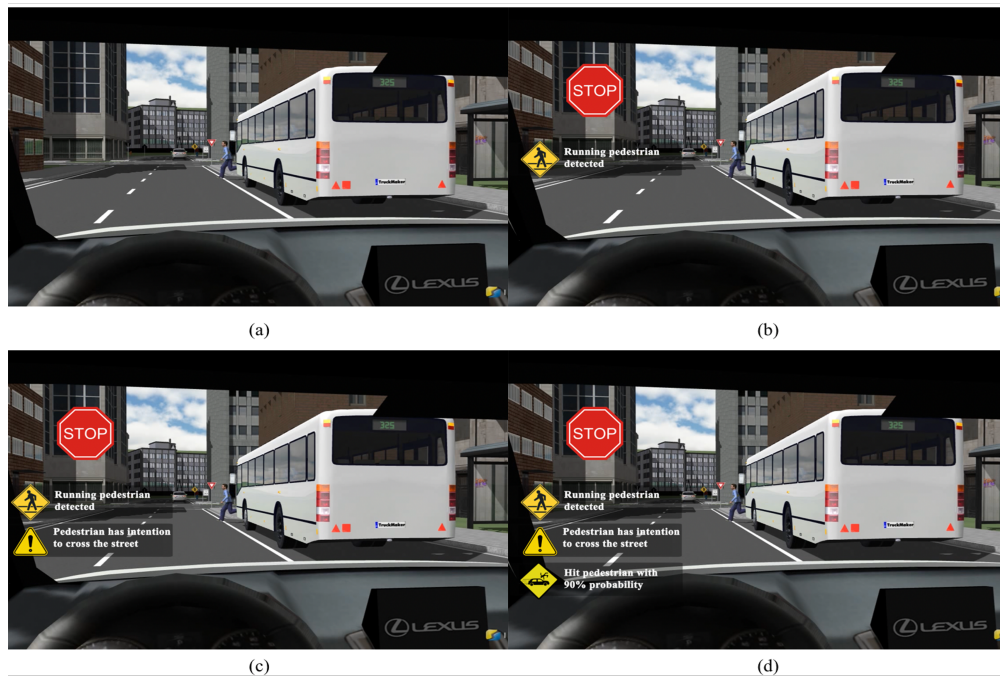


Figure 2. Presented explanations S2 in (a) control, (b) SA L1, (c) SA L2 and (d) SA L3 conditions (see S3 L3: <https://youtu.be/GNL2cMK5Lyk>).

first part of each simulation video and answered three questions about their SA about the driving scenario (see Table 3). Then, they watched the second part of the video where they could see what happened actually. After each scenario, the participants evaluated their situational trust in AVs using the STS-AD scale and rated the given explanation(s) using the explanation satisfaction scale. After finishing all the six scenarios, the participants were required to report their mental workload about the explanations.

Scenario Design

Participants' trust in AVs' scenarios was investigated by manipulating their SA using three SA levels (Endsley, 1995) in different scenarios. All the situations were extracted from real driving scenarios and from Wiegand et al.'s work (2020), where they explored the necessity of the explanations in unexpected situations while driving an AV. Seven scenarios were identified and simulation videos were created to visualize the situations (see Table 4). In each scenario, the corresponding information was embedded into the video explaining the current situation before the AV started its actions. In this

