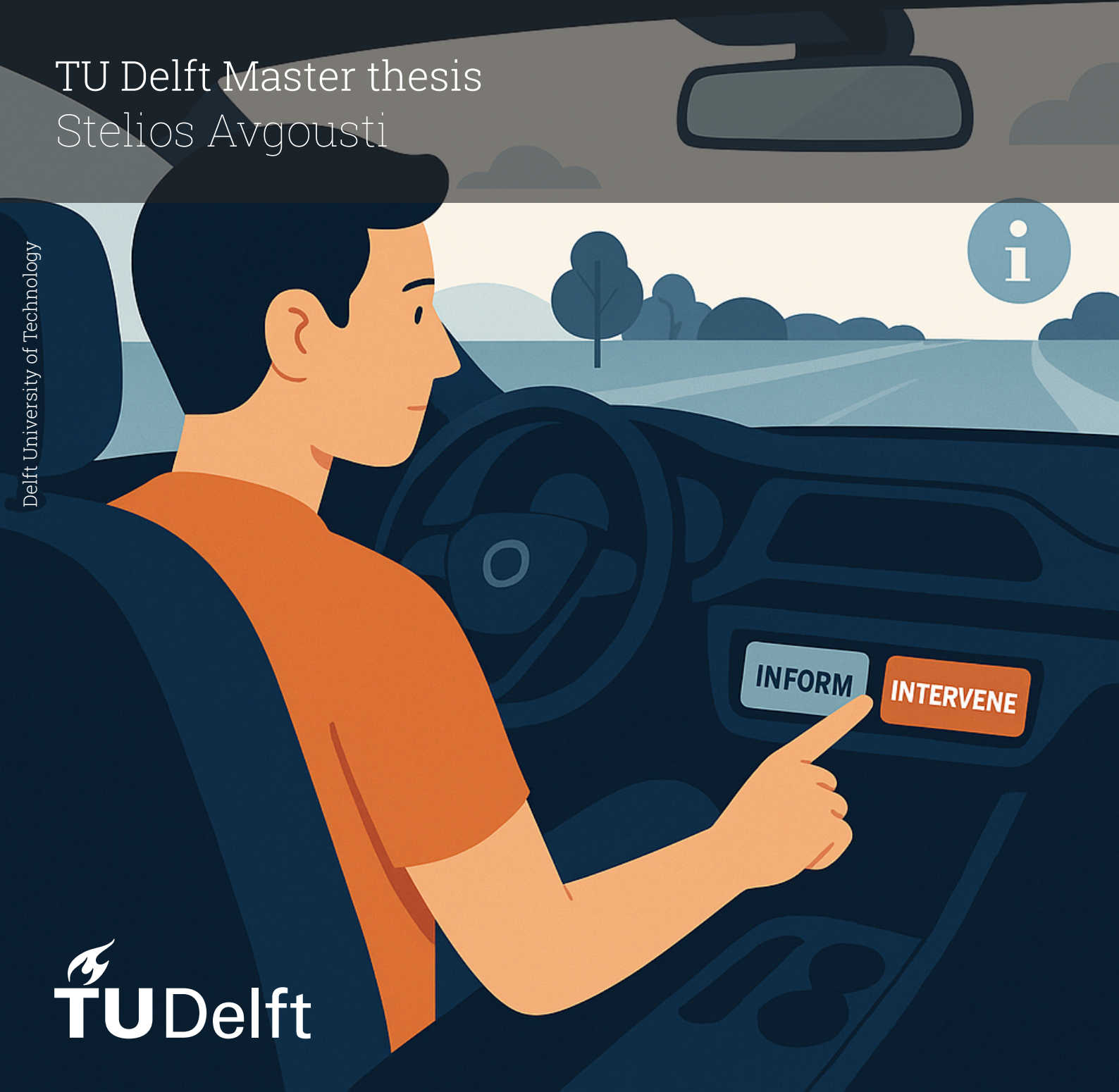
Do you trust your autonomous vehicle? A story of modality

Autonomous Vehicles and Trust

TU Delft Master thesis

Stelios Avgousti



Delft University of Technology

TU Delft

Do you trust your autonomous vehicle? A story of modality

Autonomous Vehicles and Trust

by

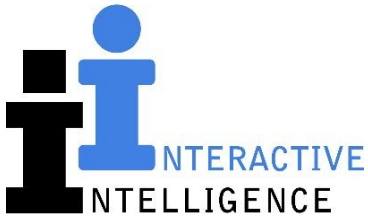
Stelios Avgousti

Instructor:	M. Tielman
Teaching Assistant:	A. George, R. Verhagen, C. Jorge
Project Duration:	2024-2025
Faculty:	Faculty of Computer Science and Engineering, Cognitive Robotics Lab, TU Delft

Cover: Canadarm 2 Robotic Arm Grapples SpaceX Dragon by NASA under CC BY-NC 2.0 (Modified)

TU Delft

The work in this thesis was conducted at the:



Interactive Intelligence
Department of Intelligent Systems
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

AND

Department of Cognitive Robotics
Faculty of Mechanical, Maritime and Materials Engineering
Delft University of Technology

Abstract

As vehicles advance toward full autonomy, SAE Level 5 systems are typically designed and envisioned without driver controls. While this promises convenience and safety, it also raises challenges around user trust and the discomfort of having no agency. This study investigates how explanation modality (vocal vs. text) and optional control mechanisms influence trust and intervention behavior in a simulated SAE Level 5 context. Thirty-six participants completed three VR driving scenarios that varied in explanation and control design. Trust and intervention behavior were measured alongside thematic analysis of open-ended feedback. Results showed that vocal explanations increased trust more than text, though not significantly. However, the presence of control buttons significantly enhanced trust among participants who perceived them positively. These participants also intervened less often, though the effect was not statistically significant. Exploratory analyses revealed that self-reported comfort with automation was associated with higher trust and lower intervention rates. These findings challenge the SAE Level 5 assumption of no user input. Even minimal, optional control features can foster trust and reduce unnecessary interventions. The study underscores the value of designing autonomous systems that maintain transparency and user agency, supporting safer and more acceptable human-AI interaction.

Acknowledgements

I believe this is the only place where the writing style and tone can be different. This is the only section of the report where I can be personal and write in a more welcoming tone. I was doing my thesis while working and socializing which was really demanding at first. I could use all the help I could get, and thankfully the right people were there to guide me through it. There are many people to thank and appreciate. First of all, I would like to thank all the participants of this experiment which took time out of their schedules to participate and help me finish what I started. I want to thank Myrthe Tielman for being my advisor and being open to anything I proposed. I want to thank Carolina Jorge, Reuben Verhagen and Ashwin George for always being available and giving me really constructive feedback and helping me achieve a better project. Also Id like to thank Andreas Achilleos that helped me with coding the open-ended questions as a secondary coder. I want to thank Federico Sari, for being my unofficial mentor and helping me during the whole duration of my thesis, which without, it might've taken way longer to finish. I want to thank Guus van Heijningen, for giving me some incredibly needed off time from work so I could focus on finishing. Lastly I want to thank my parents and my girlfriend, for supporting me mentally this whole time and never doubted me for a second. This has been an incredible experience for me from but im glad its over, as I can now move into a new chapter of my life. Thanks again.

Contents

List of Figures

List of Tables

1 Introduction

1.1. SAE Level 5 Autonomous Vehicles

SAE International defines six levels of vehicle automation (Levels 0–5), specifying whether the human driver or the system is responsible for tasks such as steering, braking, and monitoring the driving environment [34]. Level 0 corresponds to no automation, while Level 5 indicates full automation under all driving conditions. The transition to Level 5 autonomous vehicles (AVs) marks a paradigm shift in mobility, promising significant improvements in safety, efficiency, and accessibility [24]. Human error accounts for over 90% of road accidents globally; eliminating this factor through full automation could substantially reduce fatalities and transform transportation systems [39, 14, 12].

However, the success of AVs hinges not only on technical robustness but also on user trust. Trust is shaped by factors such as transparency, perceived control, and user experience [11, 19]. Control in automated driving refers to the authority to influence the vehicle's trajectory or behavior. In SAE terms, this spans a spectrum: full human control (Levels 0–2), shared or conditional control (Levels 3–4), and full system control with no human intervention (Level 5) [10].

The journey toward fully autonomous vehicles has been marked by key milestones, notably the DARPA Urban Challenge, which catalyzed major advancements in AV technology [40, 4]. These early systems demonstrated the feasibility of autonomous driving in structured environments but also revealed the difficulty of achieving human-like decision-making in dynamic urban settings [23]. While current research aims for fully autonomous Level 5 systems, practical deployments still relied on human supervision, highlighting the need for continued work in perception, planning, and robust system integration. Therefore, it is essential to design systems that balance full autonomy with mechanisms that foster trust and engagement.

Achieving SAE Level 5 autonomy today involves not only refining algorithms and sensor systems but also addressing human factors that affect adoption, particularly how users perceive and trust a system that functions entirely without their oversight.

1.2. Trust, Explainable AI, Cognitive Load

Central to the success of Level 5 AVs is maintaining user trust while preserving full automation. Optional controls, such as a button that signals attention, allow users to feel engaged without compromising the AV's autonomy [39, 16]. These controls mitigate feelings of helplessness and foster trust by enabling users to communicate with the system during high-stakes situations [31]. However, the presence of such controls also introduces the risk of wrong/unwarranted interventions, disrupting the vehicle's intended operation. A wrong (unnecessary) intervention is where a user would deviate from the pre-defined behaviour that the AV had planned in situations where it would have been both safe and efficient. Such user-initiated overrides do not avert hazards, instead, they degrade comfort, slow traffic, or create new conflict situations. Trust is a critical determinant of AV acceptance and effective operation. Literature identifies a dual challenge: addressing under trust, which hinders user engage-

ment, and over trust, which can lead to misuse of the technology [42, 6]. Trust calibration, aligning user expectations with the system's actual capabilities, has emerged as a critical strategy to mitigate these issues. Systems lacking clear communication exacerbate distrust, leading to disengagement or inappropriate interventions [5]. Conversely, clear and contextually relevant feedback has been shown to build trust and user satisfaction [21, 19].

Explainable AI (XAI) plays a vital role in fostering trust by providing clear, actionable insights into AV decision-making [2, 1]. XAI reduces user uncertainty by clarifying why the system behaves as it does, helping align user mental models with system behavior and lowering the chance of unnecessary interventions. AV explanations can be divided into ”how” and ”why” messages. ”How” messages (e.g., ”The car is stopping”) inform users of system actions, offering straightforward operational feedback. ”Why” messages (e.g., ”Obstacle ahead”) explain the reasoning behind system behavior, fostering a deeper understanding and trust [21, 41, 42]. Studies suggest that combining both ”how” and ”why” messages provides the most comprehensive explanations, although it may increase cognitive load in certain contexts [6]. The modality of explanation delivery further impacts trust calibration. Vocal cues provide immediate, easily processed information, reducing cognitive effort and enhancing situational awareness, especially in time sensitive situations [5, 37]. In contrast, text-based explanations allow for more detailed and deliberate information processing, supporting thoughtful decision making [42].

Cognitive load, the mental effort required to process information, is a critical factor in designing effective explanation mechanisms. High cognitive load caused by complex or ambiguous information can overwhelm users, hinder comprehension, and increase stress, potentially leading to inappropriate actions [37, 28]. Tailoring explanation modalities to minimize cognitive load is essential for effective trust calibration. For example, while vocal explanations can quickly deliver critical information, they may unintentionally prompt hasty user actions if overused. Conversely, text explanations can support more deliberate decision-making but may increase cognitive load in fast-paced scenarios [5].

1.3. Research goal

Traditional research on AV trust has largely focused on enhancing the user's ability to take control during system failures, assuming imperfections in the AV [36]. This focus overlooks a critical aspect in SAE Level 5 systems, which are designed to operate autonomously without requiring human input. In this context, user driven errors, such as unwarranted use of control mechanisms, emerge as a more critical concern than system failures.

Current research has not yet fully explored how the combination of explanation modalities and optional control mechanisms influences user trust and intervention behavior in fully autonomous systems. Unlike Level 4 systems, where a fallback driver must assume control outside a limited operational design domain, Level 5 vehicles are expected to self-recover from all edge cases. This shifts the focus from system failure to driver-induced disturbances, such as unnecessary interventions. This study aims to investigate how different explanation modalities (vocal vs. text) and optional control mechanisms influence user trust and the frequency of unnecessary interventions in SAE Level 5 AVs. By examining this dynamic, the study aims to inform the design of AV systems that effectively balance autonomy and user engagement

1.3.1. Research Question

This thesis aims to answer the following research question:

How do explanation modalities (vocal vs. text explanations) and the availability of control mechanisms influence user trust and intervention behavior in SAE Level 5 autonomous vehicles?

To address this question, an experimental study was designed in which participants experienced different driving conditions varying in explanation modality and control availability. Trust levels, intervention behavior, and subjective feedback were collected to examine how these factors interact and affect user experience.

1.3.2. Chapter Overview

The remainder of this thesis is divided into five chapters. Chapter 2 reviews the relevant literature on trust in automation, explainable AI, cognitive load, and control mechanisms in autonomous vehicles. Chapter 3 explains the experimental design, including the driving simulator setup, participant selection, and data collection procedures. Chapter 4 presents the findings, including quantitative analyses and thematic coding of qualitative feedback. Chapter 5 discusses the findings in relation to the research question and prior literature, addressing limitations and highlighting key takeaways. Chapter 5.3 outlines recommendations for future work, followed by the conclusion in Chapter 6.

2

Background

The development of SAE Level 5 autonomous vehicles (AVs) presents an unprecedented opportunity to revolutionize mobility, enhancing safety, efficiency [39, 14, 12, 24].

However, earning public trust in fully autonomous vehicles remains a key barrier to their widespread adoption. Addressing this requires understanding the technical limitations of AVs, including their susceptibility to unexpected environmental scenarios, sensor failures, and ambiguous road situations. For instance, both perceptual and planning errors have been observed when AVs encounter novel or poorly understood environments [4, 23]. In addition, users might intervene inappropriately due to misinterpreting system behavior, thereby disrupting optimal AV performance. In systems designed to minimize such errors, user-initiated mistakes, driven by mistrust or misunderstanding, can become the primary source of operational inefficiencies. This section explores foundational concepts related to trust in automation, user interaction design, feedback modalities, cognitive workload, and the dynamics of user errors in SAE Level 5 AVs.

2.1. Autonomous Vehicles: Taxonomy, Challenges and Expected Errors

Autonomous vehicles are classified using the SAE taxonomy, ranging from Level 0 (no automation) to Level 5 (full automation) [34]. Level 5 AVs offer benefits such as improved road safety, broader mobility access, and increased traffic efficiency [23, 40]. While technical reliability continues to improve, user interaction introduces a distinct class of challenges [36].

SAE Levels of Automation:

- Level 0 (No Automation): The driver is fully responsible for all driving tasks, though the system may offer warnings or momentary assistance.
- Level 1 (Driver Assistance): The system assists with either steering or acceleration/deceleration, but not both simultaneously (e.g., cruise control).
- Level 2 (Partial Automation): The vehicle can control both steering and speed but requires constant driver supervision (e.g., Tesla Autopilot, GM Super Cruise).
- Level 3 (Conditional Automation): The AV manages most driving tasks but may request driver intervention in complex scenarios (e.g., traffic jams).
- Level 4 (High Automation): The AV handles all driving functions in specific conditions or environments (e.g., urban areas), but human takeover may still be necessary outside those conditions.
- Level 5 (Full Automation): The vehicle performs all driving tasks under any condition, eliminating the need for human input or traditional controls such as a steering wheel or pedals.

AVs must navigate complex urban environments, adapt to dynamic traffic scenarios, and address challenging cases such as occlusions or unpredictable agent behavior [4]. Expected errors include sensor

malfunctions, object misclassifications, and suboptimal decisions in ambiguous situations [3]. However, as system reliability improves, user-driven disruptions become more prominent. Users may unnecessarily engage with optional control mechanisms due to uncertainty or misjudgment, interrupting autonomous operation. This highlights the importance of interaction mechanisms that maintain trust and reduce unnecessary intervention. Shared control approaches—where users can intervene or signal concerns—must be carefully designed to balance transparency and system robustness [16, 31].

2.2. Shared Control and User Interaction

User interaction design in AVs must balance full autonomy with opportunities for user engagement to maintain trust. Shared control mechanisms and adaptive decision-making can foster a sense of involvement while preserving system performance and safety [39, 19, 41].

In SAE Levels 2-4, driver and automation alternate or share the control loop, often called shared control. At Level 5, the vehicle executes the control loop independently. However, perceived control features may still be present, even though they are not formally required. These features should be one-way: users can influence the AV, but the AV never expects the human to resume continuous driving.

User concerns about relinquishing control to automation have been shown to significantly affect trust and adoption of autonomous vehicles [14]. While traditional AV systems allowed for user intervention during failures, future SAE Level 5 systems may challenge this dynamic. Understanding how the presence or absence of control interfaces influences trust remains crucial for designing effective trust calibration strategies.

Historically, shared control interfaces like steering wheels and pedals have been included in semi-autonomous vehicles to facilitate user takeovers during failures. However, in fully autonomous SAE Level 5 systems, these controls may no longer serve a functional role. It is shown that trust in automation and concerns about giving up control are critical factors influencing adoption of autonomous vehicles [12]. Their findings suggest that the removal of control interfaces could amplify users' discomfort with loss of control, thereby undermining trust. This highlights the need for trust calibration strategies that address both cognitive trust in system reliability and affective concerns about autonomy [14].

Critically, the effectiveness of shared control mechanisms depends on how well the system communicates its actions and limitations. Clear explanations are essential for minimizing unnecessary interventions and maintaining trust.

2.3. Trust in Automation

Research identifies a dual challenge in trust in automation: under-trust leads to disengagement and hesitancy, while over-trust can lead to misuse [42, 6]. Striking a balance through trust calibration, aligning user expectations with system capabilities, is therefore essential.

Studies have shown that systems lacking clear communication or exhibiting failure can exacerbate distrust, resulting in reduced cooperation or disengagement [5]. Conversely, contextually relevant explanations, particularly those conveying both intent and reasoning, can enhance trust and satisfaction in semi, autonomous systems [21, 19]. These findings underscore the importance of trust calibration: aligning user expectations with system capabilities to support effective human-agent collaboration.

Trust in automation is influenced by several key factors:

- Perceived Competence: Users must believe that the AV can handle complex driving tasks safely and effectively [22, 25].
- Transparency and Explainability: Systems must clearly convey how and why decisions are made to foster user understanding and acceptance [9, 2].
- User Control and Engagement: Offering optional control mechanisms can enhance trust by giving users a sense of agency [39]. However, misuse of these controls can undermine system performance.
- Cognitive Load : Trust calibration is closely influenced by the user's cognitive load. High cognitive

load can overwhelm users, leading to stress and impaired decision-making. Thus, designing systems that minimize cognitive load through intuitive feedback and interaction mechanisms is crucial for fostering trust and reducing unnecessary interventions [37, 28].

