KTH ROYAL INSTITUTE
OF TECHNOLOGY

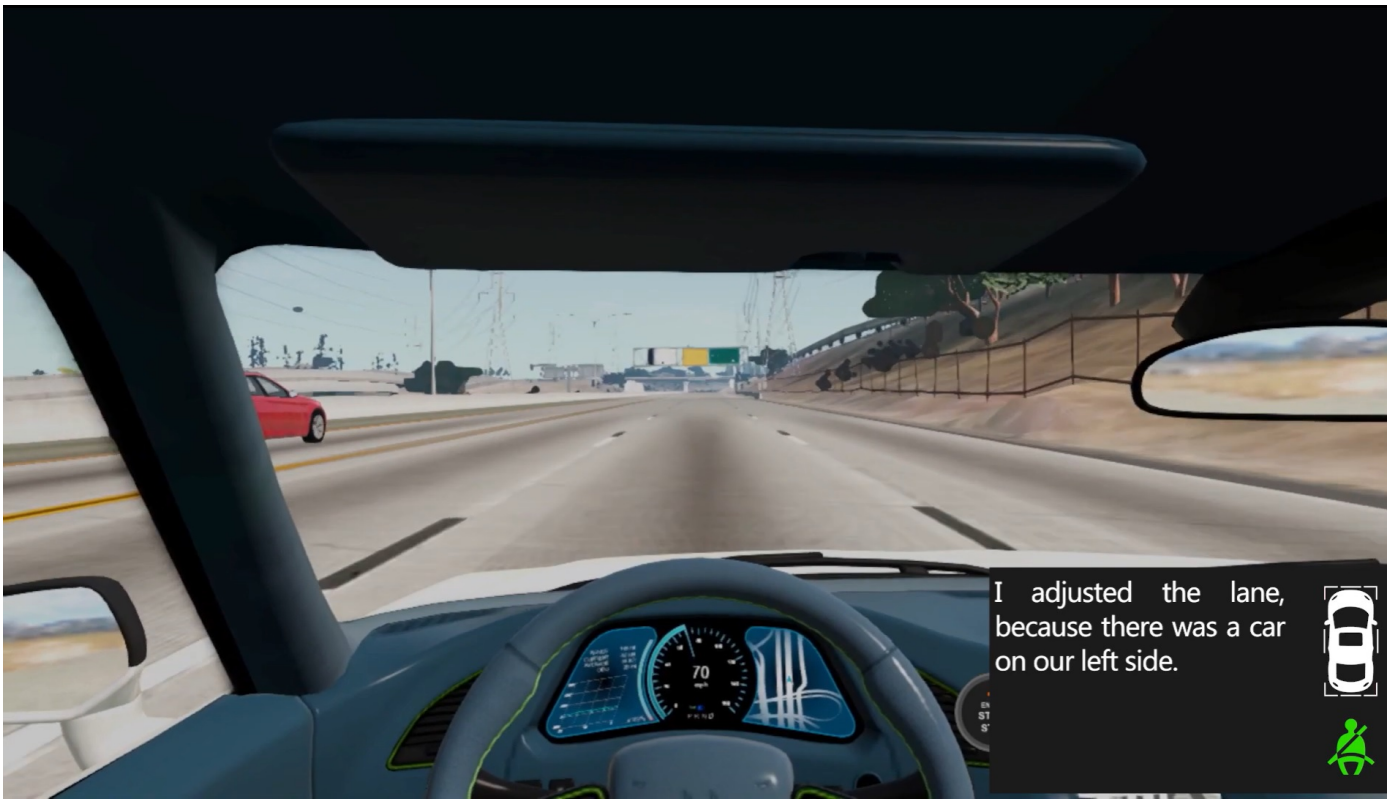
Degree Project in Computer Science and Engineering, Specialising in ICT Innovation

Second Cycle, 30 credits

Human Factors Involved in Explainability of Autonomous Driving

Master's Thesis

ABRIANSYAH ARISONI



Stockholm, Sweden 2023

KTH ROYAL INSTITUTE
OF TECHNOLOGY

Degree Project in Computer Science and Engineering, specialising in ICT Innovation

Second cycle, 30 credits

Human Factors Involved in Explainability of Autonomous Driving

Master's Thesis

ABRIANSYAH ARISONI

Stockholm, Sweden, 2023

Human Factors Involved in Explainability of Autonomous Driving

Master's Thesis

ABRIANSYAH ARISONI

Date: August 15, 2023

Supervisor: Elmira Yadollahi
Examiner: Iolanda Dos Santos Carvalho Leite
School of Electrical Engineering and Computer Science
Swedish title: Mänskliga faktorer som är involverade i förklaringen av autonom körning
Swedish subtitle: Magisteruppsats

© 2023 Abriansyah Arisoni

Abstract

Autonomous Car (AC) has been more common in recent years. Despite the rapid development of the driving part of the AC, researchers still need to improve the overall experience of the AC's passengers and boost their willingness to adopt the technology. When driving in an AC, passengers need to have a good situation awareness to feel more comfortable riding in an AC and have a higher trust towards the system. One of the options to improve the situation awareness is by giving passengers an explanation about the situation.

This study investigates how the situational risk of specific driving scenarios and the availability of visual environment information for passengers will affect the type of explanation needed by the AC passenger. The study was conducted through a series of different scenario tests presented to online study participants and focused on the human interaction to level 4 and 5 AC. This study's primary goal is to understand the human-AC interactions further, thus improving the human experience while riding in an AC.

The results show that visual information availability affects the type of explanation passengers need. When no visual information is available, passengers are more satisfied with the type that explain the cause of AC's action (causal explanation). When the visual information is available, passengers are more satisfied with the type that provide intentions behind the AC's certain actions (intentional explanation). Results also show that despite no significant differences in trust found between the groups, participants showed slightly higher trust in the AC that provided causal explanations in situations without visual information available.

This study contributes to a better understanding of the explanation type passengers of AC need in the various situational degree of risk and visual information availability. By leveraging this, we can create a better experience for passengers in the AC and eventually boost the adoption of the AC on the road.

Keywords

Autonomous Driving, Human-Robot Interaction, Explainable Artificial Intelligence Autonomous Driving, Human-Robot Interaction, Explainable Artificial Intelligence

Sammanfattning

Autonomous car (AC) har blivit allt vanligare under de senaste åren. Trots den snabba utvecklingen av själva kördelen hos AC behöver forskare fortfarande förbättra den övergripande upplevelsen för AC-passagerare och öka deras vilja att anta teknologin. När man kör i en AC behöver passagerare ha god situationsmedvetenhet för att känna sig bekväma och ha högre förtroende för systemet. Ett av alternativen för att förbättra situationsmedvetenheten är att ge passagerare en förklaring om situationen.

Denna studie undersöker hur den situationella risken för specifika körsituationer och tillgängligheten av visuell miljöinformation för passagerare påverkar vilken typ av förklaring som behövs av AC-passageraren. Studien genomfördes genom en serie olika scenariotester som presenterades för deltagare i en online-studie och fokuserade på mänsklig interaktion med nivå 4 och 5 AC. Denna studiens främsta mål är att förstå människa-AC-interaktionen bättre och därmed förbättra den mänskliga upplevelsen vid färd i en AC.

Resultaten visar att tillgängligheten av visuell information påverkar vilken typ av förklaring passagerarna behöver. När ingen visuell information finns tillgänglig är passagerarna mer nöjda med den typ som förklarar orsaken till AC:s agerande (orsaksförklaring). När den visuella informationen finns tillgänglig är passagerarna mer nöjda med den typ som ger intentioner bakom AC:s vissa handlingar (avsiktlig förklaring). Resultaten visar också att trots att inga signifikanta skillnader i tillit hittats mellan grupperna, visade deltagarna något högre förtroende för AC som gav orsaksförklaringar i situationer utan visuell information tillgänglig.

Denna studie bidrar till en bättre förståelse för vilken typ av förklaring passagerare i AC behöver vid olika situationella riskgrader och tillgänglighet av visuell information. Genom att dra nytta av detta kan vi skapa en bättre upplevelse för passagerare i AC och på sikt öka antagandet av AC på vägarna.

Nyckelord

Autonom Körning, Interaktion Mellan Människa och Robot, Förklarlig Artificiell Intelligens Autonom Körning, Interaktion Mellan Människa och Robot, Förklarlig Artificiell Intelligens

Acknowledgments

First of all, I want to say thank you to my supervisor, Elmira Yadollahi, PhD for being an incredible mentor throughout this thesis project. Thank you for always being there to offer me support, help, time, guidance and invaluable insights. I am truly fortunate to have her as my supervisor and I know this thesis project would not have been possible without her help.

The next thank you goes to Iolanda Leite, PhD, my Examiner, for her invaluable help, support, and guidance. Her expertise in the social robotics area and her constructive feedback have greatly improved my quality of work, and I am grateful for the opportunity to learn from her.

I also want to sincerely thank the Indonesia Endowment Fund for Education (LPDP) and the National Research and Innovation Agency of The Republic of Indonesia (BRIN) for their generous support in funding my master's studies. Their financial assistance has made it possible for me to pursue and complete this academic endeavour.

To fellow Social Robotics research group members, I cannot thank you enough for your support throughout my thesis project. I also want to thank Shruti Chandra, PhD, for her assistance with the Qualtrics platform.

My deepest gratitude goes to my wife, April, and my son, Nara, for their love and support. Your understanding and sacrifices have been the force that enables me to finish my study. There will be no more late-night video calls because we will get together soon.

I also want to thank my parents, Novandi Arisoni and Rosmiati Rasimeng, my in-laws, Untung Rahardja and Murtiningsih, and my siblings. Your unwavering support and care for my family during my study have been invaluable. I am thankful for your love and sacrifices.

Last but not least, I want to thank all my friends in Sweden and Indonesia for their encouragement during my master's study. Especially to the PPI Stockholm that brings a little piece of Indonesia into Stockholm.