Table 4: *Scenarios with description in this study*

Name	Scenario	Description and Link
Training	Reluctant to turn right due to a pedestrian	City: The AV stops before turning right, and a pedestrian stands on the other side of the street and moves a little. There is no crosswalk. The AV slowly turns with intermittent stopping. https://youtu.be/B3Zw7-kZzoY
S1	Long wait at the intersection to turn left	Highway: The AV approaches an intersection with a green traffic light. It stops behind the traffic light, and then moves a bit. After about 10 seconds, the AV finally turns left after an oncoming car passes. https://youtu.be/PfpxPfmePg
S2	The AV stops and the pedestrian crosses	City: While driving, the AV stops abruptly. It waits. After seconds, a pedestrian crosses the street behind the bus. The AV continues driving. https://youtu.be/i9nt3FvqbnM
S3	Unexpected stop due to an emergency vehicle	City: The AV stops. In some distance, there is a green traffic light. After a while, an emergency vehicle passes with the siren on. The AV waits for about 2 more seconds and continues driving. https://youtu.be/XmSrxEYeySo
S4	Strong and abrupt braking to reach the speed limit	City: The AV enters the city and brakes abruptly and strongly to reach the speed limit. https://youtu.be/b5jrT4Mx9bg
S5	Early lane change due to heavy traffic	Highway: The AV changes to the right lane far away from the turn and it detects heavy traffic on the defined route. https://youtu.be/0kQw498WK20
S6	The AV waits for a long time before merging	Highway: The AV slows down and stops. It needs to merge with the highway and waits for its chance with a safe distance while the AV's intention in merging lanes is not clear. Traffic is overloaded. https://youtu.be/L8I8ULMcuYw

work, explanation modality was also explored by adding voice-over to simulations. In visual+auditory conditions, an auditory message with a synthesized female voice was added to provide the same situational explanations simultaneously with the visual explanations. Figure 2 illustrates the simulations for the S2 scenario (see Table 4) correspondingly for the control, SA L1, SA L2, and SA L3 conditions. In the control condition, no explanation was given. The SA L1 condition provided information explaining the perception of the current environment, including the surrounding objects which influenced on the AV's behavior. In the SA L2 condition, additional information was used to explain how the AV understood the surrounding environment. The SA L3

condition included all the information from SA L2 and added extra information about how that might affect the AV's behavior in the future.

Data Analysis

Statistical analysis was conducted using the R language in RStudio. A two-way analysis of variance (ANOVA) was used to analyze the effects of the explanations on situational trust, explanation satisfaction, and mental workload. The alpha was set at 0.05 for all the statistical tests. Post-hoc analysis was conducted with Tukey's HSD test.

RESULTS

Manipulation Check

In this study, the effect of the provided information on SA was explored with the control condition and three SA levels, where the participant's SA was measured by the number of correct responses throughout the experiment. A two-way ANOVA test showed that there was a significant main effect of SA levels ($F(3, 333) = 38.23, p = .000, \eta^2 = .253$) and modalities ($F(1, 333) = 4.26, p = .040, \eta^2 = .009$) (see Figure 3). There was no significant interaction effect between SA levels and modalities ($F(2, 333) = 0.28, p = .752$). The post-hoc analysis showed that SA was significantly higher in SA L1, L2, and L3 conditions compared to the control condition, and significantly higher in the visual + auditory modality ($p = .040$) compared to the visual-only modality. Figure 3 illustrates the mean SA scores across different experimental conditions.

Situational Trust

The means of the STS-AD over all six scenarios were calculated and analyzed with a two-way ANOVA. Results showed that the main effect of SA levels was significant ($F(2, 294) = 3.93, p = .020, \eta^2 = .029$) whereas the main effect of modalities ($F(1, 294) = .07, p = .789, \eta^2 = .000$) and the interaction effect ($F(2, 294) = 1.31, p = .272, \eta^2 = .007$) were not significant (see Figure 4). The post-hoc