2.4. Trust as a Risk-Taking relationship

Trust in automation is not simply a static attitude. It is fundamentally a risk-taking relationship, as described in Mayer et al.'s model and extended by Johnson and Bradshaw [18]. Trust involves "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor." This definition highlights that trust inherently involves accepting vulnerability and risk. The user must rely on the system with the expectation that it will act competently and in alignment with user goals. In an autonomous vehicle (AV) context, this means a rider trusts the AV to drive safely and effectively, risking their safety on the assumption that the vehicle will not fail. Crucially, trust is manifested through action: a user's trusting behaviors (or lack thereof) indicate their willingness to take risks. For example, choosing to rely on the AV's decisions without intervening reflects a high degree of trust, whereas frequently overriding or doubting the AV indicates low trust. Prior research shows a strong link between trust and reliance on automation: when people trust an automated system, they are far more likely to use it and not intervene, and when trust is low they tend to disuse or override the automation [27]. It is then safe to say that trust directly drives risk-taking behaviors, where trusting the AV means the user willingly takes the risk of not taking control, conversely leads to the user taking control to mitigate risk.

An essential aspect of trust as a risk relationship is the role of user perception and individual differences. System performance alone does not determine trust. A user's prior experiences, personality, and perceptions strongly influence how much risk they are willing to take on the system [18]. Merritt and Ilgen [27] demonstrated that people's perceptions of an automation's attributes (e.g. how safe, reliable, or capable it seems) can account for over half of the variance in their trust, above and beyond the system's actual reliability. In other words, two users might experience the exact same driving performance from an AV, yet develop different trust levels because they interpret that performance differently, perhaps due to one person's greater technophobia or another's familiarity with similar systems. Moreover, they showed that trust has both a dispositional component (a baseline tendency to trust machines) and a situational component that is shaped by interaction history. When administered before any interaction, the trust survey reflected general propensity to trust technology. After interaction, it reflected situational, experience-based trust.

To ground this study theoretically, we adopt the risk-based trust framework from Mayer et al., as extended to human-automation teams by Johnson and Bradshaw [18]. The model identifies key antecedents of trust in a trustee, namely the trustee's perceived ability, integrity, and benevolence, which together shape the trustor's willingness to be vulnerable. Johnson and Bradshaw extend this by framing trust in human–machine interaction as relational and built on interdependence: trust emerges from structured relationships involving observability, predictability, and directability. In autonomous driving scenarios, while the AV does not itself "trust", the interaction design can foster mutual adjustment. The human may provide input in rare cases, while the AV handles routine driving. The Risk-Taking Trust Relationship framework suggests that providing avenues for user engagement increases the user's willingness to trust, because it reduces the perceived stake of vulnerability (the user knows they have a fallback if needed) [38]. In other words, even in a fully self-driving car, trust is not a passive state. It involves the user deciding whether to sit back and let the car drive (taking the risk of non-intervention) or to step in. Therefore we can assume that in a Level 5 context the primary risk may shift from technical failure to mis-calibrated human risk-taking, pressing any input when no hazard exists.

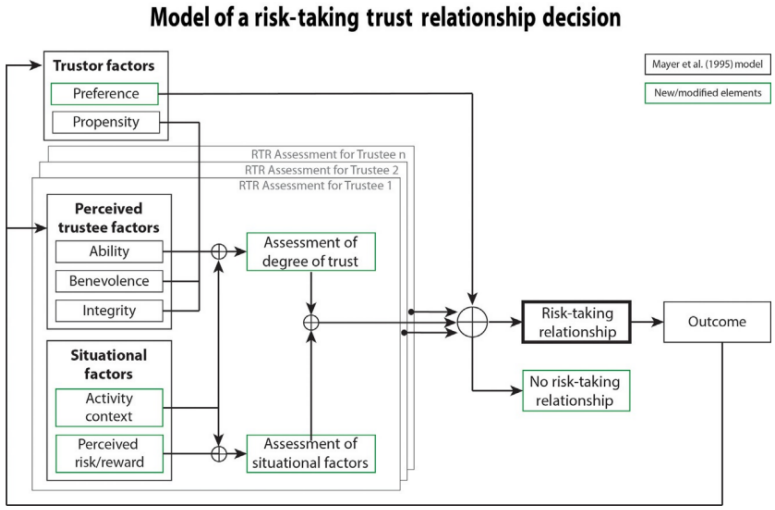


Figure 2.1: Model of a risk-taking trust relationship decision taken from Johnson and Bradshaw [18]

2.5. The Role of Explanations

Empirical studies show that transparency and user control are key factors influencing trust in automation. When the AV clearly explains its decisions or actions, It can reassure users about the system's behavior and decision-making rationale, addressing the antecedents of trust. Providing explanations (either via auditory or visual) makes the system's behavior more transparent, which is known to foster trust by helping users understand why the automation is doing something [20]. Similarly, giving users a form of control or override can bolster trust by granting the user a sense of agency. Participants expressed that the ability to override increased their comfort, particularly in uncertain scenarios [38]. Therefore, incorporating explanation modalities and optional control mechanisms directly ties into the theoretical antecedents of trust: explanations improve predictability and transparency, and control options reduce the perceived risk of trusting the automation.

Explanainable AI (XAI) frameworks enable systems to communicate why decisions were made, improving user trust and understanding [7]. Different explanation types (e.g., causal, intentional, or rationale-based) impact trust development and mental model formation in users [9, 35]. This becomes particularly important in maintaining user safety and managing expectations in systems where AV behavior may otherwise appear unpredictable. Explanations are often categorized into 'how' (descriptive) and 'why' (causal) messages. "How" messages (e.g., "The car is stopping") inform users of system actions, while "why" messages (e.g., "Obstacle ahead") provide reasoning, fostering deeper understanding and trust [21] [41] [42]. Combining both explanation types improves user understanding, though this may increase cognitive load, as discussed earlier. The modality of feedback also impacts trust. Vocal feedback supports fast, low-effort communication, making it effective in time-sensitive scenarios. [5, 37]. Text feedback, on the other hand, allows users to process information more deliberately but may be less effective in urgent scenarios [42]. This transparency into decision-making is the core goal of XAI [7] [35].

Cognitive load significantly affects how users respond during automated driving. High workload—whether from demanding monitoring tasks or secondary activities—can impair performance and lead to inappropriate actions [37, 28]. Consequently, the risk of overwhelming users with excessive or poorly designed information must be carefully managed. Tailoring the communication modality, such as using augmented reality interfaces to provide "how" and "why" explanations, has been shown to enhance trust and situation awareness in safety-critical scenarios [26]. This interplay between workload, trust, and explanation design highlights the importance of developing interfaces that minimize user errors and foster appropriate reliance on automation.

2.6. Knowledge Gaps and Research Direction

Despite extensive research on trust, explanations, and shared control in AVs, the combined effect of explanation modalities and optional control mechanisms on user trust and intervention behavior in SAE Level 5 systems remains understudied.

This study addresses this gap by investigating how different explanation modalities (vocal vs. text) and optional control mechanisms influence user trust and the frequency of unnecessary interventions. Understanding these dynamics is key to designing AVs that maintain full autonomy while encouraging appropriate user engagement and contributing to trust-centric AV design.

3

Methodology

This chapter outlines the methodology used to investigate the research question. A controlled experiment was designed using a driving simulator to expose participants to different combinations of explanation output and available control mechanisms. Participant behavior and attitudes were observed across these conditions to test hypotheses about how trust develops in response to explanation and control variations. The following sections describe our hypothesis, experimental design, participant sample, materials and measures, and the data analysis approach.

3.1. Hypotheses

Given the cognitive processing differences between vocal and textual explanations, the role of user control in trust calibration, and the influence of individual differences such as cognitive load, propensity to trust AI, and prior experience with AVs, the following hypotheses are formulated.

3.1.1. Primary Hypotheses

The main objective of this study is to evaluate how explanation modality affects user trust and intervention behavior.

- H1: Modality → Trust : *Vocal feedback enhances trust more than text-based feedback due to its lower cognitive load.*
- H2: Modality → Intervention Behavior : *Users receiving vocal explanations will press the intervention buttons less frequently, as they feel more reassured.*

3.1.2. Secondary Hypotheses

Beyond modality, other factors, such as the availability of control mechanisms, the presence of explanations, and individual user differences may influence trust and intervention behavior. These will be explored using qualitative data from post-experiment interviews, which will provide context for the quantitative findings by capturing user preferences, perceived explanation effectiveness, and overall trust levels. While the secondary hypotheses are not the primary focus, the collected data may provide valuable insights into their effects and contribute to a broader understanding of user behavior in fully autonomous vehicles.

- H3: Control Mechanism → Trust : *Providing intervention buttons increases trust, as users feel they have a degree of control over the AV's decisions.*
- H4: Explanations → Trust : *The presence of explanations increases trust, as users now feel they have transparency with what the AV is thinking.*

Furthermore, individual biases and situational factors may play a role in shaping user trust and intervention behaviors. These factors will be explored using the pre-study survey in cohesion with the rest of the gathered data:

- *Age does not correlate with how much trust people will have in an AV system*
- *Gender does not correlate with how much trust people will have in an AV system*
- *Users with a naturally higher propensity to trust AI will exhibit higher baseline trust in AVs.*
- *Users familiar with AVs will have higher trust levels compared to those without prior exposure.*
- *Experienced drivers may intervene more often due to risk sensitivity.*
- *Trust and intervention behavior are expected to be negatively correlated (higher trust, fewer interventions)*

3.2. Experimental Design

The study employed a mixed-design structure. The between-subjects factor was explanation modality. The within-subjects factors were explanation availability and user control availability. All participants experienced a progression of trials varying in explanation and control, allowing within-subject comparison of their presence vs. absence. By varying control and explanation within subjects, each participant experienced both the absence and presence of interventions and explanations.

Each participant underwent three trials in the driving simulator, corresponding to increasing levels of system transparency and user control:

- **Trial 1: No Explanations, No Control:** The AV drove autonomously without providing any explanations for its decisions, and the participant had no means to intervene (no buttons). This trial establishes a baseline for the behavior when the AV is effectively a “black box” and the user is entirely hands-off. This is used as a get-to-know trial.
- **Trial 2: No Explanations, Control Available:** The AV still provided no explanations of its actions, but now the participant had access to the “Inform” and “Intervene” buttons (user control mechanisms). This condition isolates the effect of giving the user a way to influence or override the AV, without any explanations. We can observe if merely having control options (even without explanations) changes the user’s tendency to intervene. This round is used as a learning round for understanding how the user interaction buttons work with the vehicle. The participant is instructed to use the buttons as much as they can so they have a good idea how to use the system.
- **Trial 3: Explanations + Control:** In the final trial, the AV provided real-time explanations for its decisions, either via text displayed on the interface or vocally through a headset, depending on the participant’s assigned modality group, and the participant also had the control buttons available. This is the fullest condition, combining transparency (through explanations) with user agency. It is in Trial 3 that we expect the modality effect to manifest: differences in outcomes between the text explanation group and the voice explanation group should become measurable here, since both groups receive explanations but in different formats. We also expect the cumulative effect of having had the ability to intervene (from Trial 2 onward) and now also explanations to influence trust. Button usage was logged exclusively during Trial 3 to assess actual intervention behavior under the combined influence of control and explanations.

The order of trials was the same for all participants (progressing from least transparency/control to most), as this simulates a scenario where an initially fully autonomous system is gradually augmented with user-centric features. To control for potential learning effects or ordering confounds, we made sure that the driving scenarios in each trial were of comparable difficulty and risk. Each trial had a distinct driving route and situation (e.g., different traffic conditions or events), but the set of scenarios was balanced in complexity so that no trial was inherently more challenging than the others. The total experimental phase lasted approximately 20 minutes per participant: 5 minutes for Trial 1, 5 minutes for Trial 2, and 10 minutes for Trial 3, see Figure 3.1.

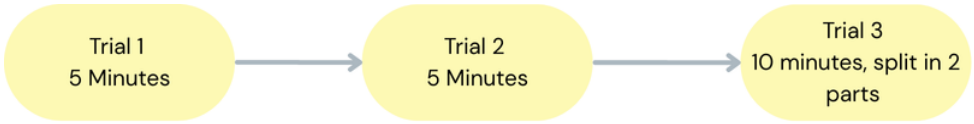


Figure 3.1: Trial Times and Order

3.2.1. Experiment Procedure

Each participant followed the standardized procedure below. Upon arrival, participants signed the informed consent form (A.1). Next, they completed a pre-study survey (A.2) used for analyzing individual differences. Following that, the participant will start engaging in the experimental trials. Participants first experienced the baseline condition, followed by the second and third trials. The third trial is considerably longer than the first two, due to its longer duration and potential for VR sickness, it was split into 2 equal parts where the user can continue from trial 3.1 to 3.2 seamlessly. Next, the participant is required to fill in an overall trust survey (A.3) that only concerns the last trial where both control and explanations are present. Finally, the participant is required to fill out a general post-experiment questionnaire (A.4) that targets all three trials. This questionnaire has open-ended questions which gives us the ability to assess how explanations and control mechanisms influence user trust in a more qualitative manner.

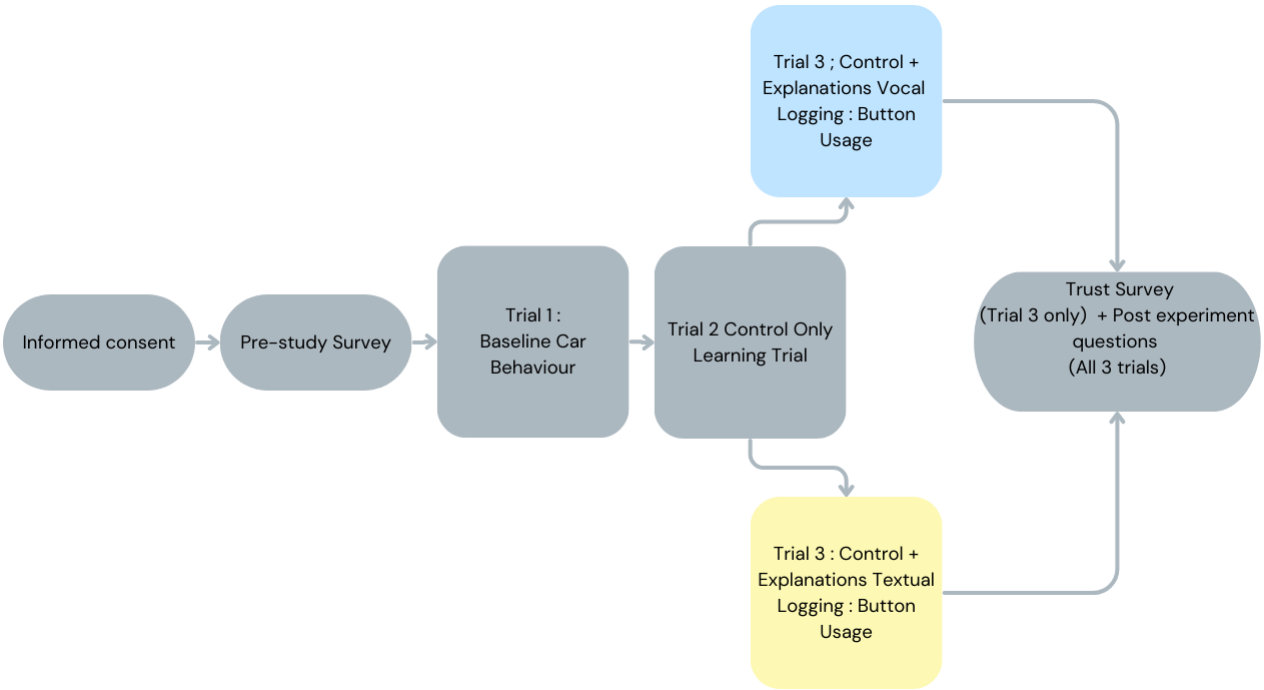


Figure 3.2: Experimental Procedure

3.3. Participants

We recruited N = 36 participants through personal connections and online advertisements, all of whom possessed a valid driver's license or a student driving license, and had normal vision and hearing. Participants were randomly assigned to either the Text Explanation group (n = 18) or the Vocal Explanation group (n = 18), ensuring equal group sizes. Within each group, participants experienced both control conditions as described. The sample included a mix of genders and ages (ranging from early 20s to mid 40s), reflecting a variety of driving backgrounds. Before the experiment, they were briefed about the general nature of a "virtual autonomous driving experience," though they were not told the specific hypotheses to avoid expectation bias. Participation is voluntary, and no financial or material compensation was provided. The study has received ethical approval from the Human Research Ethics Committee of Delft University of Technology. Important demographic and relevant background infor-

mation was also collected through a pre-study questionnaire (A.2). It included items on age, gender, propensity to trust AI, prior AV experience (e.g., whether they had been in an autonomous vehicle), and general driving experience (e.g., years of driving). This information was used to characterize the sample and to analyze whether factors like prior AV exposure or driving experience influenced trust in the simulation. Table 3.1 provides an overview of the participants characteristics.

	Vocal group		**Text group**	
	Amount	Percentage	Amount	Percentage
Gender				
Men	13	72%	11	61%
Women	5	28%	7	39%
Other	0	0%	0	0%
Age				
18–22 years old	5	28%	4	22%
22–26 years old	9	50%	10	56%
26–30 years old	1	6%	2	11%
30–35 years old	1	6%	2	11%
35–45 years old	2	11%	0	0%
Been in a car which was driverless				
Yes	6	33%	5	28%
No	12	67%	13	72%
Driving experience				
Not Experienced	1	6%	1	6%
Slightly Experienced	0	0%	4	22%
Experienced	3	17%	4	22%
Moderately Experienced	8	44%	5	28%
Very Experienced	6	33%	4	22%
Comfort with Autonomous Vehicles				
Strongly Disagree	1	6%	1	6%
Disagree	1	6%	2	11%
Slightly Disagree	4	22%	3	17%
Neutral	4	22%	5	28%
Slightly Agree	4	22%	2	11%
Agree	3	17%	5	28%
Mostly Agree	1	6%	0	0%

Table 3.1: Participant characteristics for Vocal and Text groups.

3.4. Materials

The experiment requires a combination of software tools and input mechanisms to simulate a Level 5 autonomous vehicle (AV) driving experience. The following materials are used:

3.4.1. Driving Simulator: CARLA with JOAN Framework

The experiment was conducted on a high-fidelity driving simulator in the Cognitive Robotics lab. The simulator consists of an office chair as the driver's seat, with a Varjo XR-3 headset providing a full field of view. We used the Carla Simulator (an open-source driving simulation environment) to create realistic urban and highway driving scenarios for the AV [8]. Carla is based on the Unreal Engine which is a powerful and versatile 3D computer graphics game engine developed by Epic Games. The autonomous driving behavior of the simulated vehicle was governed by a script that followed traffic rules and responded to other vehicles/pedestrians in a human-like manner. The vehicle **did not** adhere to Traffic lights and Speed limit signs. Importantly, the simulator was programmed to introduce a few pre-determined events (such as a sudden appearance of a car at an interesection, or an obstacle on the road) in each trial to test participant reactions and trust-related behaviors. These events were designed such that the AV's correct action might be non-obvious to a human driver (thus testing if the user would trust the AV's decision or try to intervene). The simulator also recorded telemetry data (location of other cars, speed, etc.), but our primary behavioral measures were the button presses of each participant. To structure the experiment within CARLA, the JOAN framework is used [29]. JOAN is a specialized experiment framework designed to facilitate human-in-the-loop studies in CARLA. It allows for controlled AV behavior, event scripting, and logging of participant interactions, making it an essential tool for conducting this study. The simulator runs on a Windows machine with an NVIDIA GPU, ensuring smooth rendering of the environment and minimal input lag.