I am grateful to each and every individual and organization mentioned above for having such a significant effect on my work and making this master's degree journey possible.

Stockholm, August 2023
Abriansyah Arisoni

Contents

List of Figures

List of Tables

List of acronyms and abbreviations

AC	Autonomous Car
AI	Artificial Intelligence
HRI	Human-Robot Interaction
XAI	Explainable AI

Chapter 1

Introduction

Autonomous Car (AC) with powerful Artificial Intelligence (AI) systems are becoming increasingly common on our roads. These vehicles can make complex decisions without human input, using advanced algorithms and data to navigate and respond to various situations. According to SAE International, AC is divided into six levels of driving automation, level 0 - level 5. Level 0 is where the cars do not have any driving automation, and the driver has complete and sole control of the primary car control. Level 5 is where the car has full driving automation and can operate anywhere, and under all road conditions in which a conventional vehicle can be reasonably operated by a typically skilled human driver [1].

In critical systems where safety is prioritized, such as AC, having a confound situation awareness and understanding of the current situation is important in order to make an informed decision [2]. When riding in AC, the explanation provided can make humans believe that the AC can recognize the environment and base their actions on this recognition, thus deeming the AC trustworthy [3]. Explanations given by an AC need to be comprehensive but should not increase the cognitive load of the user, while at the same time still satisfying users' need of understanding AC actions [4].

In human-robot interaction, one factor that can improve coordination between both parties is having an accurate view of the task and each party's involvement. [5]. It means that the user can see the work performed by the robot. Even though the accurate view of the task is important (or in the case of AC driving is the environment visual information), a recent study shows that the need for the user to process visual attention can be considered as an additional cognitive workload [6]. Apart from the visual information, it is also important for humans to get an explanation if the AC shows an unexpected

behavior [4]. This risk of unexpected behavior can impact the user's trust towards the system [7].

Based on studies mentioned above, the following research questions was derived for this study:

RQ1: How the situational risk and visual information available to the passenger will influence the type of explanation needed by AC's passenger?

RQ2: How the visual information available to the passenger and the type of explanation will influence the passenger's trust towards AC?

1.1 Ethics and Sustainability

All of the responses collected from participants were just used for this study and should not have raised ethical questions. The questionnaire presented to participants contained a pre-defined response, meaning the response was managed in terms of appropriateness. Before being executed, all the plans and decisions regarding this study were presented to the project supervisor. This study was not designed to handle the sustainability issues that might have emerged in the future.

Chapter 2

Background

2.1 Explainable AI (XAI)

While AC has the potential to improve safety and efficiency on our roads, it also raises important questions about explainability and transparency. As AC make important decisions and perform autonomous tasks, it is essential that we are able to understand their actions through explanation [8]. That is why AI researchers and professionals have focused on Explainable AI (XAI) to help them better trust and understand the AI model [9]. With industry 4.0 growing rapidly with AI and other technology continous expansion as a leading industry innovator, AI and XAI are now important in the industry realm [10]. Stakeholders also show their concern in the application of AI. In their General Data Protection Regulation (GDPR), European Union introduced a 'right to explanation' provision in 2016, thereby intensifying the importance of XAI [11][12].

The XAI has been around as long as early AI [13]. XAI created to produce details or reasons to make AI function clear or easy for human to understand [14]. Explainability in XAI is important because it can help us understand, increase trust, and confidence in managing powerful AI applications [15][16][17].

2.2 XAI in Autonomous Driving

XAI is essential for safety-critical systems such as defense, health care, law and order, and autonomous driving vehicles [18]. First, we need to understand that humans expect explanation when they are confused and surprised by the behavior of others [4]. Driving in an AC will potentially expose humans

to various uncertain situations or behavior. Thus, in terms of AC, the need for it stems from the increasing concerns for transparency and accountability of autonomous vehicles [19]. Recent studies shows that many potential AC customers still feel reluctant and hesitant towards the adoption of AC [20] [21]. XAI is also particularly important, as it can help promote trust and acceptance of this technology and provide valuable insights into the performance and reliability of these systems [22]. The need for explanations also arises because users need accessible information on which factors have been considered when the system takes action [23]. XAI produces details or reasons to make its functioning clear or easy to understand [14].

2.3 Explanation Type

There are different types of explanations in order to put context into an AI system. In 2015, Koo *et al.* [24] investigated three different types of explanations in AC in relations to driver understanding, trust, and performance on a semi-autonomous vehicle. The first explanation was *How* messages, that informed the user about how the car is acting. The example for this *How* message is ”The car is braking”. Second type of explanation that they tested was the *Why* message, that presented situational information and explained the reasons for engaging automation, such as ”Obstacle ahead.”. The last type of explanation they tested was the combination of *How* and *Why* messages that alert how the car was acting and why the car was making those actions. An example of this combination message is ”Car is braking due to obstacle ahead”. In this study, they found that ”why” messages led to poor driving performance from the users. ”How” messages were preferred by drivers and led to better driving performance. Combination of ”why” and ”how” messages resulted in the safest driving performance but created negative feelings in drivers. They argued that providing detailed explanations might increase the cognitive overload of drivers.

Ha *et al.*. [3] investigated three types of explanations in relations to trust towards AC. They use no explanation, simple explanation, and attributional explanation that are based on attribution theory. For simple explanation, they gave a description of why and how certain actions were executed without a subject. An example of a simple explanation is ”Stopped after identifying the sudden appearance of a pedestrian in the road”. In the attributional explanation, they provide the description of why and how the AC acted. An example of this attributional explanation is ”The autonomous vehicle stopped after identifying the sudden appearance of a pedestrian in the road”. After

doing the experiment, they unveiled that in the high level of perceived risk scenario, attributional explanations were not effective for increasing trust towards AC, compared to a simple or no explanations. They argued that attributional explanations might enhance cognitive overload when users are faced with a high degree of risk situation.

In the same year similar to Ha *et al.*, Schraagen *et al.* [25] conducted a study that investigated the effect of different types of explanations in relation to explanation satisfaction and trust. In this study, they differentiate the explanations into three types; causal (causes), intentional (reasons), and mixed (combination of causal and intentional) explanation. For the causal explanation, they gave explanations such as ”I slow down because of the person on the left side”. For the intentional explanation, the example of the message will be ”I slow down because I believe the woman wants to cross the street”. And the mixed explanation example will be ”I slow down because there is a person on the left side and I believe she wants to cross the street”. After running the online study, they found out that participants were least satisfied with causal explanations. They also found that intentional explanations were effective in creating high levels of trust towards the system, and mixed explanations makes the user had higher understanding of the system and resulted in the least changes in trust over time.

2.4 Situational Risk

Situational or external risk is the uncertainty that is related with the driving situation [26] [27]. In a study conducted in 2020, Stuck *et al.* [7] stated that perceived situational risk is someone’s belief of the probability and/or feeling that a specific situation has potential negative outcomes based on someone’s knowledge and experience with the task. In the study, they also found that the presence of risk and participants’ perceived situational risk can impact behavioral trust towards automation.

In the same study mentioned in the previous section, Ha *et al.* [3] examined certain explanation types against different levels of risk that were implemented in the simulator program. They created four autonomous driving situations with different levels of risk in terms of the weather (clear day and snowy night) and driving speed (faster than 40 km/h and slower than 40 km/h). Users in experiment reported that participants perceiving higher risk in unnatural situations such as driving with a slow speed on a clear day, and they perceived lower risk in natural situations such as driving with a slow speed on a snowy night. Ha *et al.* [3] suggest that effective feedback and explanation needs to be

designed based on the perceived risk rather than the actual risk of situations.

Li *et al.* [28] conducted a study that examined if participants perceived risk differently in various scenarios. In this study, they designed nine driving scenarios that identified risk based on driving speed (high, medium, and low speed), traffic (trucks, cyclists, and heavy traffic), and abnormal behaviors (other swerves, other merges, and subject swerves). They found out that participants were able to distinguish and identify different levels of risk associated with the driving scenarios within certain categories. Results of this study showed that participants reported the highest level of trust, perceived automation reliability, and the lower level of perceived relational risk when driving in a low-risk situation.

2.5 Visual Information Availability

Visual information in Human-Robot Interaction (HRI) is deemed important because humans may alter their opinions at any time based on their own mental processes, thoughts and motivations, as well as what they see around them [29]. Furthermore, multiple studies also stated the importance of visual information in HRI. Tabrez *et al.* [30] stated that the key aspect of effective teamwork is maintaining awareness of what teammates are likely to do or need. Barnes *et al.* [31] stated that in soldier-robot teaming, the soldier that has shared imagery from the robot showed considerable performance gains.

In 2013, Gergle *et al.* [5] conducted a study to examine two coordination processes that are impacted by visual information, situation awareness and conversational grounding. In one of the experiments, they manipulate the timing of visual information availability given to participants. They found out that immediate visual feedback can improve collaboration and enhance situation awareness in participants. Gergle *et al.* [5] also stated that according to situation awareness theory, visual information increases coordination by providing actors with an accurate perspective of the task state and each other's activities. They also found that participants become more active in coordinating communication if the shared visual information is unavailable.

Chapter 3

Methodology

3.1 Experimental Design

Based on the literature review above, we designed an experiment to investigate the relations between explanation type, visual information availability, and situational risk to passengers' satisfaction with the explanation provided by the AC.

Therefore, we conducted an online study where participants watched AC simulated scenarios videos with various degrees of risk and measured their satisfaction with the AC's explanation. In this study, each participant watched a total of 6 videos that showed different degrees of risk, visual information availability and the type of explanation provided to the users. Figure 3.1 shows the overview experimental design in this study. All participants experienced different degrees of situational risk during the study. Participants in the first group were presented with videos that contained available visual information and causal type of explanation. Participants in the second group were presented with videos that contained available visual information and intentional type of explanation. Participants in the third group were presented with videos that contained no available visual information and causal type of explanation. Furthermore, participants in the fourth group were presented with videos that contained no available visual information and intentional type of explanation. In each group, we do counterbalance to minimize the order effect that might happen in this study. A more detailed view of the counterbalance we did in this study can be seen in Fig. A.1.

