Handover from AI to Humans in L3 Automated Vehicles: A Pilot Study

Abbas Kerem Doğan

Faculty of Behavioral, Management and Social Sciences (BMS)

Department Cognitive Psychology Ergonomics (CPE)

Master's Thesis: Human Factors and Engineering Psychology

1st Supervisor: Simone Borsci

2nd Supervisor: Rob van der Lubbe

Table of Contents

Abstract

In the world of automated driving, the shift from Level 2 (driver is still driving) to Level 3 (driver overlooks the automated driving) will mark an important change in the role of the driver. In L3 automated vehicles, drivers are "passive" as they are only expected to drive after moments of handover. This means that drivers are expected to take control of the driving from AI only when AI requests a takeover. The underlying psychological factors that might affect the human-AI interaction during handover are crucial to understand because unsuccessful takeovers can lead to accidents. To our knowledge, the present study is the first that attempted to examine the relationship between a time constrained takeover request, situational trust, mental workload, sleepiness, state anxiety, and success in taking over (takeover performance of drivers). In this pilot, we identified significant effects of mental workload and state anxiety on taking over success. We found situational trust to be a significant mediator of mental workload and state anxiety. Through Bayesian Structural Equation Modeling (BSEM) and model selection approaches, we identified a model that showed a significant explanation of success in taking over. Importantly, the study found supporting evidence for the relevance of voice chatbot simulations as a continuous measurement tool in VR studies.

Keywords: Autonomous Vehicles, Handover, Human-AI Interaction, Situational Trust, Mental Workload, Sleepiness, State Anxiety, Chatbots, Bayesian Structural Equation Modeling, Model Selection

Introduction

Automated vehicles (AVs) have the potential to revolutionize transportation systems as they are expected to reduce accident rates by enhancing vehicle safety and reducing driver mental workload (Xing et al., 2021). However, AV usage can be considered controversial because of the gap between driver expectations of AV and the actual capability of the system (Merriman, 2021). Compared to actively controlling the vehicle during manual driving, the driver is responsible for passive monitoring of the road and taking over during emergencies during automated driving (AD) (Society of Automotive Engineers International, 2018). The interaction between the driver and AV, as well as how the AV communicates the current state of driving to the driver, appears to be a crucial factor in this context. Effective communication between the driver and the AV is essential so that the driver is aware of the vehicle's current state and can take control in case of an emergency.

Currently, the Society of Automotive Engineers (SAE, 2021) defines six levels of automation. Each level of automation offers a different relationship between the driver and the vehicle (Hopkins & Schwanen, 2021). In levels 0 (no automation) to 2 (partial automation), the driver must drive and monitor the vehicle (SAE, 2021). Automation in these levels is referred to as driver support features such as automatic emergency braking or lane centering. From level 3 (conditional automation) to 5 (full automation), the driver has a more passive role (SAE, 2021; Avetisyan, Ayoub, & Zhou, 2022). In these levels, the driver is not driving, and only in L3 automated vehicles will have the function to handover to the driver (SAE, 2021).

For the most part, the future of automated vehicles includes fully automated ones that do not require manual intervention from the driver (L4 and L5) (Meyer, Dokic & Müller, 2015), however technical and developmental constraints remain, especially in challenging situations (Van Brummelen et al., 2018). Moreover, takeover requests (TORs) continue to be important for investigation, and even if fully automated ones are available, it is difficult to imagine them

not including handover (Woide et al., 2022). Further signifying the importance of examining ways to improve human-AV interaction during moments of handover, which this study focuses on.

Currently, the automated car industry has been working on the jump from L2 to L3. This marks a significant jump considering that in L2 the driver is still driving the car whereas in L3 driver is no longer driving the car (SAE, 2021). As of writing this paper, Mercedes-Benz is the world's first manufacturer to receive approval from German transport authorities for its L3 Drive Pilot. Moreover, they recently announced at CES 2023 that they are the first manufacturer to receive an L3 certification in the United States of America, from the state of Nevada (Autocrypt, 2023). L3 AVs are becoming part of the transportation systems as we are conducting this research on human-AI interaction during moments of handover in L3 AV systems.