The External Environment

Figure 3.3 shows the experimental setup used during the study. The participant is seated in a quiet, enclosed room to minimize distractions, wearing a Varjo XR-3 headset. The simulation is mirrored on a secondary screen for monitoring purposes. A standard keyboard is used to interact with the vehicle via left Ctrl and right Ctrl buttons. A steering wheel is also present, but was not used in this experiment. The environment was designed to ensure consistent conditions across sessions. On the desk the headset used for the Vocal explanations is also visible.

Figure 3.3: Participant seated in the lab using the VR-based driving simulator setup.

The Simulator Environment

Figure 3.4 shows a typical city scene where the AV encounters a partial obstruction and explains its behavior via an on-screen message. The speedometer indicates the car moving with 10 km/h. Participants experiencing the Voice condition would receive the same explanation through a headset.

Figure 3.4: City scenario with AV-generated explanation and speed display during navigation.

In contrast, the highway view (Figure 3.5) depicts an unobstructed environment with higher speeds (38 km/h), fewer dynamic elements, and a more predictable AV trajectory. Just below the speedometer, the interface highlights label mappings (e.g., “Slow down,” “Break,” “Stop”, ”Don’t Stop”) corresponding to the participant’s input buttons.

Figure 3.5: Highway driving scene showing speed and control label interface.

A wider perspective of the city layout is illustrated in Figure 3.6, showing parked vehicles, pedestrian

zones, and urban infrastructure. This context helped elicit decisions under ambiguity.

Figure 3.6: Wider city landscape used for urban navigation scenarios.

Finally, Figure 3.7 captures a complex scenario in which the AV detects a lane-blocking van. After stopping and displaying an “Obstacle Ahead” message, the AV waits for an approaching vehicle to pass, overtakes the van via the opposite lane, and then safely merges back. This was designed to test trust in nuanced AV decision-making.

Figure 3.7: Dynamic event with lane blockage and AV-generated explanation.

3.4.2. Control Mechanisms: Inform and Intervene Buttons

Participants can interact with the AV through two dedicated control mechanisms:

- Inform Button: Allows users to notify the AV that they perceive a situation requiring additional caution. Pressing this button signals the system to begin slowing down earlier or makes the

car continue with less speed but does not override its decision-making. The car resets to its standard behaviour after 7 seconds. The button can also be used to make the vehicle move from a stationary state, to a moving state of 15 km/h max. The car resets its standard behaviour after 7 seconds.

- Intervene Button: Enables participants to directly override the AV's behavior. If the car is stopping, pressing this button will force it to continue driving. Conversely, if the car is proceeding, pressing this button will force it to stop. The car resets its standard behaviour after 7 seconds.

Both buttons are mapped to keyboard inputs to standardize interactions across participants. Left control was set for the "Inform" and Right control for the "Intervene". The keyboard buttons had a distinction so the participant could easily distinguish them from the rest of the keys while wearing a virtual reality headset. See Table 3.2 which shows how each button press affects the vehicle depending on its state at the time. These interaction mechanisms allow participants to actively engage with the AV while assessing their trust and intervention tendencies.

Scenario	Inform (Left Ctrl)	Intervene (Right Ctrl)	Do Nothing
Car sees an obstacle which blocks the road	Car slows down faster than planned	Car ignores the obstacle and continues its predetermined course	Car follows the normal stop & go routine
Car detects a partially blocked road, decides not to stop, but continue slowly	Car slows down and continues at an even lower speed	Car stops completely, then resumes after a few seconds	Car follows its normal behavior (no stop, just continue)
Car is driving normally	Car temporarily slows down for a bit	Car stops completely and resumes by itself	Car continues driving normally

Table 3.2: Inform, Intervene, and Do Nothing responses for different scenarios.

3.4.3. Explanation Mechanisms: How and Why Messages

In Trial 3, the autonomous system provided explanations in one of two forms. For participants in the Text condition, a heads-up display on the simulator screen would show a short text message whenever the AV made a significant decision. The text was displayed in a subtitle-like overlay at the top of the screen for a few seconds. For participants in the Voice condition, the simulator played a spoken explanation through the car's audio system (using a pre-recorded human voice) conveying the same content. The content of the explanations was scripted to be identical in meaning across modalities, and was kept concise to avoid overloading the user with information. The rationale for testing voice vs text comes from interface design considerations in vehicles: auditory messages can allow the user to keep their eyes on the road, whereas text might require visual attention and reading. Although the prior studies we build on do not directly compare voice to text, the rationale comes from HMI theory: auditory messages may reduce visual distraction, while text could introduce cognitive demand. Our experiment will shed light on whether modality also affects the user's trust and intervention frequency. These explanations follow the How-Why framework, adapted from prior research on Explainable AI (XAI) in AVs [33, 26]. Although Why-only explanations are reported to induce the highest trust, we chose to use How + Why messages because they are shown to enhance safety most effectively. The explanations used are:

- Information that conveys only machine behavior (automation-centered communication). How-Only: "The car is stopping."
- Message that conveys only the situational reasoning for automation (context-centered communication). Why-Only: "Obstacle ahead"
- How + Why: "The car is stopping. Obstacle ahead".

The participant experiences 2 events of where it has to do something out of the ordinary in each trial except Trial 3 where they experience 5 events. One event is where the car slows down because of a warning but continues and avoids the obstacle, and the other scenario is one where the car has to stop completely because of an obstacle and then continue. For this three core messages were used, each

with slight variations to avoid repetition.

- Stopping completely: ”The car is stopping. Obstacle ahead”
- Continuing slower: ”Road slightly blocked, Slowing down”
- Slowing down on highway : ”Congestion on Lane, Slowing down”

3.4.4. Qualtrics

Qualtrics was used as both a data management tool and a survey platform throughout the study. All participant surveys, including the pre-study demographic and propensity questions, the post, Trial 3 trust survey, and the final open-ended interview questions—were created and administered using Qualtrics. After the driving trials, participants accessed the survey through a provided link and completed their responses digitally. The resulting data were downloaded via Qualtrics, facilitating efficient data organization and export for statistical analysis.

3.4.5. Calendly

Calendly was employed to coordinate the scheduling of in-person experimental sessions. Participants were provided with a link containing a brief description of the study and a calendar showing available appointment slots. Each session was allocated 45 minutes to allow sufficient time for instructions, the experiment itself, and the completion of surveys. Calendly also automated email reminders to participants, including details about the experiment’s location and timing.

3.5. Task

Each participant undertook three driving trials using a high-fidelity simulator developed in CARLA. The simulation was conducted in a customized virtual city, adapted from an existing CARLA map and modified to enhance realism. Two versions of the map were used to account for the two experimental groups: one with narrated and one with text-based explanations. The autonomous vehicle (AV) behavior was pre-scripted and trajectories were manually recorded by driving the route, saving positional (x, y, z), acceleration, and yaw data. These data were used in a custom steering and speed-following algorithm developed for this study, enabling the AV to navigate smoothly between trajectory points while adjusting speed to the context of the environment.

Throughout the simulation, dynamic traffic was present to create realistic driving conditions, including other vehicles moving alongside or in front of the AV. In city areas, the AV maintained speeds between 10 and 35 km/h, while highway segments allowed speeds up to 70 km/h. Participants were seated in the vehicle and given two keyboard-based control buttons, ”Inform” and ”Intervene,” used in the second and third trials.

In Trial 1, the AV operated fully autonomously without any participant control or explanations. The AV autonomously navigated around two events: a slow-down situation and a full blockade requiring a stop and overtake maneuver.

In Trial 2, participants could use the Inform and Intervene buttons to influence the AV’s behavior but did not receive explanations for the AV’s actions. Participants were instructed to experiment with the buttons frequently during this trial to become familiar with their function. Excessive button usage could lead to a minor collision, demonstrating the potential consequences of over-intervention.

In Trial 3, participants retained control through the buttons and additionally received ”How” and ”Why” explanations about the AV’s behavior. Participants were instructed to use the buttons at their discretion, pressing them only when they felt intervention was necessary. This final trial was longer and featured multiple blockade scenarios to allow for a comprehensive assessment of trust and intervention behavior. Participants received no information from the experiment designer during the final trial.

Button usage was logged exclusively during Trial 3 to capture behavioral data aligned with the primary experimental hypotheses.

3.6. Measurements

To evaluate the impact of explanation modality and control mechanisms, this study employs a combination of subjective and objective measurements.

- **Subjective measurements**: Gather self-reported trust levels using validated trust questionnaires at the end of the experiment. This includes pre-study variables and four open-ended questions covering all trials.
- **Objective measurements**: Track actual behavior through button-press frequency (intervene/inform usage) during experimental trials.

This mixed-method approach provides a robust picture of user trust, combining survey data with behavioral indicators. The trust questions were adapted from the Foundations for an Empirically Determined Scale of Trust in Automated Systems [17], complemented with constructs from the Situational Trust Scale for Automated Driving (STS-AD) [15]. Potential item order bias was mitigated following insights from Gutzwiller et al. [13].

3.6.1. Subjective Measurements

Self-Reported Trust: Trust in the AV was assessed through structured surveys administered at one key point, after Trial 3 (Final Trust Survey). This survey assessed the participant's overall trust in the AV after experiencing all conditions but participants were specifically tasked with rating the last trial. For example, participants rated statements like "The AV is reliable", "I can depend on the AV to make safe decisions", or "I felt wary of the AV's behavior" on a 7-point Likert scale from "Strongly Disagree" to "Strongly Agree" . Choosing "Strongly Disagree" would award 1 point, where as choosing "Strongly Agree" would award 7 points. Some items were phrased positively and others negatively to capture trust and distrust aspects. Items were scored (with negatively worded items reverse-coded; see Equation 3.2) and averaged into a composite trust score. A higher score corresponds to greater trust in the AV 3.1. In addition to the final post-trial survey, we also administered the Pre-Study Trust Survey before any driving began.

$$\bar{t} = \frac{1}{n} \sum_{i=1}^{n} q_i \tag{3.1}$$

Equation 3.1: Mean trust score.

$$q^* = \begin{cases} q, & \text{if positively worded} \\ 8 - q, & \text{if negatively worded} \end{cases} \tag{3.2}$$

Equation 3.2: Reverse coding for negative questions.

The pre-study trust score serves as a baseline for each participant which also helps us reason for any effect visible in the experiment (A.2). Finally, after Trial 3 and the final survey, we conducted a short semi-structured interview where participants could qualitatively describe their experience, explain their preferences (e.g., did they prefer having control or not, did they trust the explanations), and any other observations. These qualitative responses help interpret the quantitative data and provide illustrative quotes about the trust relationship.

3.6.2. Objective Measurements

Intervention Behavior: The primary objective measure of trust-related behavior was the usage of the Inform and Intervene buttons across trials. The frequency and circumstances of button presses were logged as indicators of the participant's willingness to take control. We interpret higher intervention rates as evidence of lower trust as the user did not feel comfortable letting the AV handle the situation. Conversely, if a participant rarely or never pressed the intervene or the inform button and chose to "do nothing" during critical moments, it suggests they trusted the AV's autonomy and accepted the risk of not intervening. Because intervention behavior can influence trip duration (e.g., frequent interventions prolong the trip), raw counts of button presses were preferred over rate-based measures to avoid confounding time and frequency. For each participant, the total number of presses was computed (look at 3.3), where I is the amount of informing and V the amount of intervening, and compared between the two modality groups (Text vs. Voice) .

$$T_i = I_i + V_i \tag{3.3}$$

Equation 3.3: Total amount of interventions.

Combining these measures, we have a rich dataset: subjective trust ratings (pre and post) and behavioral intervention data for each participant. In the next section, we describe how we analyzed these to test our hypotheses.

3.7. Data Analysis

The data analysis plan integrates statistical tests for the quantitative hypotheses (H1 and H2) and qualitative analysis for the exploratory hypotheses (H3 and H4), using appropriate tools to ensure thoroughness and reliability. All quantitative analyses will be performed using Python with relevant libraries (e.g., Pandas for data handling, SciPy/Statsmodels for statistics), which enables reproducible computation and visualization. Prior to hypothesis testing, the dataset will be screened for any data-entry errors or missing values and addressed accordingly. Basic descriptive statistics will be computed for all measures, and distributions of key variables (trust scores, button-press counts) will be visualized (e.g., via histograms or boxplots) to check for outliers and inform assumption checks.

3.7.1. Quantitative Analysis of Primary Outcomes (H1 & H2)

- To test H1 (Modality → Trust), an independent-samples t-test will be used to compare the Trust in Automated Systems survey scores between the two explanation modality groups (voice vs. text) after Trial 3. The t-test will determine whether the mean trust score of participants who received vocal explanations differs significantly from that of participants who received textual explanations. A higher mean in the vocal condition, if observed, would support H1's expectation that vocal explanations (which does not require reading and thus imposes lower visual-cognitive load) leads to greater user trust in the SAE Level 5 AV compared to text-based feedback.
- For H2 (Modality → Intervention Behavior), an independent-samples t-test will similarly be conducted to compare the total number of intervention button presses in Trial 3 between the voice and text explanation groups. The number of button presses (combining "Inform" and "Intervene" actions) during Trial 3 serves as an objective measure of the participant's intervention behavior. This test will assess whether participants with vocal explanations tend to intervene (press the buttons) less frequently than those with textual explanations, as H2 posits. A lower mean intervention count in the vocal explanation group would indicate that voice explanations reassure users and reduce unnecessary interventions (consistent with the hypothesis that they felt more confident in the AV's decisions).

While the hypotheses are directional in nature (H1 is assumed to have higher trust for the voice condition; H2 is predicted to have fewer interventions), we still use a two-tailed Welch's t-tests with a significance threshold of $\alpha = 0.05$, as its possible textual information to exarcerbate more trust. Assumptions of normality and homogeneity of variance will be checked for each comparison. Shapiro-Wilk tests (and normal Q–Q plots) will be used to verify whether trust scores and intervention counts are approximately normally distributed in each group, and Levene's test will examine the equality of variances between the voice and text conditions. If these assumptions are violated (for example, if the intervention counts are heavily skewed or variances are unequal), appropriate non-parametric alternatives will be employed. In particular, a Mann-Whitney U test would replace the t-test if normality is not met, and Welch's t-test adjustment would be applied if variances are unequal. Alongside p-values, we will report effect sizes to quantify the magnitude of any observed differences. For the t-tests, Cohen's d will be calculated (using pooled standard deviation) and interpreted using conventional benchmarks (≈0.2 for a small effect, 0.5 medium, 0.8 large). This provides additional insight into the practical significance of modality effects on trust and intervention behavior, beyond mere statistical significance.

3.7.2. Qualitative Analysis for Control Effects (H3 & H4)

Hypotheses H3 and H4 pertain to the influence of providing control buttons and explanations on user trust. Because the study's design did not include separate quantitative measures of trust for Trials 1 and 2 (trust was only formally surveyed after Trial 3), these hypotheses will be explored qualitatively through participants' reflections in post-experiment interview questions along with their quantitative data. At the end of the study, participants provided open-ended feedback about their experience across all three trials, including their sense of control, preferences for having (or not having) the inform/intervene buttons, the presence of explanations and which version of the car they preferred and why. These

narrative responses will be analyzed using a thematic analysis approach to glean insights on H3 and H4.

The qualitative analysis will involve transcribing the open-ended responses and then coding them to identify recurring themes related to control and trust. Key themes of interest include: perceptions of trust in the AV when no user control was available (Trial 1) versus when control buttons were introduced (Trials 2 and 3), and perceptions of trust in regards to the explanations between the trials. For H3 (Control Mechanism → Trust), we will examine participants' comments for indications that having the ability to intervene increased their confidence or comfort with the AV (e.g., statements expressing greater peace of mind or trust knowing they could take over if needed). For H4 (Explanations → Trust), we will look for any comments that indicate that having the explanations in general increased their trust towards the system. Four themes are created in total: Positive, Negative, Mixed and Neutral which are being decided based on two survey questions respectively. For H3 the questions that are taken into account are 2 and 4, and for H4 are questions 2 and 3 from the Post-study survey (Look at A.4). While our analysis used predefined sentiment combinations to structure responses into themes (Positive, Negative, Mixed, Neutral), the themes themselves were grounded in patterns that emerged during an initial review of participant feedback. Thus, our approach is best described as a hybrid thematic analysis: inductive in identifying response patterns, and deductive in organizing these patterns around our hypotheses (H3 and H4).

Each theme was derived by combining sentiment responses from specific questions as mentioned above:

- Button Theme (H3) used responses from Question 2 and Question 4.
- Explanation Theme (H4) used responses from Question 2 and Question 3.
- Sentiments between Q2 and Q4/Q3 are interchangeable

Q2 Sentiment	Q4/Q3 Sentiment	Resulting Theme
Positive	Positive	Positive
Positive	Negative	Mixed
Negative	Positive	Mixed
Negative	Negative	Negative
Neutral/Unclear	Neutral/Unclear	Neutral
Neutral	Positive	Positive
Neutral	Negative	Negative

Table 3.3: Thematic encoding rules based on sentiment combinations from Q2 and Q3/Q4. Sentiment between the questions is interchangeable

Two independent coders assigned thematic labels to each response using the scheme mentioned above (3.3). To reconcile differences and create a single final label per response, a rule-based label fusion function was applied. This function respected coder agreement where present, and applied consistent rules for disagreement. Key fusion rules included:

- Identical Labels kept as a final label
- Conflicting positive and negative was kept as Mixed
- Positive with neutral or mixed would always be mapped to the latter.
- Negative with neutral or mixed would always be mapped to negative
- Mixed and neutral would be mapped to neutral

This ensured a reproducible, transparent, and non-arbitrary consolidation of qualitative codes. The result was one fused theme label per participant for both control (button) and explanation perceptions.

To complement the qualitative coding, the following quantitative analyses were performed:

- Independent-samples **t-tests** were conducted to compare the trust scores and intervention counts between participants who expressed *positive* views and those who expressed *non-positive* views (i.e., combined Mixed, Neutral, and Negative categories).
- **Cohen's *d*** was calculated to estimate the effect size of these differences, providing insight into the practical magnitude of any observed effects.

This combined approach enables a more robust analysis of H3 and H4 by integrating subjective perceptions with statistical comparisons. While small subgroup sizes, particularly for negative and neutral categories, limit generalizability, this method provides a more interpretable and consistent framework for evaluating H3 and H4 alongside the primary hypotheses (H1 and H2). The qualitative component supplements these findings by offering illustrative quotes and thematic insights that contextualize the observed statistical patterns. For example, explaining why participants in certain conditions felt more or less compelled to intervene.