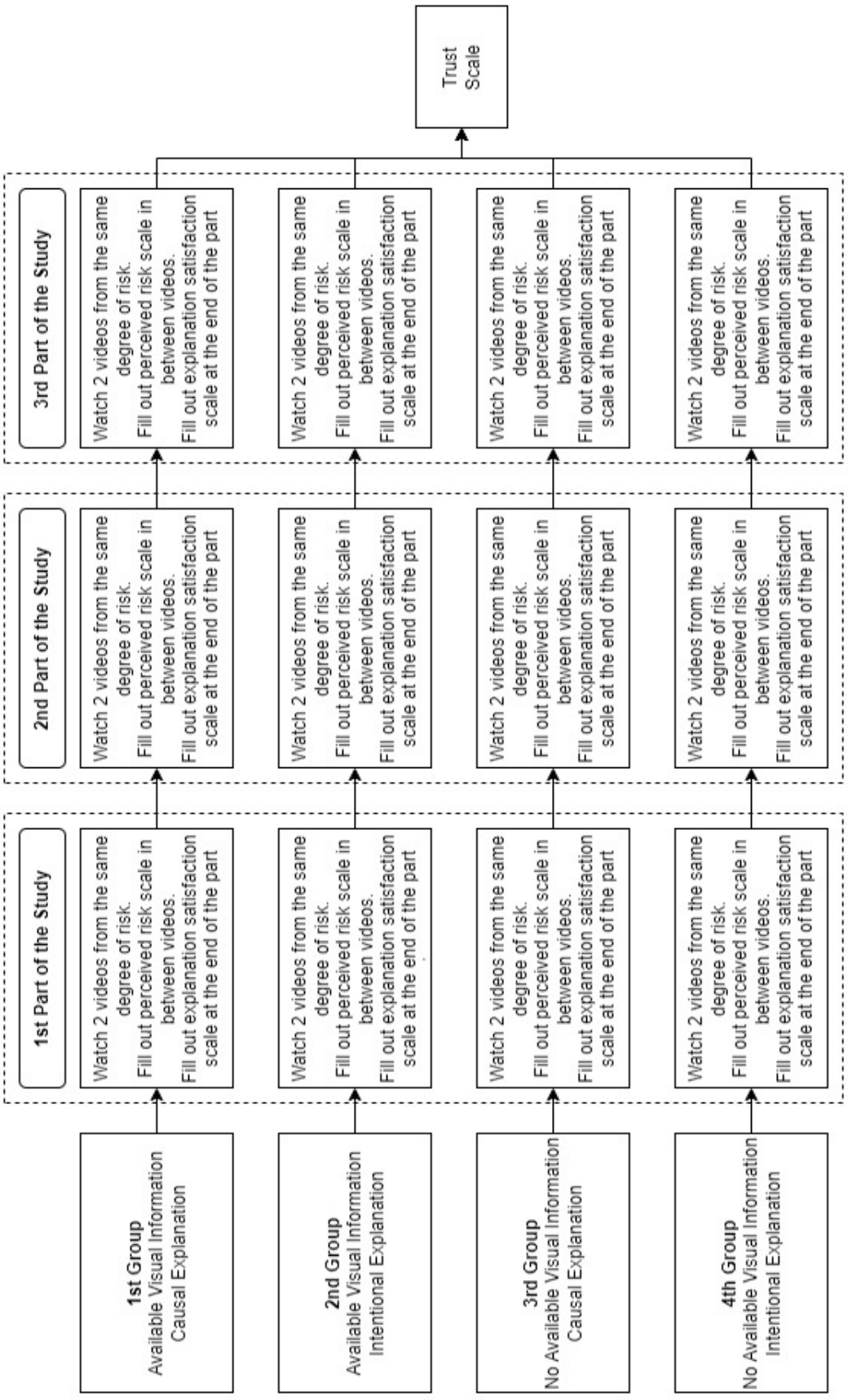


Figure 3.1: Experiment Design

3.1.1 Experiment Procedure

The online study starts with participants signing the declaration of consent that states they are at least 18 years of age, voluntarily agree to participate, are aware that they can terminate participation in this study at any time for any reason, they understand that their response is recorded for research purpose, and they are aware that their data will be made available to other researchers in an anonymized dataset. After signing the declaration of consent, participants would fill out the general demographics questionnaire. This demographic questionnaire recorded participants' age, gender, possession of a driving license, driving experience, and if they have any prior experience with an AC. Participants also fill out a questionnaire regarding risk perception [32]. This questionnaire was created to measure if they are a risk-taker person or not.

Following the initial questionnaire, participants were presented with three study parts. In each part, participants watched two videos created with a certain degree of risk, explanation type, and visual information availability. After the participants watched each video, they answered two questions that asked about what happened in the video. This was done to check if the participants had a good understanding of the situation. Questions asked to participants were "Which of the following best describes the situation on the video you just watched?" and "Apart from the one you rode in, how many cars were visible in the video?". After answering questions about the situation shown on the video, participants were asked to fill out the perceived risk questionnaire [33] to check if the participants perceived the situation according to our degree of risk. In addition to that, after each part, participants were asked to fill out a questionnaire about satisfaction [34] towards the explanation provided by the AC. After completing three parts of our study, participants will be asked to fill out the questionnaire about their trust [34] towards the AC after watching the presented videos.

3.2 Degree of Risk

The degree of risk presented to the participants is divided into three degrees: high, medium, and low. High-risk scenarios will present accidents that the ACs might have in a real-life setting. One thing to remember is that in high-risk scenarios, the AC is not at fault. The accidents happen because other cars on the road, driven by humans. In the medium-risk scenarios, participants were presented with a series of near-miss accidents that might happen in real life. As in the high-risk scenarios, the AC are not at fault, and the events happen

because of the human error from another driver on the road. For the low-risk scenarios, participants were presented with day-to-day driving scenarios. The degree of risk variable was presented within the subject. Table 3.1 shows the different scenarios for different degrees of risk that were presented to participants. This degree of risk is an independent variable presented within-subject to the participants.

Table 3.1: Driving Scenarios

No.	**Degree of Risk**	**Visual Environment Available to Passenger**	**Visual Environment Not Available to Passenger**
1	High (1st Video)	Side collision caused by other drivers suddenly changing a lane	Multiple-Vehicle Collision (gets rear-ended by other car)
2	High (2nd Video)	Truck suddenly merge into our lane and hit us	Changing lane and get rear-ended by a car
3	Medium (1st Video)	While in the highway, AC suddenly move and tried to avoid a car that suddenly merging to its lane	AC suddenly accelerate because the car in the blindspot suddenly move and try to change into our lane
4	Medium (2nd Video)	AC suddenly brake to avoid another car that merges into our lane	AC suddenly brakes to avoid another car that merges into our lane from the right side. AC cannot change lanes because there is also a car approaching on its left side.
5	Low (1st Video)	Move to slow lane to exit the highway	Merging into the highway
6	Low (2nd Video)	Overtake a car on the highway	Overtake a car on the highway but waiting for another car to pass

3.3 Visual Information Availability

In this study, participants were split into two categories of visual information availability reflected in the simulated videos. In the first category, participants have a clear view of what happens in the videos. This visual information was also supplemented by the explanation that was given by the AC. In the second category, participants do not have a clear view of what is happening in the videos. They have to rely on the explanation that was provided by the AC to comprehend the situation. This independent variable of visual information availability was presented between subject in this study. Table 3.1 shows different scenarios where participants have visual information or not. From the table, we can see a clear difference in visual information availability of the cause of AC's actions. For instance, in the initial video depicting a high-risk situation, passengers clearly see a side collision caused by other drivers that suddenly change lanes from the side of the car. Conversely, in the video with no visual environment for passengers, the AC experiences a rear-end collision, making it impossible for participants to see the approaching car from behind.

3.4 Explanation Type

The third independent variable manipulated was the type of explanation given to passengers. In this study, participants were provided with two different types of explanations that we adapted from Schraagen *et al.*'s study [25]. The first type of explanation was the causal explanation. The causal explanation explains the cause of AC's actions in the video. The second type of explanation was the intentional explanation. This type of explanation explains the intention or reason behind the AC's actions in the video. Table 3.2 shows the different explanation messages in the first video, while Table 3.3 shows the different explanation messages in the second video. In this study, the type of explanation variable was presented between the subject.

3.5 Simulation Videos

All scenarios were simulated through Beam.NG software [35] and went through Adobe Premiere Pro for post-editing to add text-based and voice-based explanations. In the Beam.NG, we used the eSBR 800 car, a stock electric car provided by the software. We created the scenarios by manipulating the car using the steering wheel that was connected to a laptop. Figure 3.2 shows the

Table 3.2: Type of Explanation on First Video

No.	Degree of Risk and Visual Information Availability	Causal Explanation	Intentional Explanation
1	High risk and visual information available	I am adjusting the lane, because there is a car on our right side	I am adjusting the lane, because there is a car that suddenly merging into our lane.
2	Medium risk and visual information available	I adjusted the lane, because there was a car on our left side	I adjusted the lane, because there was a car that suddenly merged into our lane
3	Low risk and visual information available	I am changing lane because there is an exit ahead	I am changing lane to follow our route
4	High risk and visual information not available	I brake because there is an accident ahead	I brake to avoid a collision ahead
5	Medium risk and visual information not available)	I accelerate because there is a car on our left back side	I accelerate to avoid side collision
6	Low risk and visual information not available	I am slowing down because there is still cars on our left side	I am slowing down to wait for the opportunity to merge into highway

example of one of the videos that were presented to participants during the study.