In L3 AVs, the role of the driver shifts to a passive one most of the time, becoming a "passenger" (Avetisyan, Ayoub, & Zhou, 2022). Even though AVs designed at these levels are highly automated, drivers are expected to react in time during handovers. Even with L4 and L5 vehicles, when the possibility of handovers is included in AVs, drivers will continue to use the handover function from AI to humans, especially if there is a disagreement regarding the vehicle's behavior. The interaction between human-AV during handovers remains a serious concern if drivers are out-of-the-loop, meaning they may not be fully aware of the vehicle and traffic situation when a handover is necessary. This can create a risky situation for traffic safety (Woide et al., 2022). Thus, even though the driver plays a passive role during automated driving, they should maintain sufficient awareness of the road to effectively hand over when needed.

The degree to which drivers are out of the loop can be linked to trust (see 'Trust in automation' section for definition of trust in AV systems); if drivers over-trust the system, they will no longer be engaged or aware of the situation and may show signs of sleepiness (Kundinger et al., 2019). Thus, the drivers will be out of the loop. On the other hand, if drivers

do not trust the system enough, they will not utilize the automated vehicle to its full potential (Walker et al., 2019). A lack of trust can also be linked to an increase in driver engagement (Novakazi, 2020), situation awareness (Detjen et al., 2021), and mental workload (Yousfi et al., 2021). Moreover, although, these vehicles offer a cleaner, safer, and more efficient driving experience than manual driving, trust in AVs is an important factor in the acceptance of the technology, and lower levels of trust are often associated with lower levels of acceptance (Ayoub, Yang, & Zhou, 2021). Thus, trust in automation should be considered a complex phenomenon affecting human-automation interaction and the adaption of AVs.

Trust in Automation

Trust in automation is defined as "the attitude of a user to be willing to be vulnerable to the actions of an automation based on the expectation that it will perform a particular action important to the user, irrespective of the ability to monitor or to intervene" (Körber et al., 2018, p. 19). Trust in automation relies on the user (e.g., user bias) and system characteristics (e.g., predictability) (Walker, 2021).

Three are three factors of trust in automation: "(1) human-related factors (e.g., culture, age, gender, personality, experience, workload, and knowledge about AVs), (2) automation-related factors (e.g., reliability, uncertainty, and the user interface), and (3) environmental-related factors (e.g., risk, the reputation of original equipment manufacturers)" (Ayoub, Yang & Zhou, 2021). Hoff and Bashir (2015) offered a three-layered trust model concerning the variability of trust in automation: dispositional, situational, and learned trust. Dispositional trust includes factors such as personality traits, age, gender, and culture. Situational trust is concerned with internal and external variability, a category of trust that depends on human-automation relationships in distinct contexts. Learned trust comprises pre-existing knowledge and dynamic knowledge.

The length of takeover request has been identified as an important factor that can affect trust in automation (Sheng et al., 2019; Hoff and Bashir, 2015). Yousfi et al. (2021) found that takeover duration has a significant effect on drivers' trust in AI in AVs. In their study, the participants were separated into two groups, the first group had to comply with the handover in 4 seconds while the second group had 26 seconds. Drivers had higher levels of trust when they had a longer take over duration (26 seconds). It is a key challenge to determine the effect of takeover duration on situational trust.

Moreover, trust has been outlined as a factor that affects other psychological factors during automated driving such as engagement (e.g., Woide et al., 2022), sleepiness (e.g., Kundinger et al., 2019), situation awareness (Peterson et al., 2019), mental workload (Yousfi et al., 2021), and state anxiety (e.g., Le et al., 2020).