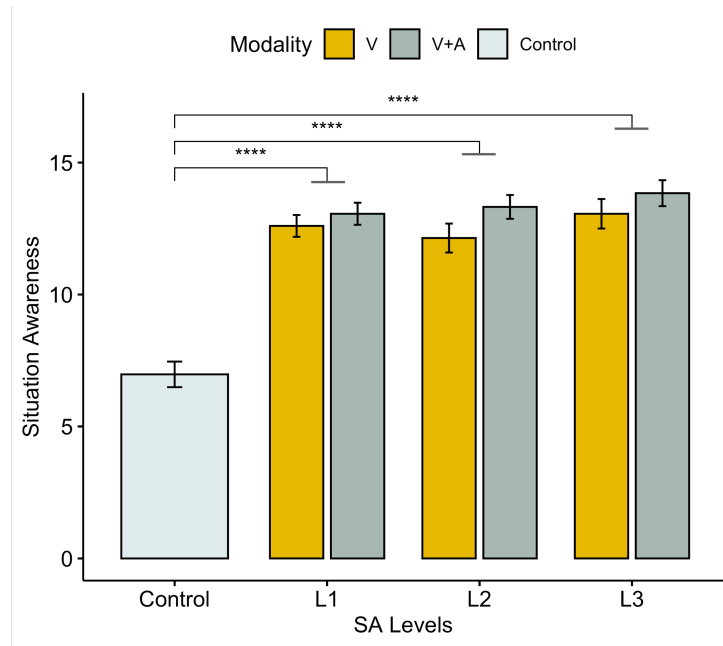


Figure 3. Mean SA scores at different conditions and explanation modalities with standard error, where ‘****’ indicates $p < 0.001$.

analysis showed that STS-AD in SA L2 was significantly higher than in SA L1 ($p = .036$). Specifically, STS-AD in Text SA L2 was significantly ($p = .040$) higher than that in Text + Voice SA L1. And STS-AD was significantly higher ($p = .047$) in SA L2 than that in SA L3. Specifically, STS-AD in Text SA L2 was marginally ($p = .052$) higher than that in Text SA L3. Compared to the control condition, it was found that only SA L2 was significantly higher ($p = .011$) mainly due to the visual-only modality ($p = .026$). As for the visual + auditory modality, the difference was not significant ($p = .131$).

Explanation Satisfaction

With regard to explanation satisfaction, the two-way ANOVA showed a significant interaction effect ($F(2, 294) = 4.53, p = .012, \eta^2 = .030$). The post-hoc analysis showed that the participants were significantly more satisfied with the given explanations in the SA L1 ($p = .014$) and SA L2 ($p = .043$) conditions compared to the SA L3 condition when explanations were presented in the visual-only modality. Furthermore, in the SA L3 condition, when a comparatively large amount of explanation information was presented, a significant effect of explanation modality was found that the visual +

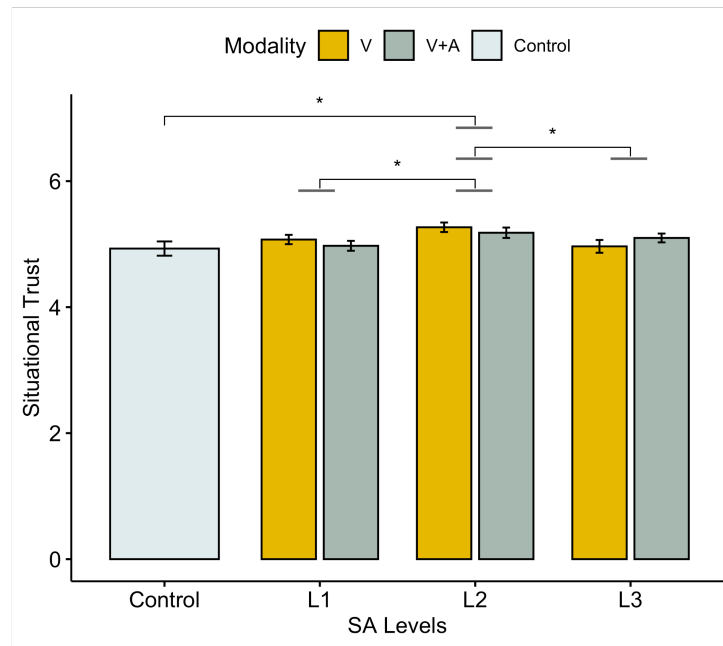


Figure 4. Overall mean and standard error of situational trust measured by the SA levels and modalities, where ‘*’ indicates $p < 0.05$.

auditory condition resulted in a higher satisfaction score compared to the visual-only ($p = .009$) condition (see Figure 5).