3.6 Measures

In this study, we used multiple validated scales from previous studies. To measure the explanation satisfaction, we used the explanation satisfaction scale that consists of 8 items. This scale measures how passengers feel they understand the system or process being explained to them[34]. For participants' trust towards the AC, we use the trust scale specifically created

Table 3.3: Type of Explanation on Second Video

No.	Degree of Risk and Visual Information Availability	Causal Explanation	Intentional Explanation
1	High risk and visual information available	I change lane because there is a truck on our left side	I change lane because there is a truck that suddenly merge into our lane
2	Medium risk and visual information available	I brake because there was a car on our right side	I brake because there was a car that suddenly merge into our lane
3	Low risk and visual information available	I am changing lane to overtake a car in front of us	I am changing lane because it is safe to increase speed
4	High risk and visual information not available	I am changing lane because there was a car on our left back side	I am changing lane because there was a car that suddenly merged into our lane.
5	Medium risk and visual information not available)	I brake because there is a car on our right side that merged into our lane, and there is a fast car approaching on our left side	I brake to avoid side collision by staying in our lane
6	Low risk and visual information not available	Waiting for the car on the left to pass, I am changing lane to overtake a car in front of us	Waiting to overtake, I am changing lane because it is safe to increase the speed

for an XAI [34]. This trust scale consists of 7 items. In addition to the explanation satisfaction and trust scale, we also adapted 5 items from the perceived situational risk scale [33] to validate if the passengers perceived different degrees of situational risk in each part of the study. The last scale that we used was the willingness to share the road. This scale was adapted from a previous study about the willingness of the driver to share the road

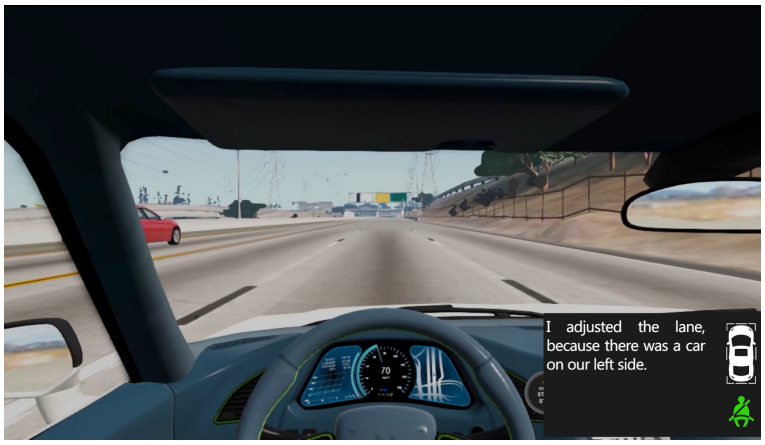


Figure 3.2: Simulation Video

with cyclists [36]. All of the scales we use are Likert scale that uses 5 points to measure the explanation satisfaction and trust towards the system.

3.7 Hypotheses

To understand the type of explainability needed in **RQ1**, we can look at the previous study, which found that AC passengers were least satisfied with causes (causal) explanation and reasons (intentional) explanation were most effective in establishing high levels of trust towards the system [25]. Nonetheless, we must acknowledge that ACs shifted humans' core tasks away from driving and toward secondary tasks like entertainment and receiving calls [37]. This shift changed how humans need to stay alert without causing cognitive overload and affecting users' perceptions [3][11]. Previous study also found that providing a more detailed explanation can potentially increase the driver's cognitive workload [24]. It is interesting to observe if there is a change of explanation type needed when passengers were faced with a higher degree of situational risk that take more cognitive load of the passengers. To further understand the type of explainability needed by users, we can measure users' feeling of satisfaction on the explanation itself AC[34].

For **RQ2**, we need to know that people have a need to keep track of other people perceive and know in social interaction [38]. People also tend to took a robot's perspective, especially when the robot displays nonverbal behaviors [39]. A Previous study shows that when people cannot see the cause of the robot's actions in human-robot interaction, they rate the robots as more unpredictable and less competent than when the cause is visible [40]. During

human-robot interaction, people will perceive a robot as more trustworthy if the robot's action is more predictable [41], thus having a more accurate understanding of the decision-making process, which are more desirable traits in a human-robot teaming [30]. By keeping that in mind, hypotheses are crafted below:

H1.a : In situations where passengers have limited visual information access, they generally exhibit a greater preference for receiving causal explanations, resulting in increased satisfaction.

H1.b : In situations where passengers have access to visual information, they typically show a preference for intentional explanations, which contributes to increased satisfaction levels.

H2 : The combination of visual information and intentional explanation will enhance passengers' trust in autonomous vehicles

3.8 Participants

A total of 364 participants between 18 to 65 years of age, with a mean of 39.17 (SD = 12.49), were recruited through the Prolific platform and split into 4 different groups, as seen in Figure 3.1. Since the videos show the AC driving on the right side of the road and the speedometer in the AC shows miles per hour to indicate speed, we only recruited participants based in the United States of America. Considering that our study took around 15 minutes to complete, we paid £3.00 to every participant that finished the study through Prolific.

A preliminary power analysis shows a minimum sample size of 64 participants for each category is needed to detect a medium-sized effect (d = 0.5) [42] with effect size $\alpha = 0.05$ and power = 0.8. A total of 361 participants were recruited through Prolific and deemed sufficient according to the power analysis.

Chapter 4

Results

4.1 Data Validation

After running the experiment, we did multiple analyses to see if our data was valid. First, we did the descriptive statistics of each group. Second, we calculate the alpha to measure the internal consistency of the scales [43] that we used in this experiment. Third, we checked if there was an order effect in this experiment. We divided each group into six categories that differed in the order of degree of risk in the videos they watched. Lastly, we analyze if participants perceived different degrees of risk from each video. All the analyses were done using SPSS software [44].

4.1.1 Internal Consistency

Each group's perceived risk and explanation satisfaction scales show excellent consistency ($\alpha \geq 0.9$). For the trust scale, the analysis shows that the scale has a good internal consistency ($0.9 \geq \alpha \geq 0.8$) in group 1, acceptable internal consistency ($0.8 \geq \alpha \geq 0.7$) in group 2 and 4, and questionable internal consistency ($0.7 \geq \alpha \geq 0.6$) in group 3. Considering that the trust scale in other groups shows that the scale has good or acceptable consistencies, we will still analyze further the trust towards an AC in group 3. The willingness to share the road scale across the group shows that the scale's internal consistency was unacceptable ($0.5 > \alpha$). Thus, we will not do further analysis based on the willingness to share the road scale. Detailed α value for every scale used in each group can be seen in Appendix C.

4.1.2 Order Effect

To check if this experiment presented an order effect, we conducted the Kruskal-Wallis test [45]. McKight *et al.* [46] stated that the Kruskal-Wallis test can be used to assess if there is a difference between three or more independently sampled groups with non-normally distributed data. In this experiment, we did not find a significant difference in explanation satisfaction across participants categories in each group. Based on that, we can conclude that concatenate each category and analyze them together in each group. The summary of the Kruskal-Wallis test can be seen in Appendix D.

4.1.3 Perceived Risk

In this study, we want participants to perceive different degrees of risk in each part of the online study. We calculated the means of a perceived risk scale that participants filled out after watching each video to validate that. After comparing the means, we found that participants in all groups perceived different degrees of risk for each part of the online study. Figure 4.1 shows the perceived degrees of risk for each video participants in group 1 watched. The perceived risk summary for group 2, 3, and 4 can be seen in Appendix E

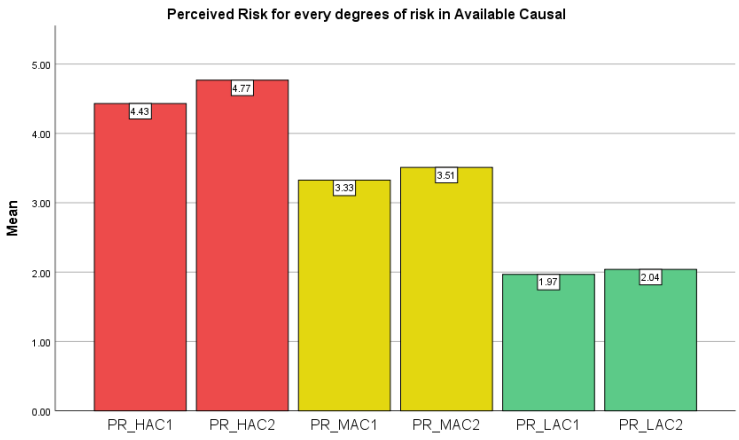


Figure 4.1: Perceived Risk in Group 1

4.2 Explanation Satisfaction

After conducting the validity check for our experiment data, we analyze the explanation satisfaction in each group. We analyzed the satisfaction using

Mann-Whitney U Test to investigate if there was a significant difference between two groups with non-normally distributed data [47]. In all tests we conducted, a significance level of 0.05 and 95% confidence interval were used in the test options.

We split the analysis into two parts. In the first part, we examine the explanation satisfaction by comparing groups with the same information availability. In the second part, we examine the explanation satisfaction by comparing groups with the same explanation type. Table 4.1 shows the separation between the two parts of explanation satisfaction analysis.