Trust and Handover Requests as Mediators of User Experience in AVs

As we consider the user experience of autonomous vehicles, the quality of takeover and trust in automation stand out as two important factors. These factors can affect other aspects of the human-automation interaction, such as engagement, sleepiness, situation awareness, mental workload, and state anxiety. Understanding the interaction between these factors is an essential step for enhancing the overall user experience in automated vehicles.

Before the start of this thesis, we conducted a systematic literature review that examined the effect of trust in automation on driving experience during takeover requests (Appendix A). We reviewed 73 journal articles regarding the effect of trust on driver engagement, sleepiness, situation awareness, mental workload, and state anxiety. The systematic literature review found only three journal articles that investigated the relationship between trust in AI and situation awareness in automated driving, and two articles regarding the relationship between trust in AI and mental workload. These early findings suggested that more research in this area is needed to comprehend the complexity of user experience in automated vehicles. Here we will also

review non-journal articles to take a deeper look into the effects of trust and takeover request in human-automation interaction.

Engagement

Engagement is defined as the state of emotional involvement and commitment (Schaufeli, 2013). Previous research suggests that engagement is an important factor in automated driving, and it can determine compliance in unexpected and complex situations (Körber et al., 2018).

Muslim, Leung, and Itoh (2022) conducted an experiment in which the participants experienced four different situations with different instructions from an automated vehicle regarding congestion (i.e. blockage on the road) and how the system is going to handle such an event. These situations are (1) the automated vehicle adjusts the speed after detecting the congestion, (2) instructs a handover and passes the congestion, (3) asks the driver to push a button so that the vehicle can pass the congestion automatically, and (4) informs the driver and then passes the congestion automatically. It was found that even though the participants trusted the first and second systems more than the third and fourth systems, they performed better in the third and fourth situations. The third situation was found to improve driver engagement and shorten reaction time.

Handover and the information provided during handover can also affect driver engagement. If the information can be verified by the driver, engagement will be supported (Woide et al., 2022). This reveals that drivers' trust relies on perception based on how the car communicates in a certain situation (Woide et al., 2022), pointing towards the importance of situational trust. Another study conducted by Wilson et al. (2020), found further support that drivers' trust in automation can lead them to "switch off" while driving. "Passenger-type viewing" during L2 automated driving was observed, an example is one driver shutting her eyes for seconds and looking outside the side window for several seconds. Their findings provide

evidence that higher levels of trust in automation can lead to less engagement on the road (Körber et al., 2018; Hergeth et al., 2016).

Regarding the relationship between trust and engagement in automated vehicles, Hergeth et al. (2016) state two important findings: (1) over-trust can lead to lower levels of engagement, as discussed above, and (2) trust can be objectively measured with gaze behavior. Drivers who reported higher levels of trust had lower frequencies of road monitoring, and the authors not only suggested that trust can be objectively measured with gaze behavior but also that it might be more direct than other behavioral measurements. Supporting this idea further, it was also found that drivers who report higher levels of trust in automation are also more involved with non-driving-related tasks (NDRTs) than those who report lower levels of trust (Körber et al., 2018).

Sleepiness

Dinges (1995, p.4) defines sleepiness as the "neurobiological processes regulating circadian rhythms and the drive to sleep". In a study conducted by Kundinger et al. (2019), higher levels of trust in automated vehicles was found to lead to higher levels of sleepiness. Researchers have found that drivers can "accept to fall asleep due to high trust in automation" in automated vehicles. However, the current literature on the relationship between trust and sleepiness is limited and requires further research especially in L3 AVs.

Situation Awareness

Endsley (1995, p.36) defines situation awareness as "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future". Merriman et al. (2021) argue that this definition suggests situation awareness as a factor that both human and non-human agents can have.