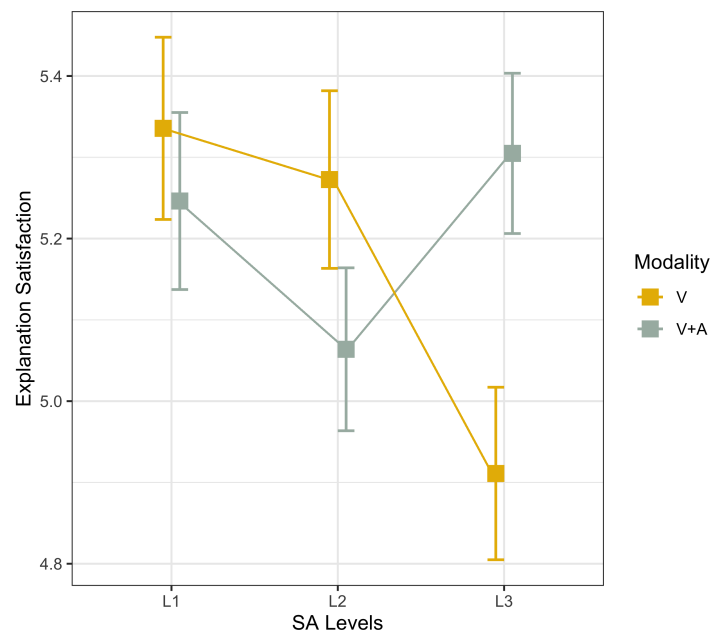


Figure 5. Interaction effect of SA levels and modalities with standard error on explanation satisfaction.

Mental Workload

The participants' self-reported mental workload was analyzed using the mean values of all the six DALI factors. As shown in Figure 6, we found a significant main effect of SA levels ($F(2, 294) = 3.70, p = .026, \eta^2 = .024$) that participants' mental workload was significantly higher ($p = .018$) in the SA L2 condition than that in the SA L1 condition and than that in the control condition ($p = .009$). Specifically, we found that participants' mental workload in the Text SA L2 condition was significantly ($p = .016$) higher than that in the Text SA L1 condition and was significantly ($p = .012$) higher than that in the control condition. Thus, the significant differences were mainly caused by the visual-only modality.

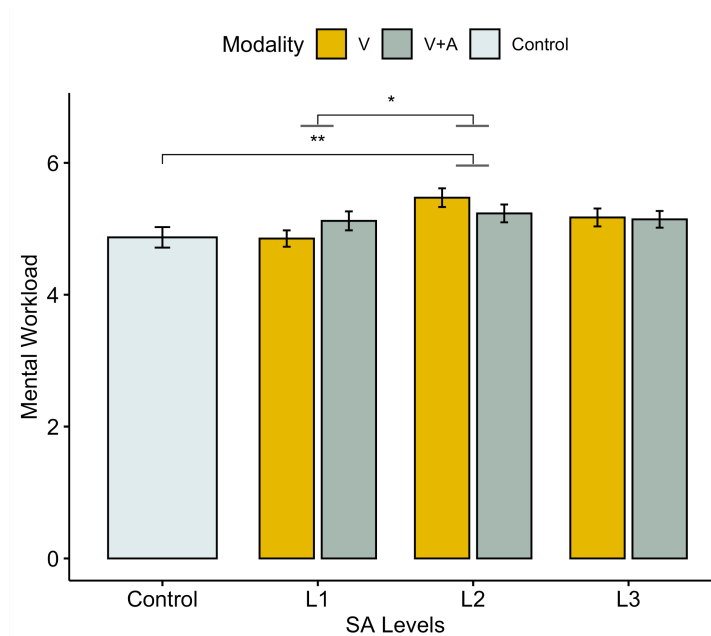


Figure 6. Overall mean and standard error of mental workload measured by the SA level and modality, where ‘*’ indicates $p < 0.05$ and ‘**’ indicates $p < 0.01$.