Table 4.1: Summary of Explanation Satisfaction Analysis

Variable	**Group(*N*)**	**Risk**	**SD**	**Mean Rank**	**Z**	***p*-value**
Information Available	Causal Expl.(91)	High	1.1667	87.73	-.851	.395
	Intentional Expl.(90)			94.31		
	Causal Expl.(91)	Medium	.7803	83.94	-1.861	.063
	Intentional Expl(90).			98.14		
	Causal Expl.(91)	Low	.7959	90.35	-.173	.863
	Intentional Expl.(90)			91.66		
Information Not Available	Causal Expl.(92)	High	.8162	95.71	-.961	.337
	Intentional Expl.(91)			88.25		
	Causal Expl.(92)	Medium	.7731	100.43	-2.192	.028
	Intentional Expl.(91)			83.47		
	Causal Expl.(92)	Low	.6807	100.90	-2.320	.020
	Intentional Expl.(91)			83.00		
Causal Explanation	Info Avail.(91)	High	1.0223	81.92	-2.584	.010
	Info Not Avail.(92)			101.97		
	Info Avail.(91)	Medium	.7677	88.87	-.805	.421
	Info Not Avail.(92)			95.09		
	Info Avail.(91)	Low	.7125	90.12	-.492	.623
	Info Not Avail.(92)			93.86		
Intentional Explanation	Info Avail.(90)	High	1.0186	88.59	-.618	.537
	Info Not Avail.(91)			93.38		
	Info Avail.(90)	Medium	.7890	103.41	-3.221	.001
	Info Not Avail.(91)			78.73		
	Info Avail.(90)	Low	.7641	98.89	-2.041	.041
	Info Not Avail.(91)			83.20		

In the first test, we compared groups with the same information availability. We found there was a significant difference (p<0.05) in the situation with medium and low risk, where there is no visual information available for

passengers. We found that participants were significantly more satisfied with the causal explanation than the intentional explanation when no visual information was available for the passengers to see what happened. Although there is no significant difference in the high-risk situation where there was no visual information available for passengers to see, participants were relatively more satisfied with the causal explanation compared to the intentional explanation. Figure 4.2 shows each group's mean plot. We also found that although there was no significant difference between causal and intentional explanations in the situation where there was visual information available for passengers to see, we can see that participants were relatively more satisfied with the intentional explanation compared to the causal explanation.

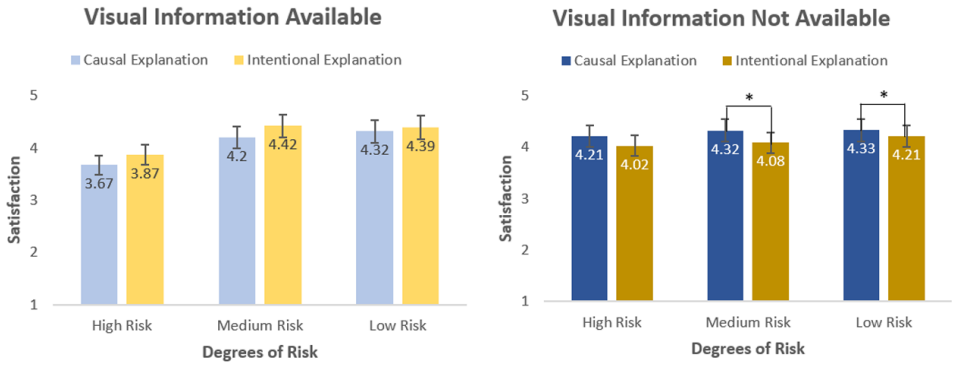


Figure 4.2: Explanation Satisfaction from Information Availability Perspective

In the second test, we compared groups with the same explanation type. After conducting the test, we found that there was a significant difference (p<0.05) in satisfaction with the high-risk situation when passengers were presented with the causal explanation. Participants were significantly more satisfied with the causal explanation when there was no visual information compared to when there was visual information for them to see. In addition to that, even though we did not find significant differences in the medium and low-risk situations, participants were relatively more satisfied with the causal explanation when there was no visual information compared to when there was visual information for them to see.

We also found that there was a significant difference (p<0.05) between groups in the medium and low-risk scenario, where participants were more satisfied with the intentional explanation where there was visual information compared to when there was no visual information for participants to see.

However, in the high-risk situation, even though there was no significant difference, we found that participants were relatively more satisfied with the intentional explanation when there was no visual information available compared to when visual information was available for them to see. Figure 4.3 shows the comparison of mean between groups that had the same explanation types between them.

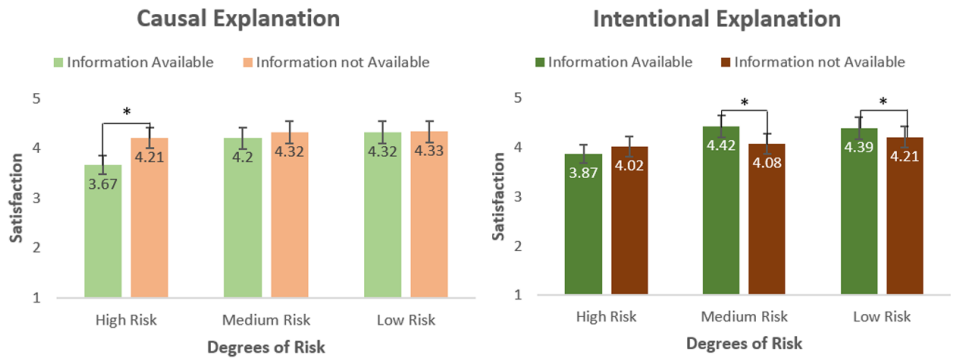


Figure 4.3: Explanation Satisfaction from Explanation Types Perspective

4.3 Trust towards AC

To analyze the trust towards AC, we test the mean of trust across the groups using the Kruskal-Wallis test [45]. We use the Kruskal-Wallis test to see if there is a significant difference in trust towards AC between the groups that experience different situations and explanations. In this test, we use a significance level of 0.05 and a confidence interval of 95%. Figure F.1 shows the test summary where we found no significant difference (p>0.05) in trust across the groups.

The test also gives us the mean-rank from each group. Group 1 has a mean rank of 174.66. Group 2 has a mean rank of 170.56. Group 3 has a mean rank of 195.12. Moreover, Group 4 has a mean rank of 189.39. Figure F.2 shows the mean-rank chart of the trust from the Kruskal-Wallis test.

Next, we can plot the mean of trust from each group to analyze the data further. Figure 4.4 shows that participants generally have a higher trust towards the AC when they are in a situation where they do not have visual information available, compared to a situation where they do have visual information available for them to see. Although it is not significant, we can also see that participants relatively had a higher trust towards the AC when

they presented with the causal explanation, compared to when they presented with the intentional explanation.

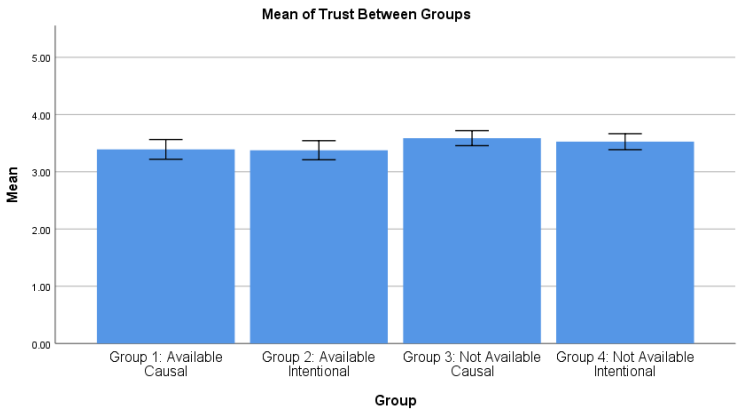


Figure 4.4: Trust Compared Between Groups

Chapter 5

Discussion

In this study, we investigate how the situational degree of risk of certain driving scenarios and the availability of visual environment information for passengers will affect the type of explanation needed by the AC's passenger. To answer that, we designed an online study where participants watched simulated videos of AC experiences. In the online study, participants will tell us how they feel about the experience through a series of validated scales.

Corresponding to the aim of this study, we develop two research questions. The **RQ1** investigates how the different degrees of situational risk, combined with the availability of visual information for the users, will influence the type of explanation needed by AC's passenger. The **RQ2** investigates how the visual information availability and the type of explanation will influence the passengers' trust towards the AC.

There are three independent variables that we manipulate in this study. The first one is the degree of situational risk. We created three different degrees of risk that act as a within-subject variable. The rest of the variables are the visual information availability and the explanation type, which act as between-subject variables.

Two dependent variables are used to answer two research questions that we crafted. For the **RQ1**, we examine the explanation type needed by measuring explanation satisfaction after participants finish one part of our study. For the **RQ2**, we measure the trust towards AC at the end of a study.

In the following paragraph, we will highlight our findings and how it relates to other studies. We will also discuss how this study can contribute to creating a better experience for AC's passengers in the future.

From the explanation satisfaction data, we can look at them from two different perspectives. The first one is to compare the groups with the

same information availability. After doing the Mann-Whitney U test based on this first perspective, we found no significant differences between the satisfaction generated by causal and intentional explanation in the situations when there was visual information available for passengers to see. However, by looking at the mean between the two groups, we found that participants were relatively more satisfied with the intentional explanation, compared to the causal explanation, in the situation when they had visual information available for them to see. This result supports the **H1.b** that we crafted. Compared to the previous paragraph, participants with visual information might have a lower cognitive workload in processing visual attention and might be able to process more detailed explanations about the reasons behind the AC's action. This second finding can be connected to the previous study by Schraagen *et al.* [25] that stated intentional explanation could create higher satisfaction towards the AC.

In addition, we found a significant difference between the causal and intentional explanation in the medium and low-risk situations when there was no visual information available for passengers to see. Passengers were significantly more satisfied with the causal explanation than the intentional explanation when they had no access to the visual information. We also found that although there was no significant difference, passengers were also relatively more satisfied with the causal explanation compared to the intentional explanation in the high-risk situation when they had no access to the visual information. This result supports the **H1.a**.

Based on the results, we can safely assume that the absence of visual information makes passengers wonder what is happening in the situation. That is why we think that participants preferred to know the cause of the AC's action in that situation compared to the reason behind the AC's action. From previous study, we also know that the need for the user to process visual attention can be considered as an additional cognitive workload [6]. The cognitive workload might especially increase where no visual information is available, whereas the users want to know what is happening in a certain situation. This finding also aligns with Koo *et al.* [24] found in their study. They found that explaining how the car is acting (the cause) is preferred compared to providing the reasons why the car is acting (intention).