When examining the relationship between situation awareness and trust in automated vehicles, it was found that higher levels of situation awareness can promote and moderate the

effect of trust (Petersen et al., 2019). Support for situation awareness was also found to be an important moderator of trust in automated vehicles. Another study that examined the same relationship also found supporting results that lower levels of situational awareness can lead to lower levels of trust and usability of the system (Schroeter & Steinberger, 2016).

It is clear to see that situational awareness affects trust, but how trust affects situational awareness is also crucial to understand. In a study done by Miller, Sun & Ju (2014), it was found that drivers who experience a comfortable ride in a fully autonomous vehicle (L5) report higher levels of trust compared to the drivers who had to manually control the handover from L2 level automated vehicles in risky situations. However, when a handover situation occurs, they perform worse on reaction time, signaling lower levels of situation awareness. This is in line with the findings of (Thill, Hemeren & Nilsson, 2014), suggesting that higher levels of trust in automation can predict lower levels of situation awareness. This is argued to be because of over-trust leading to over-reliance on the system, which then negatively impacts situation awareness (Endsley, 2018).

Predictability of vehicle behavior can also affect trust in automation and its effect on trust can be linked with situation awareness (Detjen et al., 2021). It was found that if the automated vehicle acts the way drivers expect it to act, it promotes trust and sharpens situation awareness.

Mental Workload

Mental workload has been defined as the amount of information processing demands during a task that an individual experiences (Sanders & McCormick, 1993). This reflects a relationship between the demands of the task/situation and the resources available to the individual (Wilson & Sharples, 2015).

An experiment done by Clement et al. (2022), investigated the relationship between trust and drivers' mental workload in automated vehicles. Their study found supporting evidence of

trust's effect on mental workload in which higher levels of trust signal lower levels of workload. However, the authors argue that in their study, drivers' driving experience and age played an important role in mental workload, thus it is important to investigate the effect of trust on mental workload in a group of people who are similar in age. Yousfi et al. (2021), investigated the effect of trust in automation on workload (physical, mental, and temporal) with a group of similar age. Their study suggests two critical findings: (1) trust in automation can be affected by the window of opportunity the system gives to the driver to react (the length of the window of opportunity has a positive impact on trust), and (2) there is a significant effect of trust on workload. According to their findings, higher levels of trust are associated with lower levels of physical, mental, and temporal workload. Additionally, the duration of time allocated for handover by the automated vehicle has a significant impact on mental workload, with longer durations leading to reduced mental workload. These results suggest that fostering trust in automated systems and allowing sufficient time for handover can contribute to reducing the mental workload experienced by drivers. These findings support the arguments of Du et al. (2019), stating that higher levels of trust can be linked with lower levels of mental workload. In another experiment, the relationship between trust, situation awareness, and the mental workload was examined (Avetisyan, Ayoub & Zhou, 2022). In their study design, it was found that explanatory instructions (i.e., "why" the car is acting) given by the automated vehicle resulted in high levels of trust among the participants. However, the participants also reported high levels of mental workload in this scenario. The authors explain that the workload can be explained by the mental energy spent on interpreting the "why" and the situation.

State Anxiety

Spielberger & Smith (1966) separate anxiety into two components: trait and state anxiety. They define trait anxiety as a person's personality characteristic. On the other hand, state anxiety is defined as temporary physiological reactions and conscious feelings of pressure,

dread, and worry about an event, state, etc. Moreover, emotional states, “a state of psychological arousal and of a cognition appropriate to this state of arousal” (Scachter & Singer, 1962, p.380), are proposed to be interlinked with trust (Dunn & Schweitzer, 2005). In line with these theories, Koo et al. (2015) examined the relationship between the information message of the semi-autonomous vehicle, trust, and anxiety. Their study found a negative correlation between anxiety and trust, especially when the drivers receive *why* information from the vehicle. Moreover, a significant correlation between state anxiety, trust, situational awareness, and role adaptation has been found (Lu et al., 2022). They argue that state anxiety should be considered an important factor in human-AI interaction in autonomous vehicles. The authors have also found a significant effect of state anxiety on situational trust (Lu et al., 2022). However, the current work on this relationship is limited and it should be investigated more deeply (Lu et al., 2022). Moreover, how situational trust can affect state anxiety should be studied as well to understand the relationship between the two in the context of human-AI interaction during handover situations.