DISCUSSIONS

The Effects of SA

In this study, we investigated the effects of SA explanations and modalities on situational trust, explanation satisfaction, and mental workload in AVs. First, our results partially supported that SA levels positively affected participants' situational trust (see Figure 4) and SA L2 led to the highest level of situational trust. In this sense,

situational trust appeared to be sensitive to SA. In particular, the participants' trust was significantly higher in SA L2 compared to SA L1 and L3, where the given information was either too little to foster the participants' perception and comprehension of the current situation or was redundant to notably improve trust (Mackay et al., 2019). One possible reason might be the out-of-the-loop problem, as Endsley et al. (1995) found that SA L2 was the most negatively affected level by automation, where people's understanding of the situation significantly decreased, pushing them out of the control loop. When SA L2 explanations were provided to help the participants understand the situations and bring them back to the control loop, their situational trust was significantly improved. Besides, consistent with Endsley (1995), the participants might comprehend and project the future state at the same stage in SA L2, which indicates that the participants might already receive information that is supposed to receive in SA L3. For instance, in the scenario 2 (see Table 4) comparing the SA L2 explanation (i.e., L1: "Running pedestrian detected", L2: "Pedestrian has an intention to cross the street"), and SA L3 (i.e., L1, L2, and L3: "90% risk of hitting a pedestrian") explanations, the participants might project the risk of accident at L2, hence the L3 explanation was not useful. Therefore, there was also no significant difference between SA L2 and SA L3 in terms of cognitive processing as shown in Figure 6.

With regard to the interaction effect of SA levels and modalities on explanation satisfaction (see Figure 5), the participants were more satisfied with the text explanations in SA L1 and L2 might be due to the machine-generated voice. As Tsimhoni, Green and Lai, (2001) showed that natural speech led to a better comprehension of the given information compared to synthesized speech. However, participants were more satisfied with the combined visual and auditory explanations in SA L3. This result was supported by the information processing theory (Wickens, 2008b) that it was easy to comprehend a large amount of information when more than one sensory resource (i.e., visual and auditory) was used while the participants might be annoyed to have redundant explanations with less information.

For cognitive workload, we found that participants had a higher cognitive workload in the SA L2 condition, especially the visual-only explanations, compared to the control and SA L1 conditions. One possible reason might be that the participants with explanations corresponding to SA L2 were actively interpreting the information to understand the driving scenarios, which improved their situational trust (see Figure 4). However, regardless of the extra information, SA L1 and SA L3 had similar levels of cognitive workload as the control group which might be due to the experiment design.

Implications

We proposed to explain AV behavior based on the three levels of SA and XAI theoretically to satisfy their informational needs in unexpected scenarios, and empirically explored its effects on human-AV interaction. Considering the AV as a black-box AI system, the properly-designed explanations based on the SA framework helped to define which components in the system should be explained to meet drivers' informational needs in order to understand AV's behavior. While previous studies have focused on "how", "why" and "what" information for explanations empirically (Du et al., 2021; Koo et al., 2016, 2015), this SA-based model focused more on XAI concepts and reduced the complexity of the situations to understand how the AI system came to that particular decision systematically.

During the interaction between the driver and the AV, it is important that the AV provides explanations with different levels of SA for the driver to understand its decision-making process. As pointed out by Sanneman and Shah (2020), the key point is how to map such explanations into the needed three SA levels when designing such a black-box AV system as an XAI system. At SA level 1, we need to provide explanations about what objects are perceived from the environment to explain the effects of external factors on the decision-making process. At SA level 2, we should explain how the AV understands the situation by taking the perceived objects and their actions into consideration. At SA level 3, we might consider what actions would the AV and other road users take in the near future. Our explanations attempted to be designed based on

the theory-based SA model to satisfy drivers' informational needs and benefit them by improving their trust with a minimal level of cognitive workload.