Exploratory Analysis of Individual Differences

Beyond the primary and secondary hypotheses, the study includes an exploratory analysis to investigate whether individual differences among participants might correlate with trust or intervention behavior. Several background variables were collected in the pre-study survey that could plausibly affect how they interact with the autonomous vehicle. These include demographic factors and prior experiences or attitudes, specifically:

- Age: The participant's age (in years). We will examine if age is related to trust in the AV or willingness to intervene (e.g., younger vs. older drivers' trust levels).
- Gender: Participants' gender identity. Although our sample size may limit statistical power, any differences in trust or intervention frequency between genders will be explored.
- Driving Experience: Measures of driving history, such as years of driving or self-reported driving skill and risk aversion. More experienced drivers might be more critical of the AV's decisions (potentially intervening more) or, alternatively, may trust the system if they find it behaves like a proficient driver.
- Comfort with Automation: The general comfort level with automated technologies (e.g., responses to survey items about using autopilot features or other AI systems). Higher comfort or tech-savviness could be associated with greater trust in the AV and fewer interventions.
- Prior AV Exposure: Any prior experience with autonomous vehicles (for example, having ridden in a self-driving car or used Level 4 autonomy in personal driving). Familiarity with AVs might lead to higher initial trust and potentially less frequent interventions, as users know what to expect from the system.

The relationships between each of these individual difference variables and the main outcomes (trust score after Trial 3, and intervention button count in Trial 3) will be examined using correlation analyses. If the data are approximately normally distributed and linear, Pearson's correlation will be used to compute correlation coefficients between, for example, age and trust score, or comfort-with-automation rating and number of interventions. For variables that are ordinal or not normally distributed (which may be the case for some survey ratings or skewed counts), Spearman's rank-order correlation will be employed instead. We will apply these tests in an exploratory manner (without strict a priori hypotheses), so any significant findings will be interpreted with caution.

3.7.3. Correlation Between Trust and Intervention

To investigate the relationship between the subjective trust and their objective intervention behavior, a correlation analysis between final trust scores and total button presses (inform + intervene) during Trial 3 will be conducted. Shapiro–Wilk tests will confirm if both variables were approximately normally distributed within each explanation modality group (text and voice), allowing us to use Pearson correlation coefficients. To assess whether the strength of this relationship differed by modality, a Fisher z-test will also be conducted on the correlation coefficients from each group.

In summary, the data analysis combines rigorous quantitative tests for the primary modality effects on trust and intervention behavior with qualitative and exploratory methods to capture the influence

of control mechanisms and individual differences. This mixed-methods approach ensures that the hypotheses H1 and H2 are tested with appropriate statistical power and assumptions checks (reporting any significant differences along with their effect sizes), while H3 and H4 are explored through rich qualitative insights. By triangulating quantitative results with interview-derived explanations and considering personal factors, the analysis will provide a comprehensive understanding of how explanation modality and user control features jointly shape user trust in, and interaction with, a fully autonomous vehicle.

4

Results

This chapter presents the results of the study in alignment with the predefined hypotheses. Quantitative results are reported first, targeting H1 and H2. Subsequently, qualitative insights from post-experiment interviews along with the quantitative analysis is reported in regards to H3 and H4. Finally, exploratory correlations with individual differences are presented.

4.1. Participants

Before conducting any analysis, it is important to verify that pre-study variables do not systematically differ between groups or influence the results. It is important to see if any of the pre-study variables are significantly different when comparing the two groups since the same type of participant is needed in each group.

The participants were divided evenly between the voice and text explanation groups. Statistical tests revealed no significant group differences across any of the measured variables. Chi-square tests showed no group differences in gender or prior AV experience. Mann–Whitney U tests confirmed similar distributions for age, driving experience, and comfort with automation. While minor variations existed (e.g., slightly more males in each group, or non-normal age distributions), none reached statistical significance. Thus, the participant groups were considered demographically and attitudinally comparable, minimizing the risk of confounding effects in subsequent analyses.

Detailed distributions and test results can be found in Appendix A.5.

4.2. Quantitative Results: Effects of Explanation Modality on Trust and Intervention (H1, H2)

4.2.1. Trust Scores Across Modalities (H1)

An independent-samples t-test was conducted to compare the composite trust scores (Trial 3) between the voice and text explanation groups.

The Shapiro–Wilk test confirmed that trust scores were approximately normally distributed for both the voice group ($p = 0.112$) and the text group ($p = 0.243$). Levene's test indicated no significant difference in variances ($p = 0.118$), supporting the assumption of homogeneity of variance.

There was no significant difference in trust between the voice and text groups ($t = 0.73$, $p = 0.469$). The effect size (Cohen's $d = 0.24$) indicates a small, non-significant difference, with slightly higher trust scores in the voice group.

Figure 4.1 presents a violin plot with embedded boxplots and individual data points (strip plot) comparing trust scores between the text and voice explanation groups. The distribution of trust scores appears slightly higher in the voice group, though there is notable overlap between the two groups. The boxplots indicate comparable medians and interquartile ranges, while the wider spread in the text group suggests

greater variability in trust responses among those participants.

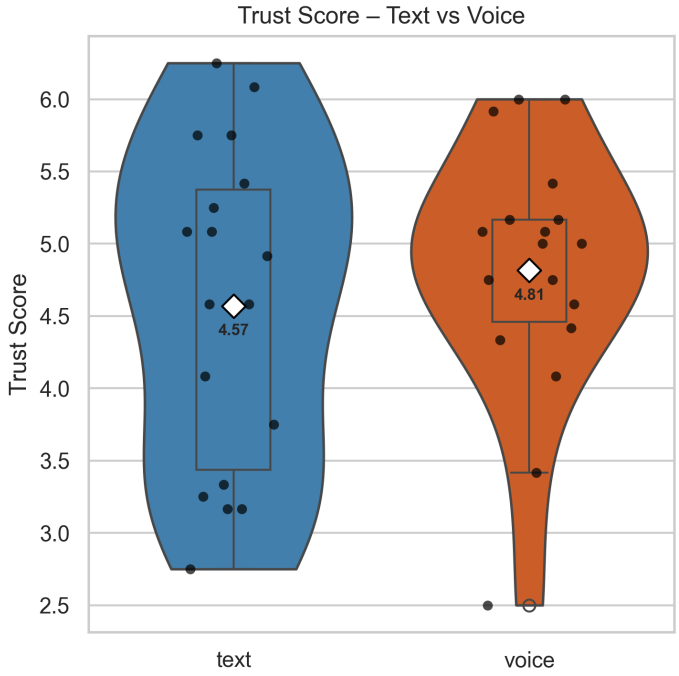


Figure 4.1: Distribution of trust scores by explanation modality (text vs. voice). Violin plots show the full distribution, boxplots indicate the interquartile range and median, and individual dots represent participant scores. Mean trust scores are annotated for each group.

To further quantify the group difference, a Gardner–Altman plot (Figure 4.2) was generated. The vertical thickness of the violin shows the kernel-density estimate of all bootstrap mean-differences, how often a given mean-difference appeared during bootstrapping. The observed mean difference in trust scores between the voice and text groups was 0.25 points in favor of the voice condition, with a 95% confidence interval (CI) ranging from -0.42 to 0.88. The bootstrap sampling distribution, shown as the grey violin, indicated that the most frequent resampled differences clustered slightly above zero. However, the 95% CI included zero, confirming that the difference was not statistically significant. This result is consistent with the previously reported small effect size (Cohen's $d = 0.24$). Interpretation: Average trust in the voice condition was 0.25 points higher than in the text condition; however, this difference was not statistically significant.

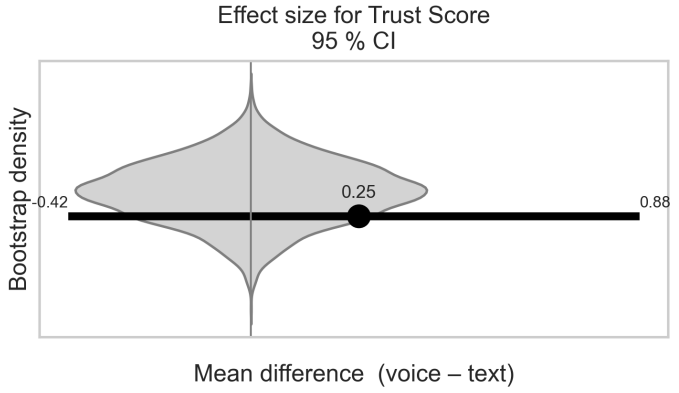


Figure 4.2: Gardner–Altman plot showing the mean trust score difference between explanation modalities, with 95% confidence interval.

4.2.2. Button Usage Across Modalities (H2)

An independent-samples t-test was also conducted to compare the total number of button presses (combining Inform and Intervene actions) between the voice and text groups during Trial 3.

Normality checks using the Shapiro–Wilk test confirmed acceptable distribution for both groups (voice: $p = 0.149$ text: $p = 0.118$). Levene's test found no significant variance difference ($p = 0.663$).

There was no significant difference in button presses between the voice and text groups ($t = 0.65$, $p = 0.519$). Cohen's $d = 0.22$ suggests a small, non-significant effect, with slightly fewer interventions in the text group on average.

Figure 4.3 shows the distribution of total intervention button presses (inform + intervene) during Trial 3, again using a violin plot overlaid with boxplots and individual participant data points. The distributions for both groups demonstrated considerable spread and overlap.

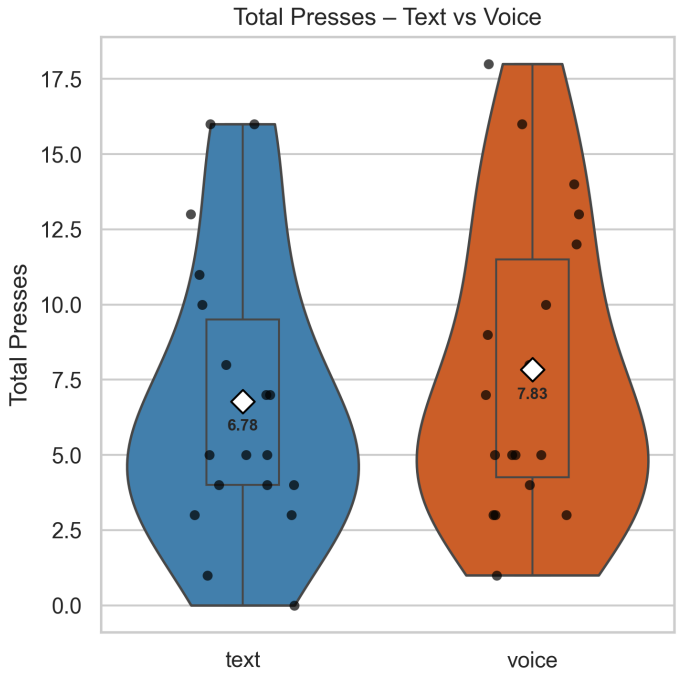


Figure 4.3: Distribution of total button presses by explanation modality (text vs. voice). Violin plots depict the full distribution of press counts, boxplots show the interquartile range and median, and individual dots represent participant data. Mean total presses are annotated for each group.

The Gardner–Altman plot in Figure 4.4 provides the effect size analysis for intervention behavior. The vertical thickness of the violin shows the kernel-density estimate of all bootstrap mean-differences, how often a given mean-difference appeared during bootstrapping. The mean difference in total button presses between the groups was 1.06 presses (voice minus text), with a 95% confidence interval from -2.06 to 4.17. As with the trust score analysis, the confidence interval crossed zero, suggesting no statistically significant difference in intervention frequency between the two groups. This is consistent with the small observed effect size (Cohen's $d = 0.22$) and the descriptive statistics showing only a modest trend toward more interventions in the voice condition. Interpretation: Participants in the voice group pressed the intervention buttons 1.06 times more on average than those in the text group; however, this difference was not statistically significant.

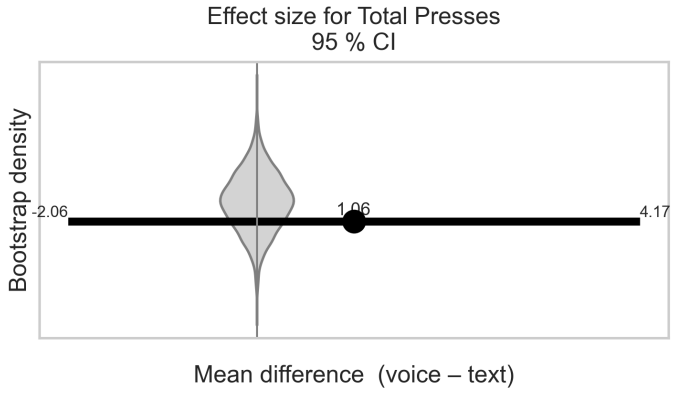


Figure 4.4: Gardner–Altman plot showing the mean difference in intervention button presses between explanation modalities, with 95% confidence interval.

Although the hypothesized benefits of vocal explanations (greater trust and reduced intervention) were not statistically supported, the observed trends align directionally with H1. These results suggest the need for further investigation using larger sample sizes or alternative study designs.

4.3. Qualitative Insights: Controls and Explanations with Trust (H3, H4)

To explore how participants' subjective experiences with control mechanisms (buttons) and explanations aligned with their trust and behavior, thematic analysis was conducted on open-ended responses from the post-experiment survey. Responses were coded into four categories: positive, negative, neutral, and mixed, separately for themes related to control mechanisms (H3) and explanations (H4). This coding captured whether participants perceived these system features as trust-enhancing, problematic, or ambivalent.

4.3.1. H3 : Control Mechanisms (Buttons) and Trust

Thematic coding of participants' views on the control buttons yielded the following distribution: 17 positive, 9 mixed, 6 negative, and 4 neutral. Participants who expressed positive attitudes towards the buttons reported a higher mean trust score ($M = 5.03, SD = 0.81$) compared to those with mixed, neutral, or negative views combined ($M = 4.35, SD = 1.06$), see Figure 4.5. An independent-samples t-test confirmed that this difference was statistically significant ($t = 2.06$, $p = 0.047$) with a medium-to-large effect size ($d = 0.678$), supporting H3.

In terms of behavior, those with positive button perceptions exhibited a lower mean number of interventions ($M = 6.18, SD = 4.39$) compared to others ($M = 8.52, SD = 4.83$), although this difference was not statistically significant ($t = -1.36, p = 0.184$). However, the effect size was moderate ($d = -0.449$) indicating a potentially meaningful reduction in interventions associated with positive control perceptions even if not statistically confirmed in this sample.

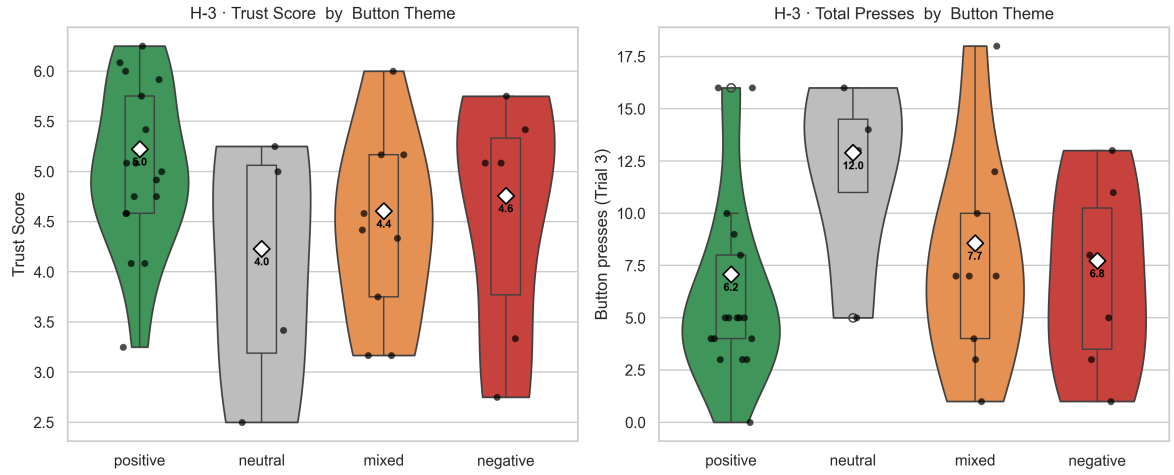


Figure 4.5: Trust scores and total intervention counts across button perception themes. Participants who expressed positive views of the control mechanisms (buttons) generally reported higher trust and exhibited fewer interventions compared to participants with mixed, negative, or neutral perceptions. Individual data points are overlaid.

4.3.2. H4 : Explanations and Trust

For the explanations theme, most participants (26) provided positive evaluations, with fewer expressing mixed (6), neutral (3), or negative (1) views. Participants with positive explanation perceptions reported higher trust scores ($M = 4.88, SD = 0.89$) than others ($M = 3.77, SD = 1.01$), though this difference was not statistically significant ($t = 1.64, p = 0.124$), see Figure 4.6. The effect size was moderately large ($d = 0.682$) indicating a potentially meaningful behavioral difference between explanation perceptions and trust.

In terms of intervention behavior, participants with positive explanation perceptions had fewer button presses ($M = 6.46, SD = 4.32$) than those with other views ($M = 8.5, SD = 3.7$). While this difference also did not reach statistical significance ($t = 1.56, p = 0.143$), the effect size was moderately large ($d = -0.649$) suggesting a potentially meaningful behavioral difference that warrants further investigation with larger samples.

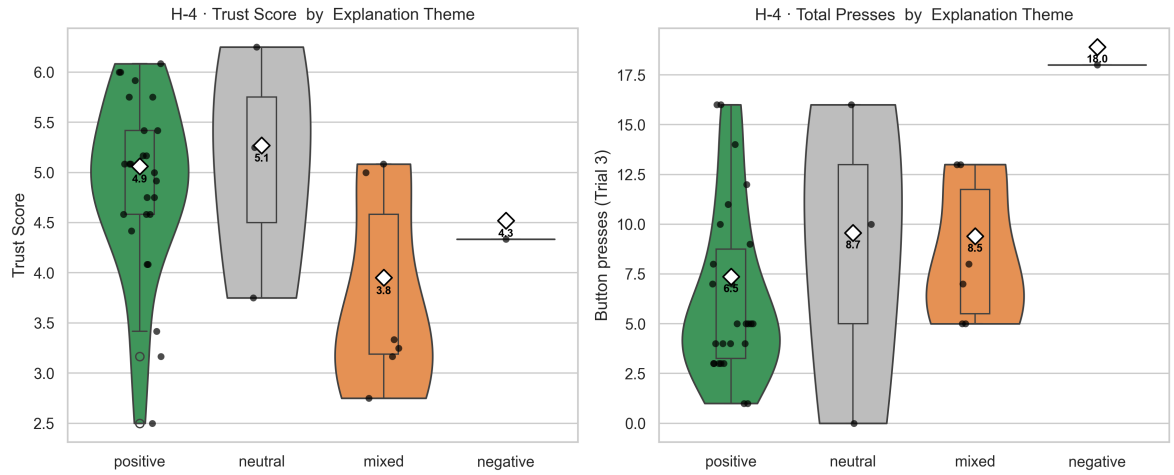


Figure 4.6: Trust scores and total intervention counts across explanation perception themes. Participants with positive evaluations of the explanations reported slightly higher trust scores and lower intervention rates than other groups, although these differences were not statistically significant. Individual data points are overlaid.