The second test we did on the explanation satisfaction data was to compare the group with the same explanation type. In this test, we found that participants were significantly more satisfied with the causal explanation in the high-risk situation when visual information was not available compared to when there was visual information available for them to see. We also found that

in the medium and low-risk, although there were no significant differences, we found that participants were relatively more satisfied with causal explanations in situations where they did not have visual information. This result supports the **H1.a** and complements the first test result. We predict that the unavailable visual information generates additional cognitive workload, thus needing a simpler explanation about how the car is acting (cause) in that situation [24]. Once the participants have visual information, the cognitive workload will be reduced, and participants might be able to process more detailed explanations.

We also found that in the situation with low and medium risk, participants were significantly more satisfied with the intentional explanation if visual information was available compared to the intentional explanation in the situation without visual information. This result also supports the **H1.b** and complements the previous test on the explanation satisfaction data. However, from the test, we found that although it is not significant, participants were slightly more satisfied with the intentional explanation when they faced a situation without visual information in the high degree of risk. This result can be explained by looking at the previous study by Gergle *et al.* [5] that stated people become more active in coordinating communication if the shared visual information is unavailable, thus creating the need for a detailed explanation in the higher risk.

Considering all aspects above, this study affirms previous research in cases where visual information was available. However, when visual information was not available, the results of this study showed a completely opposite preference for the type of explanation provided by AC. This outcome establishes the importance of visual information in how humans perceive explanations, and it should be factored into the design of explanations itself. This includes exploring methods to assess how humans perceive visual information, comparing it to how ACs detect the environment, thus giving the users a suitable explanation in certain situations.

From the trust towards AC data, we did not find a significant difference between the four groups that we observed. However, based on the mean that can be seen in Figure 4.4, we can see that participants had a slightly higher trust towards the system that gives the causal explanation in the situation where there was no visual information available (Group 3). This results rejecting the **H2** that stated the combination of visual information and intentional explanations (Group 2) will generate the highest trust between groups we observe. In fact, Group 2 generates the lowest trust by a slight margin. This could happen because of two possibilities. First, it is possible that the scale that we used is not reliable. It can be seen from the internal consistency validation

that the trust scale in Group 3 has a questionable internal consistency ($0.7 \geq \alpha \geq 0.6$). The second possible reason is that participants might base their trust rating not on the explanation but on how the car acts in that scenario. In addition to that, even if we use the validated scales developed in the previous studies, there is a lack of confirmatory testing in trust scales that exists right now. [48].

5.1 Limitations

The biggest limitation of this study is the fact that this study will be conducted online. This will limit the type of response that can be observed from the participants. For instance, we cannot observe the speed of the response from the users if the AC needs human actions. Another limitation will be the autonomous algorithm itself. This study will not use the in-house autonomous algorithm. Instead, the scenarios were built around the built-in AI system from BeamNG.drive software, limiting the flexibility of creating a scenario. This study also not exploring the optimal design of the user interface of AC that might influence users' perception and trust towards an AC.

5.2 Future Work

This study opens up many new exciting further research opportunities. First, future work can further investigate the explanation type in various degrees of risk by conducting in-person studies. The in-person studies will enable the researcher to investigate further and measure various factors in relation to the explanation type needed. This in-person study also enables the researcher to conduct a qualitative study regarding the explanation satisfaction.

Regarding technology development, we believe there is an opportunity to develop a system that can measure AC's passengers perceived situational risk and adjust the explanation given by the AC accordingly. This system can utilize multiple sensors to be able to measure the perceived situational risk of the passengers.

Overall, many improvements can still be made to create a better experience for humans in an AC and help boost the adoption of AC on the road.

Chapter 6

Conclusions

In conclusion, this study aimed to investigate the impact of situational risk and visual information availability on the type of explanation needed by AC's passengers and how these factors influence passengers' trust in the AC. The findings indicate that the availability of visual information affects the type of explanation preferred by passengers. When visual information is unavailable, passengers are more satisfied with causal explanations that explain the cause of certain AC's actions. On the other hand, when visual information is available, passengers prefer intentional explanations that provide reasons behind the AC's actions. These results suggest that visual information reduces cognitive workload (compared to when no visual information is available) and enables passengers to process more detailed explanations.

Furthermore, this study revealed that passengers are significantly more satisfied with intentional explanations when visual information is available in low and medium-risk situations. However, in high-risk situations without visual information, passengers show slightly higher satisfaction with intentional explanations. This finding implies that when faced with higher risk and limited visual information, passengers may rely on more detailed explanations to compensate for the uncertainty.

Regarding trust in the AC, no significant differences were found between the groups. However, participants showed slightly higher trust in the AC that provided causal explanations in situations without visual information. This result rejected the **H2** that the combination of visual information and intentional explanations would generate the highest trust. The result suggests that trust ratings might be influenced by factors other than the provided explanation, such as the AC's performance in the given scenario. We also highlight the need for further research and refinement of trust scales used in

evaluating human trust in autonomous systems.

Overall, this study can contribute to a better understanding of passengers' preferences and expectations regarding explanations provided by AC. By considering the situational degree of risk and visual information availability, future designs and implementations can create explanation strategies to enhance passenger satisfaction and trust, thus improving the overall user experience of AC.

References

[1] I. SAE, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles j3016_202104 (p. 41)," *SAE International*, 2021. [Page 1.]

[2] G. Wiegand, M. Eiband, M. Haubelt, and H. Hussmann, ""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*, 2020, pp. 1–11. [Page 1.]

[3] 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. [Pages 1, 4, 5, and 14.]

[4] S. Sreedharan, A. Kulkarni, and S. Kambhampati, *Explainable Human-AI Interaction: A Planning Perspective*. Springer Nature, 2022. [Pages 1, 2, and 3.]

[5] D. Gergle, R. E. Kraut, and S. R. Fussell, "Using visual information for grounding and awareness in collaborative tasks," *Human–Computer Interaction*, vol. 28, no. 1, pp. 1–39, 2013. [Pages 1, 6, and 24.]

[6] F. Fraboni, L. Gualtieri, F. Millo, M. De Marchi, L. Pietrantoni, and E. Rauch, "Human-robot collaboration during assembly tasks: the cognitive effects of collaborative assembly workstation features," in *Proceedings of the 21st Congress of the International Ergonomics Association (IEA 2021) Volume V: Methods & Approaches 21*. Springer, 2022, pp. 242–249. [Pages 1 and 23.]

[7] R. E. Stuck, B. J. Tomlinson, and B. N. Walker, "The importance of incorporating risk into human-automation trust," *TheoreTical issues in ergonomics science*, vol. 23, no. 4, pp. 500–516, 2022. [Pages 2 and 5.]

[8] Q. V. Liao, M. Singh, Y. Zhang, and R. Bellamy, "Introduction to explainable ai," in *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–3. [Page 3.]

[9] K. Gade, S. C. Geyik, K. Kenthapadi, V. Mithal, and A. Taly, "Explainable ai in industry," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 3203–3204. [Page 3.]

[10] I. Ahmed, G. Jeon, and F. Piccialli, "From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5031–5042, 2022. [Page 3.]

[11] R. W. Andrews, J. M. Lilly, D. Srivastava, and K. M. Feigh, "The role of shared mental models in human-ai teams: a theoretical review," *Theoretical Issues in Ergonomics Science*, pp. 1–47, 2022. [Pages 3 and 14.]

[12] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018. [Page 3.]

[13] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, and A. Holzinger, "Explainable ai: the new 42?" in *Machine Learning and Knowledge Extraction: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27–30, 2018, Proceedings 2*. Springer, 2018, pp. 295–303. [Page 3.]

[14] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020. [Pages 3 and 4.]

[15] D. Holliday, S. Wilson, and S. Stumpf, "User trust in intelligent systems: A journey over time," in *Proceedings of the 21st international conference on intelligent user interfaces*, 2016, pp. 164–168. [Page 3.]

[16] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "Xai—explainable artificial intelligence," *Science robotics*, vol. 4, no. 37, p. eaay7120, 2019. [Page 3.]

[17] M. U. Islam, M. Mozaharul Mottalib, M. Hassan, Z. I. Alam, S. Zobaed, and M. Fazle Rabby, "The past, present, and prospective future of xai: A comprehensive review," *Explainable Artificial Intelligence for Cyber Security: Next Generation Artificial Intelligence*, pp. 1–29, 2022. [Page 3.]

[18] P. Gohel, P. Singh, and M. Mohanty, "Explainable ai: current status and future directions," *arXiv preprint arXiv:2107.07045*, 2021. [Page 3.]

[19] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "Explanations in autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, 2021. [Page 4.]

[20] R. Hussain and S. Zeadally, "Autonomous cars: Research results, issues, and future challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1275–1313, 2018. [Page 4.]

[21] L. Meyer-Waarden and J. Cloarec, ""baby, you can drive my car": Psychological antecedents that drive consumers' adoption of ai-powered autonomous vehicles," *Technovation*, vol. 109, p. 102348, 2022. [Page 4.]

[22] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions," *arXiv preprint arXiv:2112.11561*, 2021. [Page 4.]

[23] R. Setchi, M. B. Dehkordi, and J. S. Khan, "Explainable robotics in human-robot interactions," *Procedia Computer Science*, vol. 176, pp. 3057–3066, 2020. [Page 4.]

[24] 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, pp. 269–275, 2015. [Pages 4, 14, 23, and 24.]

[25] J. M. Schraagen, P. Elsasser, H. Fricke, M. Hof, and F. Ragalmuto, "Trusting the x in xai: Effects of different types of explanations by a self-driving car on trust, explanation satisfaction and mental models," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 64, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2020, pp. 339–343. [Pages 5, 11, 14, and 23.]

[26] L. Petersen, H. Zhao, D. Tilbury, X. J. Yang, L. Robert *et al.*, "The influence of risk on driver's trust in semi-autonomous driving," 2018. [Page 5.]

[27] K. Titchener, M. J. White, and S.-A. Kaye, "driver distractions: characteristics underlying drivers' risk perceptions. in: Proceedings, 10-12 november 2009, sydney convention and exhibition centre, sydney, new south wales." 2009. [Page 5.]