Limitations and Future Work

This study also has limitations that can be examined in future studies. First, the experiment was conducted in a low-fidelity setting on MTurk due to the COVID-19 pandemic. The SA was measured with the SAGAT technique (Endsley, 1995) and we found that participants' SA was notably improved compared to the control condition. However, we could not identify significant differences among the three SA levels based on the provided explanations. One of the possible reasons might be that the data was collected on MTurk, where the scenarios were relatively short (30-45 seconds) and the fidelity was relatively low in the experiment. This potentially reduced the participants' engagement level. Another reason might be the absence of non-driving related tasks due to the difficulty in controlling participants when the experiment was conducted on MTurk, which allowed the participants to continuously monitor the ride. Nevertheless, the significant differences in SA between the control conditions and others indicated the importance of simple explanations in improving SA. Further investigations are needed to understand the effects of different explanations on SA and subsequently on trust, mental workload, explanation satisfaction, and the joint performance of the human-AV team in high-fidelity driving simulators. Second, only self-reported measures were used to evaluate the trust and mental workload. Additional measures, such as physiological measures (e.g., galvanic skin response (Du, Yang, & Zhou, 2020), eye-tracking (de Winter, Eisma, Cabrall, Hancock, & Stanton, 2019)) can be included in future studies. Third, only a limited number of scenarios were tested in the experiment with low to moderate risks. Future studies can explore more scenarios with different levels of risk. Fourth, since the experiment was conducted as a between-subjects design, the participants experienced only one of the SA levels, the results might be affected by individual differences and the low-fidelity of the experiment setting.

CONCLUSION

In this study, we designed an SA-based explanation framework to help drivers understand the driving situations and map the AV's behavior properly to the situation. By exploring participants' situational trust, cognitive workload, and explanation satisfaction, we evaluated the effectiveness of the framework in three SA levels and two modalities. Based on the results, it was partially supported that SA-based explanations improved participants' situational trust. Among three levels, SA L2 resulted in higher situational trust and mental workload regardless of the explanation modality. However, modality preferences were changed from visual to visual and audio due to the explanation amount in SA L3. Overall, the results confirmed that the properly-designed explanations based on the SA-based framework helped orient drivers in the unexpected situation and assess the AVs' behavior accurately leading to higher trust and acceptance of these vehicles.

References

- Ayoub, J., Zhou, F., Bao, S., & Yang, X. J. (2019). From manual driving to automated driving: A review of 10 years of autou. In *Proceedings of the 11th international conference on automotive user interfaces and interactive vehicular applications* (pp. 70–90).
- Clark, H., McLaughlin, A. C., & Feng, J. (2017). Situational awareness and time to takeover: Exploring an alternative method to measure engagement with high-level automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 61(1), 1452-1456. Retrieved from <https://doi.org/10.1177/1541931213601848> doi: 10.1177/1541931213601848
- de Winter, J. C., Eisma, Y. B., Cabrall, C., Hancock, P. A., & Stanton, N. A. (2019). Situation awareness based on eye movements in relation to the task environment. *Cognition, Technology & Work*, 21(1), 99–111.
- Du, N., Ayoub, J., Zhou, F., Pradhan, A., Robert Jr, L., Tilbury, D., ... Yang, X. J. (2019). Examining the impacts of drivers' emotions on takeover readiness and performance in highly automated driving. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.
- Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A. K., Yang, X. J., & Robert Jr, L. P. (2019). Look who's talking now: Implications of av's explanations on driver's trust, av preference, anxiety and mental workload. *Transportation research part C: emerging technologies*, 104, 428–442.
- Du, N., Yang, X. J., & Zhou, F. (2020). Psychophysiological responses to takeover requests in conditionally automated driving. *Accident Analysis & Prevention*, 148, 105804.
- Du, N., Zhou, F., Pulver, E. M., Tilbury, D. M., Robert, L. P., Pradhan, A. K., & Yang, X. J. (2020). Examining the effects of emotional valence and arousal on takeover performance in conditionally automated driving. *Transportation research part C: emerging technologies*, 112, 78–87.