A cross-tabulation further supported this trend: 22 out of 26 participants (84.6%) who gave positive evaluations of the explanations also selected Version 3 (Control + Explanations) as their most preferred

driving condition (Figure 4.7). Only four participants with positive views did not choose Version 3, possibly due to unrelated individual preferences or situational factors. This alignment between subjective perception and condition preference strengthens the interpretation that explanations can enhance user trust and satisfaction.

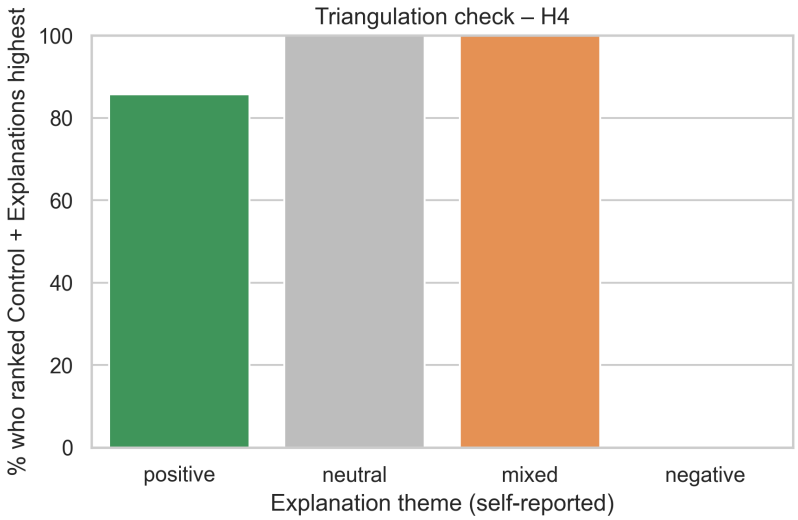


Figure 4.7: Percentage of participants ranking Version 3 (Control + Explanations) as their most preferred condition, split by explanation perception themes. Notably, 84.6% (22 out of 26) of participants who expressed positive views about the explanations also ranked Version 3 highest, suggesting alignment between subjective perception and overall user preference.

4.3.3. Summary

The qualitative themes triangulate well with the quantitative data for H3, providing converging evidence that perceiving the control mechanisms positively is associated with higher trust and potentially less intervention behavior. For H4, while the trends were consistent with the hypothesis that positive perceptions of explanations relate to higher trust and fewer interventions, the statistical comparisons did not reach significance, likely reflecting the limited sample sizes for non-positive themes.

4.4. Exploratory Analyses

Pearson or Spearman correlations were computed between participants' pre-study variables and their trust scores and button usage. These exploratory analyses aimed to generate additional insights rather than formally test hypotheses.

4.4.1. Ranking Preferences Across Versions

Participants ranked the three driving conditions based on comfort. The majority (32 out of 36 participants; 88.88%) ranked Version 3 (Control + Explanations) as the experience with the least discomfort (Rank 3), followed by Version 2 (Control only) and Version 1 (Baseline). Version 1 was rarely chosen as the preferred option. Rank 1 points to most discomfort while Rank 3 points to least discomfort

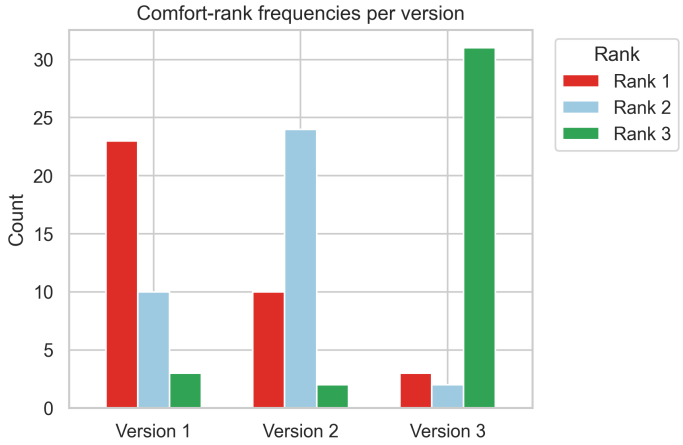


Figure 4.8: Participant ranking frequencies for each driving version. The majority ranked Version 3 (Control + Explanations) as the experience with the least discomfort (Rank 3). Versions offering user control (Versions 2 and 3) were consistently ranked higher than the baseline condition (Version 1).

A pairwise comparison of ranking choices revealed consistent preferences:

- **Version 3** was preferred over Version 1 by 32 participants and over Version 2 by 32 participants.
- **Version 2** was preferred over Version 1 by 24 participants.
- **Version 1** was preffered over Version 2 by 12 participants

While most participants favored systems offering both control and explanations, a notable subset (n = 12) preferred Version 1 (no control, no explanations) over Version 2 (control only), indicating that for some, control alone was less appealing than a minimalist baseline experience.

4.4.2. Correlations with Individual Differences

To explore potential influences of individual differences, correlations were computed between participants' pre-study variables and two key outcomes: trust scores and intervention behavior (total button presses). The analysis automatically applied Pearson or Spearman correlation methods depending on variable normality, as verified using Shapiro-Wilk tests.

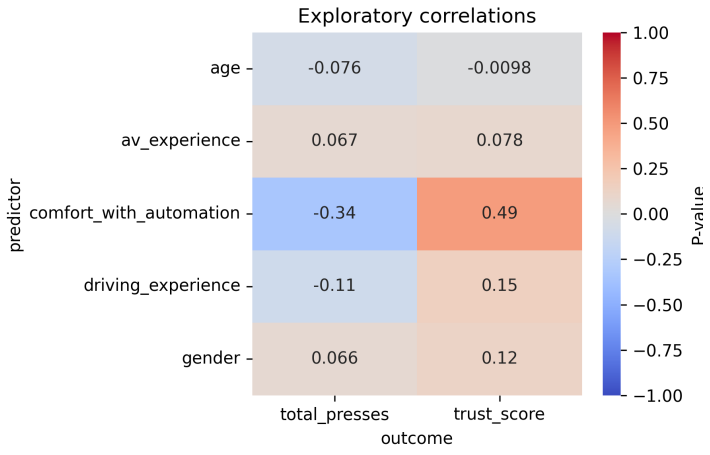


Figure 4.9: Correlation matrix showing relationships between pre-study individual difference variables and key outcomes (trust scores and total intervention presses). Statistically significant correlations were observed between comfort with automation and both trust scores (positive correlation) and total presses (negative correlation), suggesting that participants more comfortable with automation reported higher trust and intervened less frequently.

Two statistically significant relationships emerged:

- **Trust scores** were positively correlated with participants' comfort with automation ($r = 0.48, p = 0.002$), suggesting that individuals that reported that they are already comfortable with automated systems tend to report higher trust in the AV.
- **Total intervention presses** were negatively correlated with comfort with automation ($r = -0.35, p = 0.043$), indicating that participants more comfortable with automation intervened less frequently.

Other individual differences (e.g., age, gender, driving experience, AV exposure) did not show significant correlations with trust or intervention behavior. These findings highlight the role of the self reported automation comfort in shaping both subjective and behavioral responses, aligning with prior research on trust calibration.

4.4.3. Correlation Between Trust and Intervention Behavior

We explored whether trust scores were associated with actual intervention behavior.

- **Text Group** : $r = -0.283, p = 0.254$
- **Voice Group**: $r = -0.281, p = 0.258$

Both groups showed a weak negative correlation, suggesting that higher trust was associated with slightly fewer interventions, though neither reached statistical significance.

To assess whether the strength of these correlations differed significantly between groups, a Fisher z-test was performed. The result showed no significant difference in correlation strength: $z = 0.01, p = 0.995$. Thus, the relationship between trust and intervention appears consistent across both explanation modalities. Figure 4.10 displays a scatterplot of trust scores versus total button presses with linear regression lines for each group.

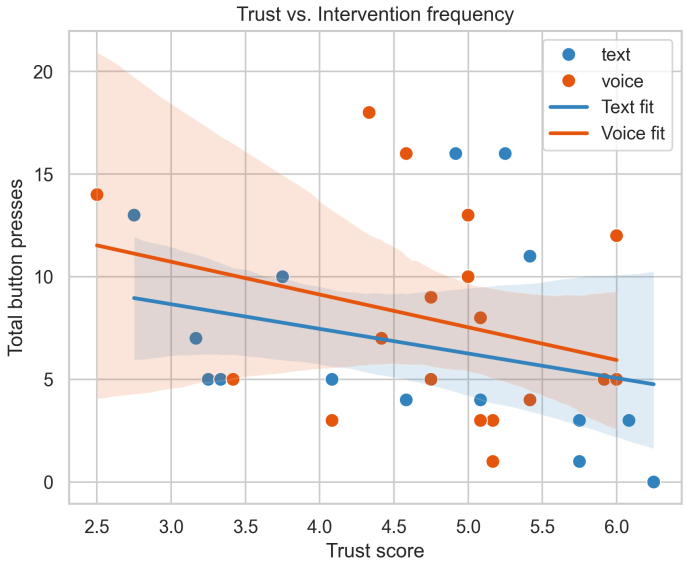


Figure 4.10: Correlation between trust scores and total button presses during Trial 3, separated by modality. While both groups show a slight negative trend, neither correlation is statistically significant.

4.5. Result Summary

Hypothesis	Descr.	Test Used	Result	p-value	Supported	Cohen's d
H1	Vocal explanations increase trust more than text explanations.	Welch two-tailed t-test	t = 0.73	0.469	No	0.244
H2	Vocal explanations reduce intervention behavior compared to text.	Welch two-tailed t-test	t = 0.65	0.519	No	0.217
H3	Positively viewing the presence of control buttons increases trust.	Welch two-tailed t-test	t = 2.06	0.047	Yes	0.678
H3	Positively viewing the presence of control buttons decreases interventions.	Welch two-tailed t-test	t = -1.36	0.184	No	-0.449
H4	Positively viewing the presence of explanations increases trust.	Welch two-tailed t-test	t = 1.64	0.124	No	0.682
H4	Positively viewing the presence of explanations decreases interventions.	Welch two-tailed t-test	t = -1.56	0.143	No	-0.649

Table 4.1: Summary of hypothesis testing results and effect sizes.

5

Discussion

In this chapter, we discuss the results of the research conducted in this thesis. We do this by revisiting the research questions and interpreting the results of the previous chapter. This chapter ends with our limitations and future work

5.1. Results

5.1.1. Modality and Trust

The primary hypothesis (H1) proposed that vocal explanations would lead to higher trust compared to text-based explanations, based on the premise that vocal communication may lower cognitive effort and foster greater understanding in automated driving contexts [35, 20]. However, the results did not yield a statistically significant difference between the two modalities . As shown in the violin plot (Figure 4.1) and further quantified through the Gardner–Altman plot (Figure 4.2), participants receiving vocal explanations reported slightly higher trust scores (mean difference = 0.25), but the 95% confidence interval (−0.42 to 0.88) included zero. This suggests that the observed difference, while directionally consistent with H1, could plausibly be due to chance.

There are several factors that may explain this non-significant finding. First, prior studies highlight that explanation modality can influence trust, but also that user preferences and cognitive workload vary significantly across individuals [35, 20] . In controlled simulation, comprising short driving scenarios with only 5 intervention relevant events, both modalities may have provided sufficient transparency, minimizing the potential benefits of vocal over text explanations and the cognitive demands may not have exceeded the threshold where modality differences become salient

Second, exploratory analyses revealed that individual factors, such as participants' comfort with automation, were strongly correlated with trust ($r = 0.48, p = 0.002$), independent of modality (Figure 4.9). This finding echoes that dispositional trust and personal attitudes often outweigh system characteristics in shaping automation trust [27]. It also aligns with earlier findings that trust in automation is influenced not only by objective system attributes but also by subjective user experience and perceived control [33, 22]. Finally, the study's sample size (N = 36) may have limited the power to detect modest effects where a total of 64 participants would ideally be present.

In summary, while vocal explanations showed a slight advantage in fostering trust, the evidence does not support a statistically significant modality effect. Instead, trust in this study appears more strongly shaped by participants' general attitudes toward automation and the overall presence of explanations and control mechanisms as further discussed in later sections.

5.1.2. Modality and Intervention

The second hypothesis examined whether explanation modality influenced participants' intervention behavior, measured through total button presses. Although the data showed a modest trend toward slightly more interventions in the voice condition compared to the text condition, this difference was

not statistically significant (see Figure 4.3 and Figure 4.4). The effect size (Cohen's d = 0.22) and the Gardner–Altman plot indicated that, while the mean difference favored the text modality (mean difference = 1.06 presses), the confidence interval crossed zero (95% CI = -2.06 to 4.17), reinforcing the lack of statistical significance.

Several factors may explain this non-significant outcome. First, while prior studies suggest that vocal explanations may lower cognitive load and enhance situational awareness [20], such advantages may not consistently translate into reduced intervention behavior. In this study, variability in participants' comprehension and usage of the intervention buttons likely moderated the influence of modality. As one participant noted, *"The multiple functions made them confusing to use, but it was nice to have some form of control"*, illustrating mixed perceptions of the buttons' utility and clarity.

Second, the simulation environment itself may have influenced intervention rates. Despite instructions clarifying that traffic lights and speed limits were not enforced, participants' real-world driving habits could have conflicted with the AV's actions. Notably, the vehicle lacked signaling (e.g., turn indicators), which may have contributed to uncertainty or perceived errors, prompting interventions unrelated to explanation modality. The fact that the experiment also takes place into a simulation, also affects the participants ability to perform the tasks as if they are in the real world. The participant could easily adopt the environment as being in a game which makes reasons to interact with the vehicle completely different as to being in a real, complex scenario.

Finally, similarly to H1, individual differences and the relatively small sample size could have limited statistical power. Even though exploratory analyses suggested that comfort with automation negatively correlated with total button presses, variability in how participants interpreted and used the buttons added noise to the outcome measure.

In conclusion, the results did not support H2: explanation modality did not significantly influence intervention behavior. Instead, intervention rates appeared more influenced by participants' understanding of control mechanisms, their subjective experience within the simulation, and individual comfort with automation.

5.1.3. Controls with Trust and Total Interventions

The third hypothesis proposed that the availability of control mechanisms, specifically the inform and intervene buttons, would increase trust and reduce the frequency of user interventions. This hypothesis was supported by the data, see Section 4, yielding a significant result. Participants who expressed positive attitudes towards the buttons reported a higher mean trust score compared to those with mixed, neutral, or negative views combined (See Figure 4.5).

In terms of behavior, those with positive button perceptions exhibited a lower mean number of interventions, although this difference did not reach statistical significance ($p = 0.184$). However, the effect size was $d = -0.449$, which suggests a moderate difference in the expected direction: participants with positive button perceptions intervened less often than others. This pattern nonetheless suggests that perceiving the control mechanisms positively may reduce the tendency to intervene, aligning with higher perceived trust.

The importance of the control buttons in shaping participants' experiences was further reinforced by open-ended feedback. In the responses (see Section A.4), the majority of participants expressed positive or mixed opinions. Many emphasized that simply having the option to inform or override the AV enhanced their sense of safety and trust. As one participant stated: *"The fact that I had the option to intervene gave me peace of mind, even though I hardly used it."* This reflects the broader phenomenon that the perception of control, even if rarely exercised, can be sufficient to bolster trust [22, 10, 39].

These findings also invite reflection on what constitutes a "wrong" intervention. In the present study an intervention was tagged as 'unnecessary' or 'wrong' if it deviated from the AV's planned trajectory without preventing a collision. Yet that engineering-centric definition excludes subjective factors such as motion comfort, perceived courtesy toward other road users, or simply the passenger's desire for a larger headway buffer. Interviews revealed clear individual differences: some participants felt any override that avoided "hard braking at the last second" was entirely justified, whereas others considered the same action trigger-happy. Thus, labelling an intervention as wrong is partly normative and partly

personal. A promising design implication is to give passengers adjustable “driving styles” or tunable safety margins, analogous to adaptive cruise-control distance settings, so that the AV’s behaviour can converge toward each user’s comfort zone, reducing the likelihood that they will feel compelled to intervene.

Additional feedback, particularly from the open-ended questions, confirmed that the availability of buttons influenced participants’ overall comfort and trust. Many who ranked versions with control mechanisms higher mentioned the reassurance and agency the buttons provided. For example: *”I trusted the system more when I knew I could inform it about issues or take action if necessary.”*

However, some participants expressed uncertainty about the functionality of the buttons, including doubts about when to use them or whether their input had a real effect. This was evident in neutral or mixed responses. One participant commented: *”I was sometimes unsure if pressing a button would really change anything.”* Such uncertainty may have introduced variability into the intervention behavior, potentially dampening the statistical strength of the behavioral findings.

Additionally, the simulation environment may have contributed to this variability. As mentioned before, the absence of traffic lights, speed limits, and turn signals, despite participants being informed about these omissions, likely added ambiguity to their decision-making. Some participants may have been unsure when an intervention was warranted or may have perceived system behavior as erroneous when it was not. These factors could explain both the higher-than-expected intervention rates and the individual differences in how the control buttons were used or perceived.

Despite these limitations, the findings strongly indicate that providing drivers with control mechanisms, especially those allowing non-invasive input, such as inform buttons, can effectively enhance trust and support appropriate calibration of intervention behavior.

5.1.4. Explanations with Trust and Total Interventions

The fourth hypothesis proposed that the presence of explanations, delivered either as vocal or text messages, would increase trust and reduce intervention behavior, as participants would better understand the AV’s decisions.

While the thematic coding revealed a consistent trend, the hypothesis was not statistically supported. Participants who expressed positive views on the explanations exhibited a higher mean trust score than those with mixed, negative, or neutral opinions (See Figure 4.6), but this difference did not reach statistical significance ($p = 0.124$). The effect size was mediun to large, d = 0.682, indicating a modest increase in trust among participants with positive perceptions of explanations.

Similarly, participants with positive perceptions of explanations pressed the intervention buttons less frequently compared to others (See Figure 4.6), yet this difference was also non-significant (p = 0.143). Still, the effect size was d = −0.649, a moderately large negative effect, meaning that participants with positive views of explanations intervened noticeably less than others. The pattern nonetheless suggests that favorable attitudes towards explanations may be associated with both higher trust and lower intervention behavior.

One important consideration is the limited sample size of participants with negative or neutral attitudes, which likely reduced statistical power and contributed to the non-significant results despite moderate effect sizes. This imbalance, combined with high variability in subjective responses, makes it difficult to draw firm conclusions from between-group comparisons.

Qualitative feedback provided valuable context for these findings. In the open-ended responses, many participants reported that explanations, particularly those offering “why” rationales, helped them make sense of the AV’s behavior and increased their trust: *”The voice explanations gave me a better understanding of what the car was doing and why.”*. Another participant stated: *”Knowing why the car stopped or slowed down made me feel more in control and more confident in the system.”*. This aligns with prior research highlighting the importance of intentional explanations (explaining ”why” actions are taken) for building trust and mental models in automated systems [30, 42].