[28] M. Li, B. E. Holthausen, R. E. Stuck, and B. N. Walker, "No risk no trust: Investigating perceived risk in highly automated driving," in *Proceedings of the 11th international conference on automotive user interfaces and interactive vehicular applications*, 2019, pp. 177–185. [Page 6.]

[29] F. Cuzzolin, A. Morelli, B. Cirstea, and B. J. Sahakian, "Knowing me, knowing you: theory of mind in ai," *Psychological medicine*, vol. 50, no. 7, pp. 1057–1061, 2020. [Page 6.]

[30] A. Tabrez, M. B. Luebbers, and B. Hayes, "A survey of mental modeling techniques in human–robot teaming," *Current Robotics Reports*, vol. 1, pp. 259–267, 2020. [Pages 6 and 15.]

[31] M. Barnes, F. Jentsch, J. Y. Chen, E. Haas, and K. Cosenzo, "Five things you should know about soldier-robot teaming," ARMY RESEARCH LAB ABERDEEN PROVING GROUND MD HUMAN RESEARCH AND ENGINEERING …, Tech. Rep., 2008. [Page 6.]

[32] A.-R. Blais and E. U. Weber, "A domain-specific risk-taking (dospert) scale for adult populations," *Judgment and Decision making*, vol. 1, no. 1, pp. 33–47, 2006. [Pages 9, 37, and 38.]

[33] R. E. Stuck, "Perceived relational risk and perceived situational risk: Scale development," Ph.D. dissertation, Georgia Institute of Technology, 2020. [Pages 9, 13, and 37.]

[34] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018. [Pages 9, 12, 13, 14, 37, and 38.]

[35] BeamNG GmbH, "Beamng.drive," https://bit.ly/beamng-aboutus. [Page 11.]

[36] S. Kaplan, I. Mikolasek, H. Bruhova Foltynova, K. H. Janstrup, and C. G. Prato, "Attitudes, norms and difficulties underlying road sharing intentions as drivers and cyclists: Evidence from the czech republic," *International journal of sustainable transportation*, vol. 13, no. 5, pp. 350–362, 2019. [Pages 14 and 38.]

[37] Y. Du, J. Qin, S. Zhang, S. Cao, and J. Dou, "Voice user interface interaction design research based on user mental model in autonomous vehicle," in *Human-Computer Interaction. Interaction Technologies: 20th International Conference, HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part III 20*. Springer, 2018, pp. 117–132. [Page 14.]

[38] S. Thellman and T. Ziemke, "Do you see what i see? tracking the perceptual beliefs of robots," *Iscience*, vol. 23, no. 10, p. 101625, 2020. [Page 14.]

[39] X. Zhao, C. Cusimano, and B. F. Malle, "Do people spontaneously take a robot's visual perspective?" in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, 2015, pp. 133–134. [Page 14.]

[40] B. R. Schadenberg, D. Reidsma, D. K. Heylen, and V. Evers, ""i see what you did there" understanding people's social perception of a robot and its predictability," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 3, pp. 1–28, 2021. [Page 14.]

[41] F. Ekman, M. Johansson, L.-O. Bligård, M. Karlsson, and H. Strömberg, "Exploring automated vehicle driving styles as a source of trust information," *Transportation research part F: traffic psychology and behaviour*, vol. 65, pp. 268–279, 2019. [Page 15.]

[42] J. Cohen, *Statistical power analysis for the behavioral sciences*. Academic press, 2013. [Page 15.]

[43] M. Tavakol and R. Dennick, "Making sense of cronbach's alpha," *International journal of medical education*, vol. 2, p. 53, 2011. [Page 16.]

[44] IBM Corporation, "Ibm spss statistics." [Online]. Available: https://www.ibm.com/products/spss-statistics [Page 16.]

[45] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952. [Pages 17 and 20.]

[46] P. E. McKight and J. Najab, "Kruskal-wallis test," *The corsini encyclopedia of psychology*, pp. 1–1, 2010. [Page 17.]

[47] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947. [Page 18.]

[48] B. C. Kok and H. Soh, "Trust in robots: Challenges and opportunities," *Current Robotics Reports*, vol. 1, pp. 297–309, 2020. [Page 25.]

Appendix A

Detailed Experimental Design

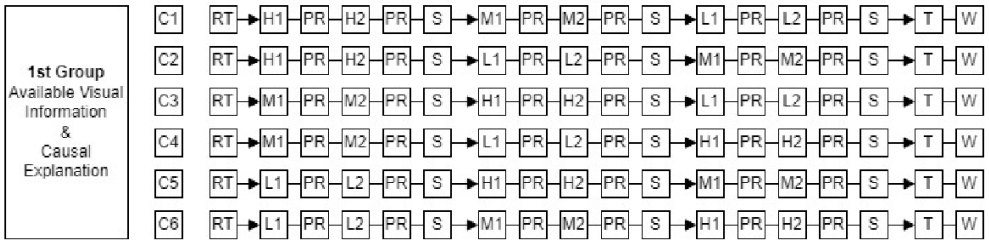


Figure A.1: Detailed Experiment Design on Group 1

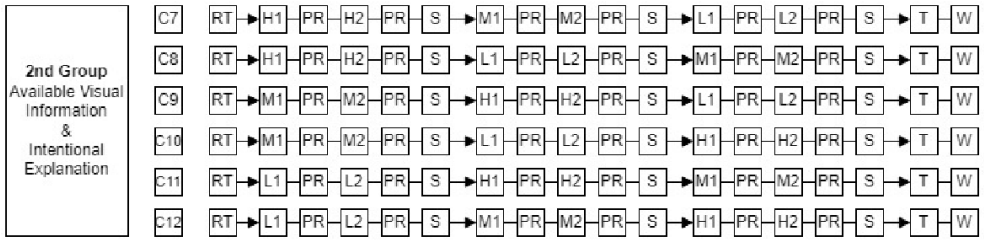


Figure A.2: Detailed Experiment Design on Group 2

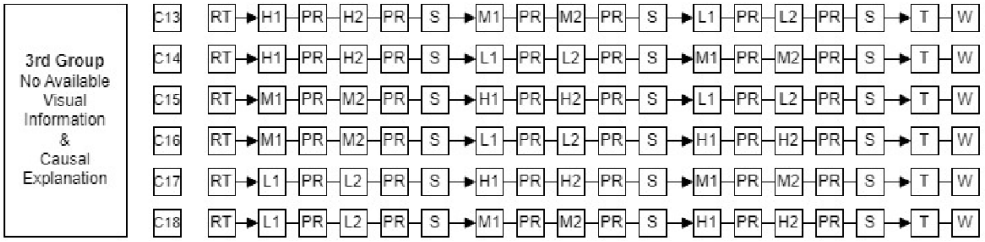


Figure A.3: Detailed Experiment Design on Group 3

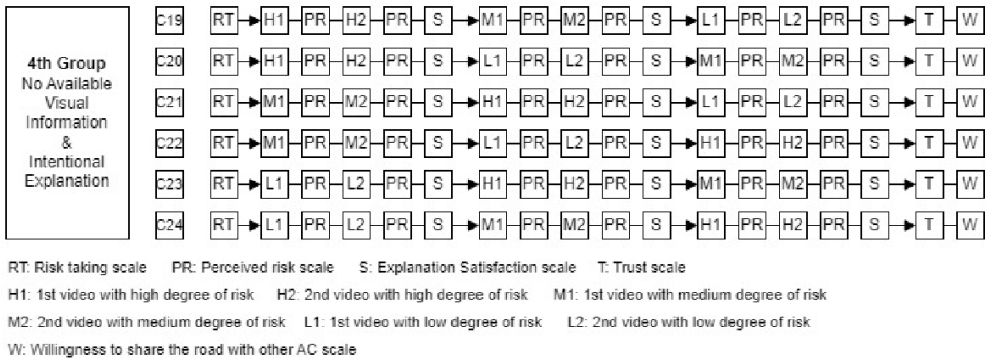


Figure A.4: Detailed Experiment Design on Group 4

Appendix B

Scales Used in this Study

Table B.1: Scales used in this study

Item	Adapted from
Risk Perception	[32]
Driving a car without wearing a seat belt	
Riding a motorcycle without a helmet	
Sunbathing without sunscreen	
Walking home alone at night in an unsafe area of town	
Perceived Situational Risk	[33]
I was very concerned about being in that situation	
I was very fearful of being in that situation	
I was very anxious about being in that situation	
I was very nervous about being in that situation	
I was very concerned about the effects that being in that situation could have on me	
Explanation Satisfaction	[34]
From the explanation, I understand the car actions	
The explanation is satisfying	
The explanation has sufficient detail	
The explanation seems complete	
The explanation tells me how the car behave	
The explanation of the car actions is useful to me	
The explanation of the car actions shows me how accurate the system is	

Table B.2: Scales used in this study (continue)

Item	Adapted from
Risk Perception	[32]
Trust Scale	[34]
I am confident in the car. I feel that it works well.	
The actions of the car are very predictable.	
The car is very reliable. I can count on it all the time.	
I feel safe that when I rely on the car I will have a nice drive experience.	
The car is efficient in that it works accurately.	
I am wary of the car.	
The car can drive better than a novice human driver.	
Willingness to Share The Road	[36]
Autonomous cars should have their own infrastructure.	
Autonomous cars are too slow/fragile to be used on the regular road.	
Autonomous and regular cars can share the same roads.	
As a driver I am/would be afraid to drive around autonomous cars.	
I am afraid of autonomous-regular cars accidents.	