Du, N., Zhou, F., Tilbury, D., Robert, P. L., & Yang, X. J. (2021). Designing alert systems in takeover transitions: The effects of display information and modality. In *Proceedings of the 13th international conference on automotive user interfaces and interactive vehicular applications* (pp. 1–13).

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the human factors society annual meeting* (Vol. 32, pp. 97–101).

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32-64.

Endsley, M. R. (2019). Situation awareness in future autonomous vehicles: Beware of the unexpected. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th congress of the international ergonomics association (iea 2018)* (pp. 303–309). Cham: Springer International Publishing.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, *37*(2), 381-394.

Frison, Anna-Katharina, Wintersberger, Philipp, Liu, Tianjia, . . . Andreas (2019). Why do you like to drive automated? a context-dependent analysis of highly automated driving to elaborate requirements for intelligent user interfaces. In *Proceedings of the 24th international conference on intelligent user interfaces* (p. 528–537). New York, NY, USA: Association for Computing Machinery.

Garcia, R. (2018). *Video shows uber operator moments before self-driving car crash that killed pedestrian.*

<https://www.usatoday.com/story/tech/nation-now/2018/03/21/fatal-uber-crash/447770>

Retrieved 2018-03-21, from

<https://www.usatoday.com/story/tech/nation-now/2018/03/21/fatal-uber-crash/447770>

Ha, T., Kim, S., Seo, D., & Lee, S. (2020). Effects of explanation types and perceived risk on trust in autonomous vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour*, *73*, 271-280.

Haspiel, J., Du, N., Meyerson, J., Robert Jr, L. P., Tilbury, D., Yang, X. J., & Pradhan,

- A. K. (2018). Explanations and expectations: Trust building in automated vehicles. In *Companion of the 2018 acm/ieee international conference on human-robot interaction* (pp. 119–120).
- Hoffman, R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable ai: Challenges and prospects. *ArXiv, abs/1812.04608*.
- Holthausen, B. E., Wintersberger, P., Walker, B. N., & Riener, A. (2020). Situational trust scale for automated driving (sts-ad): Development and initial validation. In *12th international conference on automotive user interfaces and interactive vehicular applications* (pp. 40–47).
- Koo, Jeamin, Shin, Dongjun, Steinert, Martin, . . . Larry (2016). Understanding driver responses to voice alerts of autonomous car operations. *International journal of vehicle design, 70*(4), 377–392.
- Koo, 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*, 269-275.
- Körber, M., Prasch, L., & Bengler, K. (2018). Why do i have to drive now? post hoc explanations of takeover requests. *Human factors, 60*(3), 305–323.
- Mackay, A., Fortes, I., Santos, C., Machado, D., Barbosa, P., Boas, V., . . . Sousa, E. (2019, 06). The impact of autonomous vehicles’ active feedback on trust. In (p. 342-352). doi: 10.1007/978 – 3 – 030 – 20497 – 6₃2
- Merat, Seppelt, B., Louw, T., Engström, J., Lee, J., Johansson, E., . . . Keinath, A. (2019, 02). The “out-of-the-loop” concept in automated driving: proposed definition, measures and implications. *Cognition, Technology and Work, 21*. doi: 10.1007/s10111-018-0525-8
- Norman, D. (1990, 05). The ‘problem’ with automation: Inappropriate feedback and interaction, not ‘over-automation’. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 327*, 585-93. doi: 10.1098/rstb.1990.0101