Several participants also expressed that vocal explanations were especially effective in maintaining their attention and reducing uncertainty: *”I preferred the voice messages because they felt more nat-*

ural and kept me engaged without distraction." However, some participants reported confusion or dissatisfaction with the explanations, particularly when they found them too sparse or lacking sufficient detail: *"Sometimes the explanations felt too generic, and I wasn't sure what the car was reacting to."*. Others noted that explanations occasionally lagged behind events or failed to mention specific behaviors, leading to uncertainty or skepticism about their reliability. This variability in explanation clarity and timing may have contributed to the non-significant quantitative results, despite the observed trends. Additionally, several environmental limitations as mentioned above likely introduced inconsistencies in how participants interpreted the explanations, especially for complex driving scenarios where more context would have been beneficial.

Despite these challenges, the overall pattern suggests that participants who perceived explanations positively tended to trust the AV more and intervened less frequently, consistent with previous findings that well-designed explanations enhance user understanding and trust [30, 20, 32]. While statistical significance was not achieved, the qualitative data supports the continued development of explanation strategies as a key component of trust calibration in automated driving.

5.1.5. Exploratory Findings

Beyond testing the predefined hypotheses, several exploratory analyses offered additional insights into the factors influencing trust and intervention behavior.

First, correlational analyses revealed meaningful relationships between participants' pre-study self-reported variables and their behavior and trust in the main trials. Notably, a positive correlation emerged between participants' comfort with automation and their trust scores in Trial 3 ($r = 0.48$, $p = 0.002$). Similarly, comfort with automation was negatively correlated with total intervention presses ($r = -0.35$, $p = 0.043$), see Figure 4.9. This suggests that participants who reported higher comfort with automated systems tended to trust the AV more and felt less need to intervene.

By contrast, age, gender, general driving experience, and prior AV familiarity did not show statistically significant correlations with either trust scores or intervention frequency. This suggests that situational trust development during the experiment may have played a more influential role than dispositional traits. However, the self-reported nature of the pre-study variables should be considered when interpreting these results. Participants' subjective assessments may not fully reflect their in-simulator behavior or attitudes, especially under novel or unexpected driving scenarios.

An analysis of participants' ranking preferences for the different system versions (Control Only, Explanations + Control, and Baseline) provided additional insights into how the combination of explanations and control mechanisms influenced perceived trust and usability.

While the Explanation + Control version was most frequently ranked highest, a noteworthy subset of 12 participants ranked the Baseline version (no explanations or controls) above the Control Only version. Their open-ended responses often cited a clear rationale: *"I feel more stressed when i have control but no explanations, because I feel like I HAVE to intervene. No control and explanations therefore feels safer. "*

This reasoning highlights an important nuance: control mechanisms alone, without sufficient explanatory context, can introduce cognitive uncertainty or anxiety rather than empowering the user. This aligns with prior findings suggesting that perceived control must be both meaningful and understandable to positively affect trust [22, 32, 39]. Participants seemed to prefer a lack of control over ambiguous or unexplained control opportunities.

Trust - Intevention Correlation

We examined the correlation between trust scores and total intervention presses during Trial 3. Although both the text and voice groups exhibited a weak negative correlation ($r \approx -0.28$), the associations were not statistically significant. Moreover, a Fisher z-test indicated no difference in correlation strength between the two groups ($z = -0.01$, $p = 0.995$), see Figure 4.10.

Despite the lack of statistical significance, the direction of the correlations aligns with expectations from trust theory: lower trust is generally associated with more frequent interventions. This trend, although modest, supports the conceptual framework that users' trust influences their decision to override or accept AV behavior. Larger sample sizes may be needed to detect a clearer relationship or confirm the

observed patterns with greater confidence.

In summary, the exploratory findings underscore the complex interplay between individual predispositions, system design features, and user preferences in shaping trust and intervention behavior. They also reinforce the importance of designing transparent control and explanation strategies that are not only available but also perceived as intuitive and meaningful by users.

5.1.6. Statistical Analysis Approach

The choice of statistical methods in this study was guided by the research design and data characteristics. For hypothesis testing, independent-samples t-tests were employed to compare trust scores and intervention counts between groups defined by participants' subjective evaluations (e.g., positive vs. non-positive views of control buttons or explanations).

While one-way ANOVA or non-parametric alternatives (e.g., Kruskal–Wallis with ε^2 effect size) are typically recommended when comparing more than two groups (e.g., positive, neutral, mixed, negative), they were not employed here due to the extreme imbalance across categories. For example, only one participant was coded as having a negative view of explanations (see Figure 4.6), and that same participant also had the highest intervention count, creating an outlier effect. Running ANOVA or ANCOVA under such conditions would violate key assumptions (e.g., homogeneity of variances, minimum group sizes) and likely produce unreliable results. Furthermore, adding covariates (as in ANCOVA) was not feasible given the limited sample size.

Bayesian methods were also evaluated as an alternative but were not pursued due to the confirmatory nature of the primary hypotheses and the exploratory status of many qualitative variables. Furthermore, specifying appropriate priors would have introduced additional complexity without clear benefits for the study's goals.

Using t-tests allowed for direct and interpretable comparisons between pairs of conditions, aligning with the study's hypotheses and data structure. However, it is acknowledged that the small sample sizes, especially within some coded categories, limit the power of these tests and should be considered when interpreting the results.

5.1.7. Key Take-away

The most consequential finding of this study is the strong relationship between the availability of control mechanisms, specifically the inform and intervene buttons, and participants' trust in the automated vehicles. When participants perceived these controls positively, their reported trust scores were significantly higher. This effect persisted despite the fact that actual use of the buttons was relatively infrequent and did not always correspond to lower intervention counts. This insight has important implications for the design of future automated vehicles, particularly SAE Level 5 systems, which are typically conceptualized as fully autonomous with no user controls or input. The present findings challenge this assumption by demonstrating that even minimal control affordances, when designed to be simple, non-intrusive, and easily understood, can significantly enhance user trust. Participants frequently described the mere availability of controls as increasing their comfort and reducing anxiety, even when they chose not to exercise those controls during driving.

Furthermore, the importance of perceived control aligns with established human factors literature emphasizing the role of shared control and the psychological value of having agency, even in largely automated contexts. This suggests that future SAE Level 5 systems might benefit from reconceptualizing driver roles, shifting from passive occupants to users with optional, limited interaction opportunities that support trust without undermining automation.

An important nuance that emerged in during the experiments, but is easily missed in purely quantitative metrics, is the large grey area between "preventing a crash" and being "too careful." A significant subset of interventions occurred in situations where the AV's behaviour was objectively safe but felt uncomfortably close to the participant's personal risk threshold (e.g., overtaking a stopped van with limited lateral clearance). In those cases, intervening neither avoided an accident nor unequivocally degraded safety; it simply reflected a different tolerance for proximity, braking distance, or gap acceptance. Future AV evaluation metrics might therefore incorporate a 'comfort-margin' band, not just binary crash/ no-crash

outcomes, when assessing unwanted interventions.

In summary, while explanation modalities and user factors influenced trust and intervention behavior to varying degrees, the perceived value of having control emerged as the clearest, most actionable design takeaway from this research.

5.2. Limitations

While the study yielded valuable insights into how explanation modalities and control mechanisms shape trust and intervention behavior in automated driving, several limitations should be acknowledged. First, although the simulation environment was designed to be highly immersive, it could not fully replicate the complexity and dynamic nature of real-world driving. Participants navigated a realistic vehicle interior and scenarios, but the environment lacked traffic lights, speed limits, and did not have a plethora of dynamic road actors, potentially influencing participants' perceptions and decision-making. Several participants noted that the absence of these cues made it harder to judge when intervention or informing actions were appropriate. The car's occasional wiggling at high speeds, due to simulator physics, may have also affected participants' trust or prompted unnecessary interventions.

Second, the functionality and usability of the control buttons introduced some challenges. Although participants generally valued the availability of the inform and intervene options, some feedback indicated uncertainty about how and when to use them, or whether button presses effectively influenced the vehicle's behavior. This ambiguity could have introduced variability into both the trust ratings and the frequency of interventions. Additionally, the placement of the buttons on the keyboard, and the need to press them without looking while wearing a VR headset, may have affected response behavior, especially under time pressure.

Third, while participants were informed that the scenarios aimed to mimic real-world driving, some treated the experience more like a game than a real-life situation. This is a common challenge in simulated studies and could have influenced trust calibration and intervention decisions.

Fourth, the lack of route visualization meant participants could not easily anticipate where the car was heading, possibly increasing uncertainty or leading to interventions that might not have occurred with clearer navigational cues.

Fifth, the binary nature of the intervention system (intervene or not) did not allow for more granular user input or feedback. In reality, drivers might prefer or attempt more nuanced control actions, such as requesting a slow-down or slight trajectory adjustment, rather than a full takeover.

Additionally, some participants reported experiencing virtual reality dizziness or discomfort, which could have distracted from the task and affected both trust ratings and intervention behavior.

Another limitation concerns the relatively small sample size (N = 36), which may have limited the statistical power to detect subtle effects of explanation modality or control mechanisms. The recruitment process was primarily constrained to local participants and relied on in-place exposure without broad advertisement or monetary compensation. This limited outreach reduced the pool of eligible participants and may have introduced sampling bias. While sufficient for exploratory insights and feasibility testing, future studies would benefit from larger and more diverse samples to improve generalizability and statistical robustness.

Furthermore, to reduce subjectivity, two independent coders labelled every open-ended response. When coders disagreed we collapsed the two labels with a simple rule-based "fusion" (e.g., positive + negative → mixed). Although this is preferable to relying on a single coder, it is still an ad-hoc compromise that (i) forces nuanced disagreements into coarse categories, (ii) treats all disagreements as equally important, and (iii) does not propagate the residual uncertainty into the quantitative analyses. A more rigorous approach, e.g., adjudication until consensus, or Bayesian latent-class modelling of coder uncertainty, could yield finer-grained themes and more precise effect estimates in future work.

Finally, while self-reported pre-study variables provided valuable context, these measures are inherently subjective and may not fully reflect participants' true abilities, experiences, or baseline trust tendencies.

Collectively, these limitations highlight the challenges of translating complex human-automation interaction dynamics into controlled experimental settings. They also underscore important design considerations for future studies aiming to improve ecological validity and user experience fidelity.

5.3. Future Work

Building on the findings of this study, several avenues for future research are proposed to further investigate and refine the role of control mechanisms, explanation modalities, and user interaction in SAE Level 5 autonomous vehicles.

5.3.1. Improving Control Mechanism Design and Feedback

While the presence of the inform and intervene buttons positively influenced trust, some participants reported uncertainty about their use and effectiveness. Future studies should focus on refining the design of these control mechanisms, ensuring they provide clear feedback when activated. Visual or auditory confirmation signals could help users understand when their inputs have been registered and what effects they produce, reducing confusion and enhancing the perceived reliability of the controls. Future work may also explore adaptive or multi-modal explanation strategies tailored to user preferences or context.

5.3.2. Enhancing Simulator Realism and Complexity

The simulation environment, while functional and moderately realistic, lacked certain real-world driving complexities such as traffic lights, speed limits, more dynamic actors, and route signaling. Future work should incorporate these elements to create a more immersive and representative driving experience. Increasing realism will not only improve ecological validity but also provide more accurate insights into how drivers might behave and trust automation in real-world contexts.

5.3.3. Differentiating Levels of Intervention

The current binary intervention model limited participants to either intervening or not, without the possibility of expressing varying levels of concern or control. Future research should explore more granular intervention mechanisms, such as graded inputs or suggestive feedback options. This could provide a deeper understanding of how different levels of control affect user trust, perceived safety, and interaction patterns.

5.3.4. Longitudinal and Repeated Exposure Studies

Trust and user behavior in automated driving systems are likely to evolve over time. This study captured participants' responses in single-session exposures, which may not fully reflect long-term interaction dynamics. Future studies should employ longitudinal designs or repeated exposure protocols to examine how trust, intervention behavior, and perceptions of control develop with continued experience. Such research could reveal whether initial impressions persist or change as users become more familiar with the system.

5.3.5. Outlook

Future investigations addressing these directions will not only refine our understanding of user trust and control in fully autonomous vehicles but will also inform the development of SAE Level 5 systems that are both technically robust and aligned with human expectations and preferences. By integrating nuanced control options, enhancing environmental realism, and studying long-term interactions, future research can contribute to safer, more trustworthy, and user-centered autonomous driving experiences.

6

Conclusion

This thesis explored how explanation modality (vocal vs. text) and optional control mechanisms (inform/intervene buttons) affect user trust and intervention behavior in SAE Level 5 autonomous vehicles. Using a VR simulator study with 36 participants, the results offer empirical evidence for how design choices influence user experience in fully autonomous contexts.

Although vocal explanations led to slightly higher trust than text, the difference was not statistically significant. In contrast, participants who valued the presence of control buttons reported significantly higher trust, highlighting the importance of perceived control even when actual intervention was rare. This challenges the SAE Level 5 assumption of no user input, suggesting that minimal, well-designed control affordances can improve user comfort and trust calibration.

Behavioral differences in intervention were aligned with trust trends but did not reach significance, likely influenced by limited realism, ambiguity in button functionality, and small group sizes. Exploratory results reinforced that comfort with automation predicted higher trust and fewer interventions, while open-ended feedback revealed diverse attitudes toward the balance between agency and automation. However, the study's findings must be interpreted in light of several limitations, including potential VR-induced discomfort, the simulator's lack of many dynamic traffic actors, limited route cues, and the discrete nature of intervention measures. Moreover, the simulated environment, despite efforts at realism, cannot fully replicate real-world driving conditions. Future research should address these issues by refining the simulation fidelity, diversifying user control options, and exploring how these findings generalize to more complex or real-world driving tasks.

Together, these findings contribute to the growing body of work on trust in automation and shared control. They suggest that fully autonomous systems may benefit from reintroducing minimal, optional controls to support user comfort without undermining automation. Future AV designs should incorporate intuitive feedback mechanisms, richer environments, and consider long-term exposure to capture evolving user trust

References

[1] Shahin Atakishiyev et al. “Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions”. In: *IEEE Access* 12 (2024), pp. 101603–101625. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3431437. URL: https://ieeexplore.ieee.org/document/10604830/ (visited on 12/26/2024).

[2] Alejandro Barredo Arrieta et al. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. In: *Information Fusion* 58 (June 2020), pp. 82–115. ISSN: 15662535. DOI: 10.1016/j.inffus.2019.12.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S1566253519308103 (visited on 12/26/2024).

[3] Walter Brenner and Andreas Herrmann. “An Overview of Technology, Benefits and Impact of Automated and Autonomous Driving on the Automotive Industry”. In: *Digital Marketplaces Unleashed*. Ed. by Claudia Linnhoff-Popien, Ralf Schneider, and Michael Zaddach. Berlin, Heidelberg: Springer Berlin Heidelberg, 2018, pp. 427–442. ISBN: 978-3-662-49274-1 978-3-662-49275-8. DOI: 10.1007/978-3-662-49275-8_39. URL: http://link.springer.com/10.1007/978-3-662-49275-8_39 (visited on 12/26/2024).

[4] Mark Campbell et al. “Autonomous driving in urban environments: approaches, lessons and challenges”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368.1928 (Oct. 13, 2010), pp. 4649–4672. ISSN: 1364-503X, 1471-2962. DOI: 10.1098/rsta.2010.0110. URL: https://royalsocietypublishing.org/doi/10.1098/rsta.2010.0110 (visited on 12/26/2024).

[5] Carolina Centeio Jorge et al. “Exploring the effect of automation failure on the human’s trustworthiness in human-agent teamwork”. In: *Frontiers in Robotics and AI* 10 (Aug. 23, 2023), p. 1143723. ISSN: 2296-9144. DOI: 10.3389/frobt.2023.1143723. URL: https://www.frontiersin.org/articles/10.3389/frobt.2023.1143723/full (visited on 10/25/2024).

[6] Myke C Cohen et al. “Teamness and Trust in AI-Enabled Decision Support Systems: Current Challenges and Future Directions”. In: ().

[7] Jiqian Dong et al. “Why did the AI make that decision? Towards an explainable artificial intelligence (XAI) for autonomous driving systems”. In: *Transportation Research Part C: Emerging Technologies* 156 (Nov. 2023), p. 104358. ISSN: 0968090X. DOI: 10.1016/j.trc.2023.104358. URL: https://linkinghub.elsevier.com/retrieve/pii/S0968090X23003480 (visited on 12/26/2024).

[8] Alexey Dosovitskiy et al. “CARLA: An Open Urban Driving Simulator”. In: *Proceedings of the 1st Annual Conference on Robot Learning*. 2017, pp. 1–16.

[9] Upol Ehsan et al. “Automated rationale generation: a technique for explainable AI and its effects on human perceptions”. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI ’19: 24th International Conference on Intelligent User Interfaces. Marina del Ray California: ACM, Mar. 17, 2019, pp. 263–274. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3302316. URL: https://dl.acm.org/doi/10.1145/3301275.3302316 (visited on 12/26/2024).

[10] Fredrick Ekman, Mikael Johansson, and Jana Sochor. “Creating Appropriate Trust in Automated Vehicle Systems: A Framework for HMI Design”. In: *IEEE Transactions on Human-Machine Systems* 48.1 (Feb. 2018), pp. 95–101. ISSN: 2168-2291, 2168-2305. DOI: 10.1109/THMS.2017.2776209. URL: http://ieeexplore.ieee.org/document/8125704/ (visited on 12/26/2024).

[11] Frank Flemisch et al. “Towards a dynamic balance between humans and automation: authority, ability, responsibility and control in shared and cooperative control situations”. In: *Cognition, Technology & Work* 14.1 (Mar. 2012), pp. 3–18. ISSN: 1435-5558, 1435-5566. DOI: 10.1007/s10111-011-0191-6. URL: http://link.springer.com/10.1007/s10111-011-0191-6 (visited on 11/13/2024).

[12] Chunshi Guo et al. “Driver–vehicle cooperation: a hierarchical cooperative control architecture for automated driving systems”. In: *Cognition, Technology & Work* 21.4 (Nov. 2019), pp. 657–670. ISSN: 1435-5558, 1435-5566. DOI: 10.1007/s10111-019-00559-2. URL: http://link.springer.com/10.1007/s10111-019-00559-2 (visited on 11/13/2024).

[13] Robert S. Gutzwiller et al. “Positive bias in the ‘Trust in Automated Systems Survey’? An examination of the Jian et al. (2000) scale”. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 63.1 (Nov. 2019), pp. 217–221. ISSN: 1071-1813, 2169-5067. DOI: 10.1177/1071181319631201. URL: https://journals.sagepub.com/doi/10.1177/1071181319631201 (visited on 04/07/2025).