Appendix C

Internal Consistency Analysis Summary

Table C.1: Internal Consistency Analysis Group 1

Scale	Cronbach's Alpha
Perceived Risk HAC1	0.956
Perceived Risk HAC2	0.958
Perceived Risk MAC1	0.962
Perceived Risk MAC2	0.970
Perceived Risk LAC1	0.973
Perceived Risk LAC2	0.979
Satisfaction in High Risk	0.971
Satisfaction in Medium Risk	0.940
Satisfaction in Low Risk	0.954
Trust	0.827
Willingness to Share the Road	0.312

Table C.2: Internal Consistency Analysis Group 2

Scale	Cronbach's Alpha
Perceived Risk HAI1	0.933
Perceived Risk HAI2	0.938
Perceived Risk MAI1	0.964
Perceived Risk MAI2	0.961
Perceived Risk LAI1	0.978
Perceived Risk LAI2	0.986
Satisfaction in High Risk	0.956
Satisfaction in Medium Risk	0.926
Satisfaction in Low Risk	0.937
Trust	0.790
Willingness to Share the Road	0.136

Table C.3: Internal Consistency Analysis Group 3

Scale	Cronbach's Alpha
Perceived Risk HNC1	0.934
Perceived Risk HNC2	0.949
Perceived Risk MNC1	0.969
Perceived Risk MNC2	0.965
Perceived Risk LNC1	0.956
Perceived Risk LNC2	0.952
Satisfaction in High Risk	0.929
Satisfaction in Medium Risk	0.948
Satisfaction in Low Risk	0.932
Trust	0.675
Willingness to Share the Road	0.469

Table C.4: Internal Consistency Analysis Group 4

Scale	Cronbach's Alpha
Perceived Risk HNI1	0.910
Perceived Risk HNI2	0.932
Perceived Risk MNI1	0.957
Perceived Risk MNI2	0.954
Perceived Risk LNI1	0.968
Perceived Risk LNI2	0.969
Satisfaction in High Risk	0.934
Satisfaction in Medium Risk	0.938
Satisfaction in Low Risk	0.919
Trust	0.716
Willingness to Share the Road	0.359

Appendix D

Order Effect Summary

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Satisf_HAC_Mean is the same across categories of C.	Independent-Samples Kruskal-Wallis Test	.148	Retain the null hypothesis.
2	The distribution of Satisf_MAC_Mean is the same across categories of C.	Independent-Samples Kruskal-Wallis Test	.437	Retain the null hypothesis.
3	The distribution of Satisf_LAC_Mean is the same across categories of C.	Independent-Samples Kruskal-Wallis Test	.220	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Figure D.1: Order Effect Check for Group 1

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of SATISF_HAI is the same across categories of C.	Independent-Samples Kruskal-Wallis Test	.184	Retain the null hypothesis.
2	The distribution of SATISF_MAI is the same across categories of C.	Independent-Samples Kruskal-Wallis Test	.142	Retain the null hypothesis.
3	The distribution of SATISF_LAI is the same across categories of C.	Independent-Samples Kruskal-Wallis Test	.501	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Figure D.2: Order Effect Check for Group 2

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of SATISF_HNC is the same across categories of C.	Independent-Samples Kruskal-Wallis Test	.134	Retain the null hypothesis.
2	The distribution of SATISF_MNC is the same across categories of C.	Independent-Samples Kruskal-Wallis Test	.317	Retain the null hypothesis.
3	The distribution of SATISF_LNC is the same across categories of C.	Independent-Samples Kruskal-Wallis Test	.609	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Figure D.3: Order Effect Check for Group 3

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of SATISF_HNI is the same across categories of C.	Independent-Samples Kruskal-Wallis Test	.245	Retain the null hypothesis.
2	The distribution of SATISF_MNI is the same across categories of C.	Independent-Samples Kruskal-Wallis Test	.831	Retain the null hypothesis.
3	The distribution of SATISF_LNI is the same across categories of C.	Independent-Samples Kruskal-Wallis Test	.838	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Figure D.4: Order Effect Check for Group 4

Appendix E

Perceived Risk Summary

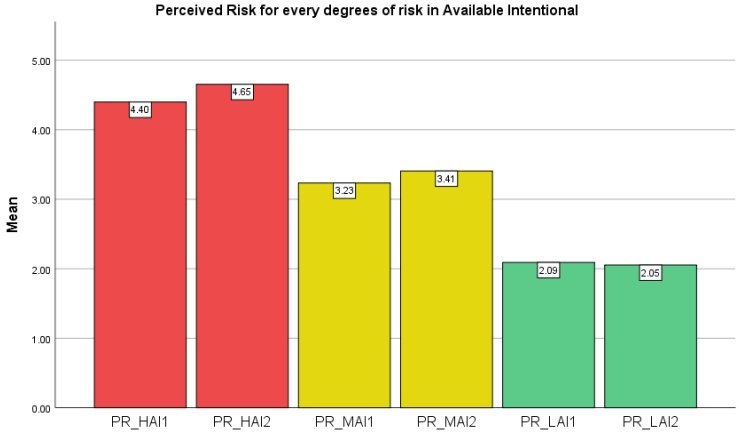


Figure E.1: Perceived Risk in Group 2

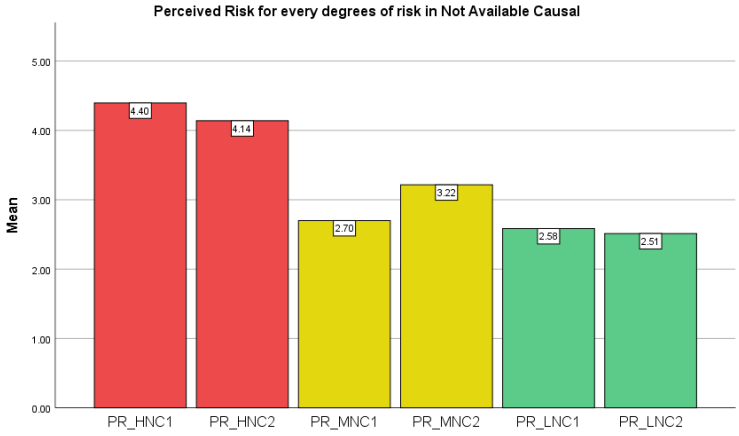


Figure E.2: Perceived Risk in Group 3

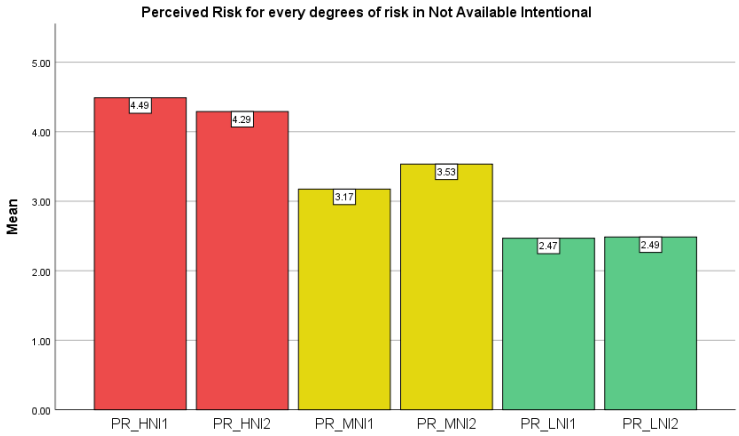


Figure E.3: Perceived Risk in Group 4

Appendix F

Trust towards AC

Independent-Samples Kruskal-Wallis Test Summary

Total N	364
Test Statistic	3.392[a,b]
Degree Of Freedom	3
Asymptotic Sig.(2-sided test)	.335

a. The test statistic is adjusted for ties.

b. Multiple comparisons are not performed because the overall test does not show significant differences across samples.

Figure F.1: Kruskal-Wallis Test on Trust

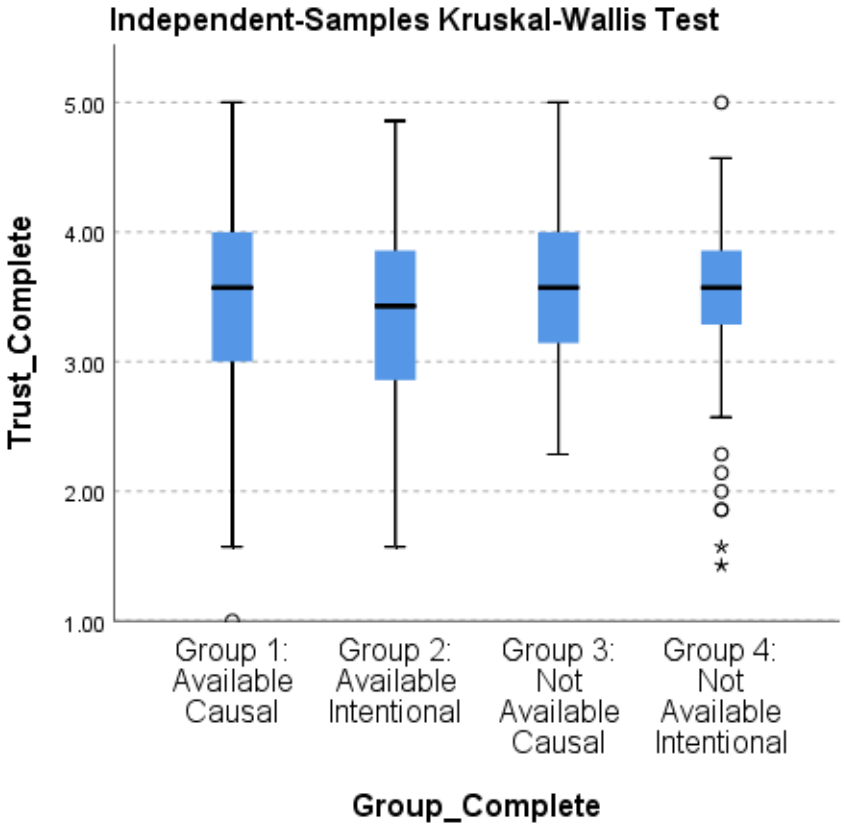


Figure F.2: Mean-rank results from the Kruskal-Wallis Test

Appendix G

Additional Material

TRITA-EECS-EX-2023:637