- Pauzié, A. (2008). A method to assess the driver mental workload: The driving activity load index (dali). *IET Intelligent Transport Systems*, 2(4), 315–322.
- Petersen, Luke, Robert, Lionel, Yang, Jessie, X., . . . Dawn (2019). Situational awareness, driver’s trust in automated driving systems and secondary task performance. *SAE International Journal of Connected and Automated Vehicles*, 2(12-02-02-0009).
- Radlmayr, Jonas, Gold, Christian, Lorenz, Lutz, . . . Klaus (2014). How traffic situations and non-driving related tasks affect the take-over quality in highly automated driving. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 58, pp. 2063–2067).
- SAE. (2021). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles*. SAE International in United States, J3016_202104.
- Sanneman, L., & Shah, J. A. (2020). A situation awareness-based framework for design and evaluation of explainable ai. In D. Calvaresi, A. Najjar, M. Winikoff, & K. Främling (Eds.), *Explainable, transparent autonomous agents and multi-agent systems* (pp. 94–110). Cham: Springer International Publishing.
- Sebok, & Wickens. (2017). Implementing lumberjacks and black swans into model-based tools to support human–automation interaction. *Human factors*, 59(2), 189–203.
- Seppelt, B. D., & Lee, J. D. (2019). Keeping the driver in the loop: Dynamic feedback to support appropriate use of imperfect vehicle control automation. *International Journal of Human-Computer Studies*, 125, 66-80.
- Shen, Y., Jiang, S., Chen, Y., Yang, E., Jin, X., Fan, Y., & Campbell, K. D. (2020). To explain or not to explain: A study on the necessity of explanations for autonomous vehicles. *ArXiv*, [abs/2006.11684](https://arxiv.org/abs/2006.11684).
- Tsimhoni, O., Green, P., & Lai, J. (2001). Listening to natural and synthesized speech while driving: Effects on user performance. *International Journal of Speech Technology*, 4(2), 155–169.
- van den Beukel, A. P., & van der Voort, M. C. (2017). How to assess driver’s interaction with partially automated driving systems – a framework for early

- concept assessment. *Applied Ergonomics*, 59, 302-312.
- Wang, Y. L., & Sus Lundgren, F. C., Lyckvi. (2020). How drivers respond to visual vs. auditory information in advisory traffic information systems. *Behaviour & Information Technology*, 39(12), 1308-1319.
- Weiner, B. (1979). A theory of motivation for some classroom experiences. *Journal of educational psychology*, 71(1), 3.
- Wickens, C. D. (2008a). Multiple resources and mental workload. *Human factors*, 50(3), 449-455.
- Wickens, C. D. (2008b). Multiple resources and mental workload. *Human Factors*, 50(3), 449-455.
- Wiegand, Gesa, Eiband, Malin, Haubelt, M., Hussmann, & Heinrich. (2020). “i’d like an explanation for that!” exploring reactions to unexpected autonomous driving. In *22nd international conference on human-computer interaction with mobile devices and services* (pp. 1-11).
- Wintersberger, Janotta, Peintner, Löcken, & Riener. (2021, jan). Evaluating feedback requirements for trust calibration in automated vehicles. *it - Information Technology*, 63(2), 111-122. doi: 10.1515/itit-2020-0024
- Wintersberger, Philipp, Frison, Anna-Katharina, Riener, A., Sawitzky, & von, T. (2019). Fostering user acceptance and trust in fully automated vehicles: Evaluating the potential of augmented reality. *PRESENCE: Virtual and Augmented Reality*, 27(1), 46-62.
- Zhou, F., Alsaid, A., Blommer, M., Curry, R., Swaminathan, R., Kochhar, D., ... Lei, B. (2020). Driver fatigue transition prediction in highly automated driving using physiological features. *Expert Systems with Applications*, 113204.
- Zhou, F., Alsaid, A., Blommer, M., Curry, R., Swaminathan, R., Kochhar, D., ... Tijerina, L. (2021). Predicting driver fatigue in monotonous automated driving with explanation using gpboost and shap. *International Journal of Human-Computer Interaction*, 1-11.
- Zhou, F., Yang, X. J., & de Winter, J. C. (2021). Using eye-tracking data to predict

situation awareness in real time during takeover transitions in conditionally automated driving. *IEEE Transactions on Intelligent Transportation Systems*.

Zhou, F., Yang, X. J., & Zhang, X. (2019). Takeover Transition in Autonomous Vehicles: A YouTube Study. *International Journal of Human-Computer Interaction*, 0(0), 1–12. doi: 10.1080/10447318.2019.1634317