[14] Sabrina M. Hegner, Ardion D. Beldad, and Gary J. Brunswick. “In Automatic We Trust: Investigating the Impact of Trust, Control, Personality Characteristics, and Extrinsic and Intrinsic Motivations on the Acceptance of Autonomous Vehicles”. In: *International Journal of Human–Computer Interaction* 35.19 (Nov. 26, 2019), pp. 1769–1780. ISSN: 1044-7318, 1532-7590. DOI: 10.1080/10447318.2019.1572353. URL: https://www.tandfonline.com/doi/full/10.1080/10447318.2019.1572353 (visited on 11/13/2024).

[15] Brittany E. Holthausen et al. “Situational Trust Scale for Automated Driving (STS-AD): Development and Initial Validation”. In: *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. AutomotiveUI ’20: 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. Virtual Event DC USA: ACM, Sept. 21, 2020, pp. 40–47. ISBN: 978-1-4503-8065-2. DOI: 10.1145/3409120.3410637. URL: https://dl.acm.org/doi/10.1145/3409120.3410637 (visited on 04/07/2025).

[16] Makoto Itoh, Frank Flemisch, and David Abbink. “A hierarchical framework to analyze shared control conflicts between human and machine”. In: *IFAC-PapersOnLine* 49.19 (2016), pp. 96–101. ISSN: 24058963. DOI: 10.1016/j.ifacol.2016.10.468. URL: https://linkinghub.elsevier.com/retrieve/pii/S2405896316320584 (visited on 12/26/2024).

[17] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. “Foundations for an Empirically Determined Scale of Trust in Automated Systems”. In: *International Journal of Cognitive Ergonomics* 4.1 (Mar. 2000), pp. 53–71. ISSN: 1088-6362. DOI: 10.1207/S15327566IJCE0401_04. URL: http://www.tandfonline.com/doi/abs/10.1207/S15327566IJCE0401_04 (visited on 04/07/2025).

[18] Matthew Johnson and Jeffrey M. Bradshaw. “The role of interdependence in trust”. In: *Trust in Human-Robot Interaction*. Elsevier, 2021, pp. 379–403. ISBN: 978-0-12-819472-0. DOI: 10.1016/B978-0-12-819472-0.00016-2. URL: https://linkinghub.elsevier.com/retrieve/pii/B9780128194720000162 (visited on 03/30/2025).

[19] Carolina Centeio Jorge et al. “Artificial Trust in Mutually Adaptive Human-Machine Teams”. In: ().

[20] Robert Kaufman, Jean Costa, and Everlyne Kimani. “Effects of multimodal explanations for autonomous driving on driving performance, cognitive load, expertise, confidence, and trust”. In: *Scientific Reports* 14.1 (June 6, 2024), p. 13061. ISSN: 2045-2322. DOI: 10.1038/s41598-024-62052-9. URL: https://www.nature.com/articles/s41598-024-62052-9 (visited on 04/22/2025).

[21] Jeamin Koo et al. “Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance”. In: *International Journal on Interactive Design and Manufacturing (IJIDeM)* 9.4 (Nov. 2015), pp. 269–275. ISSN: 1955-2513, 1955-2505. DOI: 10.1007/s12008-014-0227-2. URL: http://link.springer.com/10.1007/s12008-014-0227-2 (visited on 10/29/2024).

[22] J. D. Lee and K. A. See. “Trust in Automation: Designing for Appropriate Reliance”. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46.1 (Jan. 1, 2004), pp. 50–80. ISSN: 0018-7208. DOI: 10.1518/hfes.46.1.50_30392. URL: http://hfs.sagepub.com/cgi/doi/10.1518/hfes.46.1.50_30392 (visited on 12/26/2024).

[23] Jesse Levinson et al. “Towards fully autonomous driving: Systems and algorithms”. In: *2011 IEEE Intelligent Vehicles Symposium (IV)*. 2011 IEEE Intelligent Vehicles Symposium (IV). Baden-Baden, Germany: IEEE, June 2011, pp. 163–168. ISBN: 978-1-4577-0890-9. DOI: 10.1109/IVS.2011.5940562. URL: http://ieeexplore.ieee.org/document/5940562/ (visited on 12/26/2024).

[24] T. Luettel, M. Himmelsbach, and Hans-Joachim Wuensche. “Autonomous Ground Vehicles—Concepts and a Path to the Future”. In: *Proceedings of the IEEE* 100 (Special Centennial Issue May 2012), pp. 1831–1839. ISSN: 0018-9219, 1558-2256. DOI: 10.1109/JPROC.2012.2189803. URL: http://ieeexplore.ieee.org/document/6179503/ (visited on 10/29/2024).

[25] A. V. Shreyas Madhav and Amit Kumar Tyagi. “Explainable Artificial Intelligence (XAI): Connecting Artificial Decision-Making and Human Trust in Autonomous Vehicles”. In: *Proceedings of Third International Conference on Computing, Communications, and Cyber-Security*. Ed. by Pradeep Kumar Singh et al. Vol. 421. Series Title: Lecture Notes in Networks and Systems. Singapore: Springer Nature Singapore, 2023, pp. 123–136. ISBN: 978-981-19-1141-5 978-981-19-1142-2. DOI: 10.1007/978-981-19-1142-2_10. URL: https://link.springer.com/10.1007/978-981-19-1142-2_10 (visited on 12/26/2024).

[26] Carina Manger et al. “Providing Explainability in Safety-Critical Automated Driving Situations through Augmented Reality Windshield HMIs”. In: *Adjunct Proceedings of the 15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. AutomotiveUI ’23: 15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. Ingolstadt Germany: ACM, Sept. 18, 2023, pp. 174–179. ISBN: 979-8-4007-0112-2. DOI: 10.1145/3581961.3609874. URL: https://dl.acm.org/doi/10.1145/3581961.3609874 (visited on 12/26/2024).

[27] Stephanie M. Merritt and Daniel R. Ilgen. “Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions”. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50.2 (Apr. 2008), pp. 194–210. ISSN: 0018-7208, 1547-8181. DOI: 10.1518/001872008X288574. URL: https://journals.sagepub.com/doi/10.1518/001872008X288574 (visited on 03/30/2025).

[28] Andreas Lars Müller et al. “Effects of non-driving related tasks on mental workload and take-over times during conditional automated driving”. In: *European Transport Research Review* 13.1 (Dec. 2021), p. 16. ISSN: 1867-0717, 1866-8887. DOI: 10.1186/s12544-021-00475-5. URL: https://etrr.springeropen.com/articles/10.1186/s12544-021-00475-5 (visited on 12/26/2024).

[29] Beckers N. et al. *JOAN, a human-automated vehicle experiment framework*. Retrieved from https://github.com/tud-hri/joan. Online. 2021.

[30] Sina Nordhoff et al. “Perceived safety and trust in SAE Level 2 partially automated cars: Results from an online questionnaire”. In: *PLOS ONE* 16.12 (Dec. 21, 2021). Ed. by Sergio A. Useche, e0260953. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0260953. URL: https://dx.plos.org/10.1371/journal.pone.0260953 (visited on 10/29/2024).

[31] Marie-Pierre Pacaux-Lemoine and Frank Flemisch. “Layers of Shared and Cooperative Control, assistance and automation”. In: *IFAC-PapersOnLine* 49.19 (2016), pp. 159–164. ISSN: 24058963. DOI: 10.1016/j.ifacol.2016.10.479. URL: https://linkinghub.elsevier.com/retrieve/pii/S2405896316320699 (visited on 12/26/2024).

[32] Jakob Peintner et al. “Explaining Away Control: Exploring the Relationship between Explainable AI and Passengers’ Desire for Control in Automated Vehicles”. In: *Adjunct Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. AutomotiveUI ’24: 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. Stanford CA USA: ACM, Sept. 22, 2024, pp. 155–160. ISBN: 979-8-4007-0520-5. DOI: 10.1145/3641308.3685040. URL: https://dl.acm.org/doi/10.1145/3641308.3685040 (visited on 10/25/2024).

[33] Jakob Benedikt Peintner, Carina Manger, and Andreas Riener. ““Can you rely on me?” Evaluating a Confidence HMI for Cooperative, Automated Driving”. In: *Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. AutomotiveUI ’22: 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. Seoul Republic of Korea: ACM, Sept. 17, 2022, pp. 340–348. ISBN: 978-1-4503-9415-4. DOI: 10.1145/3543174.3546976. URL: https://dl.acm.org/doi/10.1145/3543174.3546976 (visited on 11/13/2024).

[34] On-Road Automated Driving (ORAD) committee. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. DOI: 10.4271/J3016_202104. URL: https://www.sae.org/content/j3016_202104 (visited on 10/29/2024).

[35] Jan Maarten Schraagen et al. "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* 64.1 (Dec. 2020), pp. 339–343. ISSN: 1071-1813, 2169-5067. DOI: 10.1177/1071181320641077. URL: https://journals.sagepub.com/doi/10.1177/1071181320641077 (visited on 12/26/2024).

[36] Bobbie D. Seppelt and John D. Lee. "Keeping the driver in the loop: Dynamic feedback to support appropriate use of imperfect vehicle control automation". In: *International Journal of Human-Computer Studies* 125 (May 2019), pp. 66–80. ISSN: 10715819. DOI: 10.1016/j.ijhcs.2018.12.009. URL: https://linkinghub.elsevier.com/retrieve/pii/S1071581918301277 (visited on 12/26/2024).

[37] Jork Stapel, Freddy Antony Mullakkal-Babu, and Riender Happee. "Automated driving reduces perceived workload, but monitoring causes higher cognitive load than manual driving". In: *Transportation Research Part F: Traffic Psychology and Behaviour* 60 (Jan. 2019), pp. 590–605. ISSN: 13698478. DOI: 10.1016/j.trf.2018.11.006. URL: https://linkinghub.elsevier.com/retrieve/pii/S1369847818301335 (visited on 12/26/2024).

[38] Ana Tanevska, Katie Winkle, and Ginevra Castellano. *"I don't like things where I do not have control": Participants' Experience of Trustworthy Interaction with Autonomous Vehicles*. Version Number: 1. 2025. DOI: 10.48550/ARXIV.2503.15522. URL: https://arxiv.org/abs/2503.15522 (visited on 03/30/2025).

[39] Jacques Terken and Bastian Pfleging. "Toward Shared Control Between Automated Vehicles and Users". In: *Automotive Innovation* 3.1 (Mar. 2020), pp. 53–61. ISSN: 2096-4250, 2522-8765. DOI: 10.1007/s42154-019-00087-9. URL: http://link.springer.com/10.1007/s42154-019-00087-9 (visited on 11/13/2024).

[40] C. Urmson and W. Whittaker. "Self-Driving Cars and the Urban Challenge". In: *IEEE Intelligent Systems* 23.2 (Mar. 2008), pp. 66–68. ISSN: 1541-1672. DOI: 10.1109/MIS.2008.34. URL: http://ieeexplore.ieee.org/document/4475861/ (visited on 12/26/2024).

[41] Gesa Wiegand et al. ""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*. MobileHCI '20: 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services. Oldenburg Germany: ACM, Oct. 5, 2020, pp. 1–11. ISBN: 978-1-4503-7516-0. DOI: 10.1145/3379503.3403554. URL: https://dl.acm.org/doi/10.1145/3379503.3403554 (visited on 10/29/2024).

[42] Philipp Wintersberger et al. "Explainable Automation: Personalized and Adaptive UIs to Foster Trust and Understanding of Driving Automation Systems". In: *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. AutomotiveUI '20: 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. Virtual Event DC USA: ACM, Sept. 21, 2020, pp. 252–261. ISBN: 978-1-4503-8065-2. DOI: 10.1145/3409120.3410659. URL: https://dl.acm.org/doi/10.1145/3409120.3410659 (visited on 10/29/2024).

A

Study Materials and Participant Data

A.1. Informed Consent

User Trust and Intervention Behavior in SAE Level 5 Autonomous Vehicles

Principal Researcher

Stelios Avgousti, Master's Student, Computer Science, Interactive Intelligence

Responsible Researcher

Myrthe Tielman, Assistant Professor

Institution

Delft University of Technology (TU Delft)

Invitation

You are invited to participate in a study investigating how users interact with fully autonomous vehicles (SAE Level 5). This study will involve using a driving simulator under three different conditions, followed by trust-related surveys, and will take place at TU Delft facilities.

Procedure

- Complete a pre-study questionnaire about general attitudes toward AI and past experiences with autonomous vehicles.
- Perform three simulated driving trials, lasting approximately 20–25 minutes in total.
- Complete a final survey assessing overall trust in AVs after Trial 3.
- Answer interview questions regarding the different driving trials.
- Total time commitment: approximately 45–60 minutes.

Experiment Details

The simulator will mimic real-world driving conditions. You will experience fully autonomous driving and will not need to manually control the vehicle. In certain trials, you may use provided buttons to signal or override the system's behavior. The first trial will be a familiarization round. The second and third trials will use slightly different simulation versions. Interaction data and survey responses will be recorded.

Eligibility Criteria

- At least 18 years old.

- Valid driving license.
- Proficient in English.

Voluntary Participation & Right to Withdraw

- Participation is voluntary.
- You may withdraw at any time without providing a reason or penalty.
- Upon withdrawal, your data will be deleted upon request unless it has been included in anonymized analyses.

Potential Risks and Benefits

Risks: No significant risks are expected. Some participants may experience mild discomfort such as motion sickness or fatigue. You may take breaks or discontinue participation at any time.
Benefits: You may gain insight into AI vehicles. Your participation will contribute to improving trust-calibrated autonomous vehicle systems.

Confidentiality & Data Protection

- Responses will be anonymized and stored securely on TU Delft servers.
- No personally identifiable information will be published.
- Data will be stored for five years, following TU Delft's research policies.
- Only authorized researchers will have access to your data.
- A random participant ID will be used to anonymize data.

Consent Statement

Please tick the appropriate boxes:

- I confirm that I have read and understood the information above and had the opportunity to ask questions.
- I understand that my participation is voluntary and that I may withdraw at any time without consequences.
- I understand the purpose of this study and what my participation involves.
- I consent to my anonymized data being used for research and publication purposes.
- I understand that button usage and responses to trust surveys will be recorded for analysis.
- I understand that no personally identifiable information will be linked to my responses.
- I understand that I can contact the research team if I have concerns or wish to withdraw my data before publication.

Name of Participant: _____

Signature: _____

Date: _____

A.2. Pre-Study

- **Age:** _____
- **Gender:** Male Female Other
- **Driving experience:** Likert scale 1–5 (1 = Not experienced, 5 = Very experienced)
- **Have you ever been in a vehicle where the driver didn't need to have their hands on the steering wheel? (e.g., Waymo or Tesla):** Yes / No

- **I feel comfortable relying on automated systems like self-driving cars:** Likert scale 1–7 (1 = Strongly disagree, 7 = Strongly agree)

A.3. Trust Survey

All questions use a 1–7 Likert scale (1 = Strongly disagree, 7 = Strongly agree).

1. The autonomous vehicle provides security.
2. The autonomous vehicle behaves in an underhanded manner.
3. I am familiar with the autonomous vehicle.
4. The autonomous vehicle is deceptive.
5. I am confident in the autonomous vehicle.
6. I am suspicious of the autonomous vehicle's intent, action, or output.
7. I am wary of the autonomous vehicle.
8. I can trust the autonomous vehicle.
9. The autonomous vehicle's action will have harmful or injurious outcomes.
10. The autonomous vehicle has integrity.
11. The autonomous vehicle is dependable.
12. The autonomous vehicle is reliable.

A.4. Post-study Open Questions

1. **Rank the versions of the car starting with the one you felt least comfortable with:**
   - Version 1 – No buttons or explanations
   - Version 2 – Buttons only
   - Version 3 – Buttons + Explanations

   (Participants dragged and rearranged the versions.)
2. **Why did you assign the versions in this order?** (Keep answers short and concise.)
3. **How did the explanations influence your trust in the vehicle, if at all?** (Keep answers short and concise.)
4. **What was your experience with the control buttons during the drive?** (Keep answers short and concise.)
5. **Any feedback or thoughts about the driving experience?** (Keep answers short and concise.)

A.5. Participant Distributions

A.5.1. Gender

The gender distribution across the two modality groups was relatively balanced. In the text explanation group, there were 7 female and 11 male participants, whereas the voice explanation group included 5 female and 13 male participants as seen in A.1. A chi-square test of independence was conducted to examine whether gender distribution differed significantly between the groups. The results indicated no significant association between gender and explanation modality, $\chi^2(1) = 0.12$, $p = 0.724$, suggesting that gender was approximately evenly distributed across the two groups.

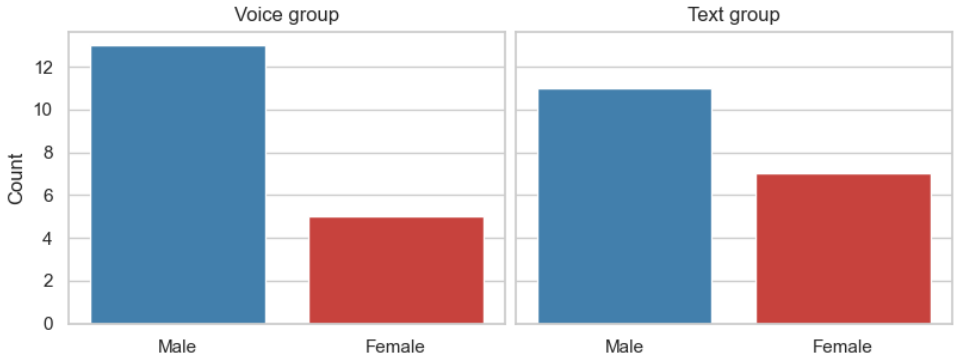


Figure A.1: The gender distribution per group.

A.5.2. Age

The age distribution of participants was examined across the two explanation modality groups. Visual inspection of histograms and density plots indicated roughly similar age distributions but suggested potential deviations from normality, see A.2. To formally assess normality, Shapiro–Wilk tests were conducted for each group. Results indicated significant departures from normality in both the voice group ($W = 0.867$, $p = 0.016$) and the text group ($W = 0.888$, $p = 0.036$). Therefore, a non-parametric Mann–Whitney U test was used to compare ages between groups. The test revealed no significant difference in participant age distributions between the voice and text conditions ($U = 142.5$, $p = 0.544$). This suggests that age was approximately balanced across the two groups.

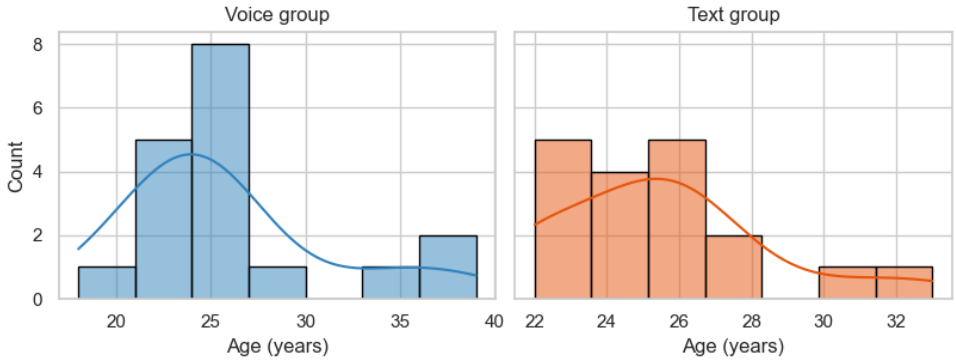


Figure A.2: The Age distribution per group.

A.5.3. Driving Experience

Participants reported their self-assessed driving experience on a 5-point scale (1 = not experienced, 5 = very experienced). To test whether driving experience differed across modality groups, a Mann–Whitney U test was conducted. The test indicated no significant difference between the voice and text groups ($U = 209.5$, $p = 0.122$). This suggests that driving experience was approximately balanced between groups and unlikely to confound the effects of explanation modality on trust or intervention behavior. The distribution is visible at A.3

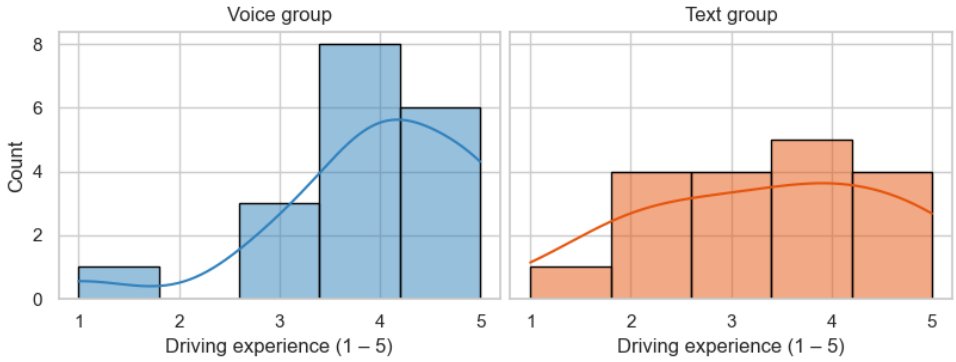


Figure A.3: The Self-reported driving experience distribution per group.

A.5.4. Comfortability with automated systems

Participants rated their comfort with relying on automated systems like self-driving cars on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree), serving as a proxy for propensity to trust automation. A Mann–Whitney U test revealed no significant difference between the voice and text groups ($U = 167.0$, $p = 0.885$). This indicates that the two groups had comparable baseline attitudes toward automation, minimizing potential confounding effects on trust or intervention behaviors observed in later trials. Distribution between the two groups is visible below (A.4).

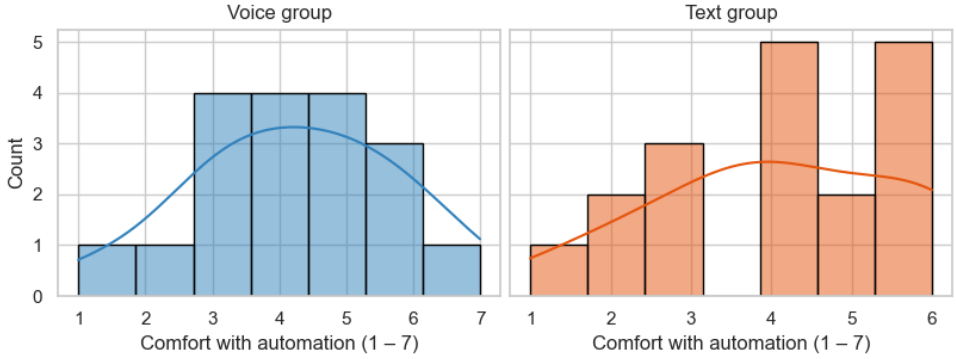


Figure A.4: The Self-reported propensity to trust AI distribution per group.

A.5.5. People that have been passengers in Avs

Participants indicated whether they had previously been passengers in a fully autonomous vehicle (no steering wheel). A chi-square test of independence revealed no significant difference in prior AV experience between the voice and text groups ($\chi^2(3) = 0.00$, $p = 1.000$). This suggests that prior exposure to autonomous vehicles was balanced across groups, reducing the likelihood that familiarity effects influenced trust or intervention behaviors during the experimental trials. The distribution between the two groups can be found A.5.

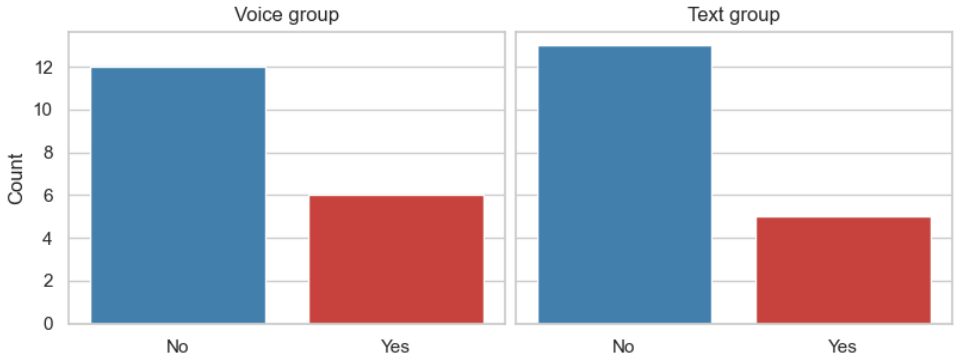


Figure A.5: People that have been in an AV before, distribution per group.

A.6. Open-ended Responses and Themes

Participant	Explanation Question	Control Question	Why Ranked in that order	Final Feedback	Btn Theme	Expl Theme	Btn Theme 2	Expl Theme 2
1	The explanations gave me more trust and made the use of buttons more easy.	In the beginning it didnt feel intuitive because you are used to brake and accelerate, but with the explanation and in the context of a self driving car it worked well.	The addition of buttons gave the idea that the car was not safe by itself. Adding explanations made it more understandable why the car behaved like it did, making it more comfortable.	I wouldn't say it was comfortable. You still need to pay a lot of attention.	mixed	positive	mixed	positive
2	It gave me insight into the vehicles decisions and I wasn't surprised by any abrupt alterations in the driving.	They felt responsive and gave me enough control over the vehicle's decisions.	Having no explanations and no buttons felt like I was simply a passenger with no impact on the outcome of the driving experience. The buttons in trial 2 gave a sense of control over the vehicles actions but it was still unclear what the vehicle was doing and why. In the third trial it felt the most comfortable since I also had input as to why we are stopping/slowing down, and at the same time I could influence these decisions.	It felt reassuring and the car was driving competently, it didn't feel unsafe at any point.	positive	positive	positive	positive
3	Explanations help me to understand the thinking process of the vehicle; thus, I know what I will expect (e.g. slowing down or stop)	It helps me to give a control when needed. It is similar to the brake/acceleration pedal.	The button helps me to feel that I still have a control over the autonomous vehicles. The explanations help me to understand or expect what will happen in the few seconds ahead.	Maybe you can add indication if the vehicle wants to turn left or right. It helps to to know which side of the road I need to look at.	positive	positive	positive	positive
4	Help me prepare for car's next move	Gave me more control	Buttons give a bit more control, explanations let me know the intend beforehand	Never been to autonomous vihacle before, felt a bit unsure	positive	positive	positive	positive

Participant	Trust Explanation	Control Buttons	Why Ranked	Final Feedback	Btn Theme	Expl Theme Btn Theme 2	Expl Theme 2	
5	it increased my awareness regarding the car's decision making but it did not make me trust it much more.	Intuitive but I could use some more time getting used to the right Ctrl button.	Having some control over the actions gives me some confidence in case I disagree with the decisions of the autonomous system. Having the explanations of what the car does also provides an extra reassurance since I do not have to guess what the car is planning to do or react	Some rapid movements of the wheel lead me to slow down the car. Maybe some more incidents or cases in the town would be interesting to see.	neutral	neutral	positive	positive
6	It helped me decide if i should take any manual actions or not	hard to understand at first but gets easier through time and practise	It s nice ti get action verification so that i dont have to think if the car is trying to stop or not. And manuak overwrite is always good, gives the more confidence in the experience	I would love to see some more complex scenarios to test the trust to the limit	mixed	positive	positive	positive
7	Helped me understand the intentions of the car, however it is annoying having to read what the car is doing while doing it, since i might start reading a bit late.	easy to understand but took some time to get used to	having an explanation makes it easier to predict the car movements and take the necessary actions to prevent accidents. Not having any input in the car driving feels dangerous and stressful	Fun driving experience	mixed	mixed	positive	mixed
8	They made me feel safer, and like I am more connected with the vehicle.	The multiple functions made them confusing to use, but it was nice to have some form of control.	I feel more stressed when i have control but no explanations, because I feel like I HAVE to intervene. No control and explanations therefore feels safer. Having control and explanations make me feel the safest, because I don't feel pressured to intervene, but I can if I want to.	The car moves a bit erratically when cornering, and gets quite close to obstacles.	mixed	positive	mixed	positive
9	It increase transparent communication	No very controllable because it is not familiar to me	when no buttons I feel easy and relax because the vehicle will probably do everything for me. when I have a button and explanation I will more in control of everything.	slow down too much	negative	positive	mixed	positive
10	By adding audio it enhance the senses used, creating a more holistic experience	Confortrable	I had more control and that gave me please of mind	Still needs a bit of work to run more smoothly but is a nice overall driving stimulation	positive	positive	positive	positive

Participant	Trust Explanation	Control Buttons	Why Ranked	Final Feedback	Btn Theme	Expl Theme Btn Theme 2	Expl Theme 2	
11	Although sometimes I could already suspect what the vehicle is doing, it felt even better to see it written.	I didn't really use them. I used them ones and I killed a biker. RIP.	I felt like no buttons and explanations has absolutely no control, Although I felt very comfortable with the autonomous car. Buttons only makes me feel a bit worried that I forget what each buttons do, but buttons and explanations make me feel even more confident!	In general I felt complete trust on the vehicle and I felt very comfortable knowing that it has full control. Of course having the buttons in case of emergency such as system failure makes me feel safer.	positive	neutral	positive	positive
12	I knew what the vehicle was going to do, so I knew if i had to intervene or not.	They were helpful in some situations, mainly the break button	Being able to control the car, and the car telling what it is going to do, gives a sense of security and makes me as a driver feel more safe	Would like the car to have a bit more speed so breaking would be nececairy more often	positive	positive	positive	positive
13	I was a bit late seeing the explanations although they did pop up in front of my eyes but i was more focused on the road than on the explanations. I feel if i got more used to it to check on the explanations that pop out i would cooperate better.	It was easy to remember what each button does since it was also written with yellow letters in the car although it was a bit tricky when it comes to actually using them but again the more practice the better it work.	Having no buttons felt less comfortable as i feel i was able to see the cars stopped in the middle of the road before the car would provide explanations to what is happening.	The fact that the car did not stop at the red lights made be a bit hesitant whether to stop the car or not so maybe if the car was automated to stop on the red lights and i knew then i wouldn't press the button to sto.	mixed	mixed	mixed	neutral
14	the explanations were 1-2 seconds too late, i had already seen the obstacle and decided to slow down, and then the car did the same. i didnt feel like intervening many times as i usually agreed with the explanations.	very normal, just felt a bit weird to start again after stopping completely.	version 3, the car behaved in the most predictable manner, there were also fewer obstacles.	i liked the experience and the car behaved in a predictable manner. i only thought the steering was a bit off centre and not realistic. car accelerated and decelerated as expected.	neutral	mixed	neutral	mixed
15	i was more prepared for the upcoming obstacles and could see if i could trust it or not. if it was legit info	they were really helpful but didn't have full control	It starts with an unknown kind of feeling of what is going on but then it becomes clearer when you're given the explanations as well	there was not a signaling of direction changing lanes with no apparent reason the messages on the screen were helpful and accurate	mixed	positive	mixed	positive
16	You are prepared on the situation, and you dont have to wait to the last moment to see whats gonna happen	Its good to keep the control in your own hands, so you will always be able of stopping if something seems to go wrong	with the 1st one you have no single control, as you do have the most control with the 3rd trial. You can stop at any time and you are informed on what the vehicle is gonna do.	the car sometimes drove quite close to some cars parked on the curb, personally i would take a bit more space to them.	positive	positive	positive	positive

Participant	Trust Explanation	Control Buttons	Why Ranked	Final Feedback	Btn Theme	Expl Theme Btn Theme 2	Expl Theme 2	
17	The explanations provided a quick explanation to why the car made a decision, providing clarity and transparency, therefore more trust.	Quite intuitive thanks to the ability to test it out during Trial 2.	Having any control of the movement of the car provides some kind of confidence to intervene if possible. Adding the explanations to the car's actions provides clarity. Having both then provides the most comfort.	Smoother steering on the highway next time	positive	positive	positive	positive
18	They increased my trust	I was feeling more comfortable because I had the option to take some kind of control	I felt less comfortable with version 2, because I had no input infortmation. The buttons may cause an accident if not used correctly. Version 1 felt more comfortable because I was only relying on the car. With version 3 I was feeling very comfortable because I had the option to interrupt based on the information given by the car	After this experiment I feel more comfortable driving in an autonomous car (Never been in one)	mixed	positive	mixed	positive
19	positively but would need more feedback to feel sure about the car experience	good	based on my experience with the 3 vehicles i felt the most comfortable when instructions where present and when i had no control or responsibility for the outcome	smooth and reliable driving except some key errors that cause my concern (getting on the pavement to make a turn)	positive	mixed	positive	positive
20	Feeling some safety that the vehicle understands what its doing	Used mostly after the car stopped completely after congestion/accident/roadblock. Only to begin it again. Other than that I generally agreed with the car and didnt have to press	I felt safer with the audio explanations knowing to anticipate how the car is going to behave. Buttons made me feel i had some control. Having no control is a bit stressful	Would like it to drive faster.	positive	positive	positive	positive
21	no they didn't	easy	im an experienced driver	Good experience, good for response practice from new drivers	negative	positive	negative	neutral
22	It made me misinterpret the timing of the execution of the action	a bit confusing but overall fine	There is a slight delay between what the car says its gonna so vs when it actually does it	the slight delay is a annoying	mixed	negative	mixed	negative
23	good and understandable	fun and explorative	Version 3 and version 1 give more comfortable compare to the version 2, it is clear and the information easy to understand or no information or control like version 1	good exploration and experiment	positive	positive	positive	positive
24	Knowing what the car is intending to do is assuring.	Overly simplistic. Lack of control.	I want to know what the car is doing and what it is going to do. And I want to be able to intervene.	I'd rather drive the car myself.	neutral	positive	mixed	positive

Participant	Trust Explanation	Control Buttons	Why Ranked	Final Feedback	Btn Theme	Expl Theme Btn Theme 2	Expl Theme 2	
25	it's nice to know what the car is thinking so if it slows down out of nowhere at least i know why. thus the trust increased	they worked really well, i just had some trouble understanding which button made the car start again after stopping. however there was no issues with the functioning	i liked having no buttons in version 1 cause i didn't have to think about anything, but i was comfortable enough with the buttons in version 3	the car was wobbly in the highway	positive	positive	positive	positive
26	helped me to predict what the car will do and to rely on it	felt easier to experience the self-driving, gave control to me if there was unexpected event	aural explanations help trusting the vehicle by giving a prediction of its actions, buttons gives human control to feel more humaine and trustworthy	stressful experience, aural explanations helped with feeling uneven	positive	positive	positive	positive
27	It gave me reassurance that it sees what is happening	Neutral because even there were only 2 buttons, they were performing 4 functions which was a little confusing	In version 2, I felt like I had to keep an eye on the road and also 'help' the system. Even though version 1 didn't have buttons, I felt a little more comfortable because I assumed that the car would have been properly tested before hitting the streets. The explanations on version 3 are a nice to have. It gives reassurance without the need for me to press buttons.	Something I missed was the car indicating the direction it would go to. It's quite important on the road	negative	positive	neutral	positive
28	More trust in the system used for the first time	Not so comfortable or intuitive	I prefer to know what the vehicle is going to do	The lateral control was quite dizzy and makes my trust low	negative	positive	negative	positive
29	They made me trust the vehicle more, as I could potentially spot any mistakes, even if it didnt make any. At the same time, if I really trusted the vehicle I would not need any explanations.	During the first trial with the control buttons things were pretty confusing. During the second one, where the car's intentions were visible, things made a little more sense.	Buttons only create unreliability for responses. The car could be already stopping and you could unknowingly cause a crash like I did. If there is no explanation I would prefer to have to trust the car.	I personally believe that the buttons introduce a weird middle ground between trust and accountability. If the car is really better at driving that humans, then real trust would imply never needing to interfere with the car's actions.	negative	mixed	negative	mixed
30	it gave me more trust.	going against my instincts. (both slow down)	First version i had to adapt, so i didnt really focus. version 2 felt better then version 3. version 3 had moments when it didnt explain, but i felt like sometimes it drove to close to the curf/car.	perhaps a trial zero for people who never did vr before.	neutral	positive	neutral	positive

Participant	Trust Explanation	Control Buttons	Why Ranked	Final Feedback	Btn Theme	Expl Theme Btn Theme 2	Expl Theme 2	
31	It made me trust the vehicle less because it didn't always stop or break fast enough for the situation.	The stop/don't stop button did not work properly in the previous simulations, meaning that when i pressed it before it would only stop and never not stop. The one time I needed a hard/quick stop it didn't stop, it just kept going and so I crashed.	Because when you did not have any control over the car, it didn't matter what you felt or did where as when I had explanations and could control the buttons the buttons felt counterintuative from the previous experiences.	The driving experience was relatively nice, a bit slow and not very natural but overall I could trust the car for the basic driving needs.	negative	positive	negative	positive
32	Explaining, added a layer of understanding and made me less wary	Easy to manage, after the first trials	The voice helped ease the transition between stages the car is going through	Sound would make this more immersive and might change the outcomes but it was still engaging.	positive	positive	positive	positive
33	The explanations helped to increase trust given that it allows for me to make the final judgement if I want to accept the decision or not.	I think it was good, after understanding that the left one is more like a "go slower for a bit" button it became better.	While the car didnt generate distrust allowing me to override the decision made it more comfortable, and also I was more aquainted with the experience by the third trial	Very relaxing, although the first one was a bit uncomfortable given that I was still getting aquainted with the experience and had no control over it.	positive	positive	positive	positive
34	make me trust car's reactions	I liked it cause it gave me more control	I prefer it when I have some level of control	it was a fun experience	positive	positive	positive	positive
35	Because I knew what the next action of the car was, so it made me more comfortable.	The controls were helpful	I assigned the version three as the most comfortable one because it was generally more comfortable when I knew what actions to expect from the car, and it felt most comfortable when I was able to control its actions.	The experience was fun and interesting.	positive	positive	positive	positive
36	It influenced my trust since the vehicle did exactly what the explanation said	got a bit confused at first but after 1-2 tries i was very comfortable with the buttons	version 1 made me feel a bit unsafe not knowing what the actions will be or having no control, version 2 made me feel more comfortable since i had control of the vehicle, version 3, at this point the vehicle gained my full trust since it showed me it is reliable and it also showed me its intentions	maybe next time some people can jump in the road out of nowhere so we can see the vehicles reaction	mixed	positive	mixed	positive