Current Challenges and Regulations

Trust in autonomous vehicles is a crucial aspect of human-automation interaction that researchers have focused on. However, two key questions remain unanswered: (a) Can trust be calibrated and maintained to improve driver performance? And (b) how can we reliably measure trust and its impact (Walker, 2021)?

Miller et al. (2016) highlight "trust fall," which refers to the divergence between trusting behavior and self-reported trust in automated vehicles. As suggested by Walker (2021), questionnaires are popular measurements for automated driving studies, but they have two distinct limitations: (a) there might be differences between what drivers report and how they behave (trust fall), and (b) they do not provide a continuous measurement, so the real-time changes are not addressed. Walker suggests using gaze behavior and skin conductance

(electrodermal activity) as real-time measures of trust, as proposed by Hergeth et al. (2016). However, these measures may introduce noisy data and reduce the validity of data (Walker, 2021).

Nevertheless, such continuous assessment cannot be purely physiological, and the drivers will be required to perform multiple subjective assessments as outlined by the Regulation (EU) 2019/2144 of the European Parliament and of the Council of 27 November 2019 (The European Parliament and of the Council, 2019). These regulations outline the necessity for advanced systems to address driver awareness, sleepiness, and distractions in autonomous vehicles. Specifically, the regulations state that "those systems do not continuously record nor retain any data other than what is necessary for relation to the purposes for which they were collected or otherwise processed within the closed-loop system" (The European Parliament and the Council, 2019, p. 11). Thus, it is important to have such a system that only collects necessary data regarding driver drowsiness and awareness. However, what is 'necessary' is not known at the moment and it is important for human factor specialists to identify.

Exploring a New Method of Continuously Collecting Drivers' Reactions with Chatbots

Chatbots are interactive software applications that can simulate natural language conversation with humans via text or voice-based communication. They can be applied in many contexts ranging from social media platforms to home devices, serving different purposes (Borsci et al., 2022). Task-based chatbots are outstanding at information requests and responding to users (Adamopoulou & Moussiades, 2020).

In the context of autonomous driving, task-based voice chatbots can serve as a tool for continuous measurement of important factors that can affect human-machine interaction, such as driver alertness and drowsiness as outlined by the existing (EU) 2019/2144 regulation by the European Commission (The European Parliament and of the Council, 2019). Continuous

measurement of vigilance is crucial in automated driving because if something happens, it is driver's fault. However, we know from cognitive studies that it is difficult to keep vigilant after a certain amount of time of passiveness (also known as *vigilance decrement*; e.g., Dinges,1995; Kundinger et al., 2019). Moreover, voice chatbot simulations can be utilized in virtual reality (VR) studies as new material to collect continuous data, possibly offering a solution to issues and challenges mentioned in the prior section. By simulating a voice chatbot, researchers and designers should be able to collect continuous data from certain moments without interrupting the participant to fill out a survey.

Purpose of the Study

To our knowledge, previous research did not explore the relationship among situational trust, engagement, sleepiness, situation awareness, state anxiety, and mental workload over time, as suggested by a preliminary systematic review of the literature that was conducted as an internship activity before the start of the thesis (Appendix A). Unfortunately, due to a 2-motnh long delay in the development of the simulation by The BMS Lab and difficulties faced in importing the eye-tracking data, we decided not to include situation awareness and engagement as variables in the current study. So, they were excluded from the models we investigated. The current study is the first of its kind to explore a theoretical model that investigates the relationship between handover duration, situational trust, mental workload, sleepiness, state anxiety, and success in takeover under different levels of complexity in handover time windows.

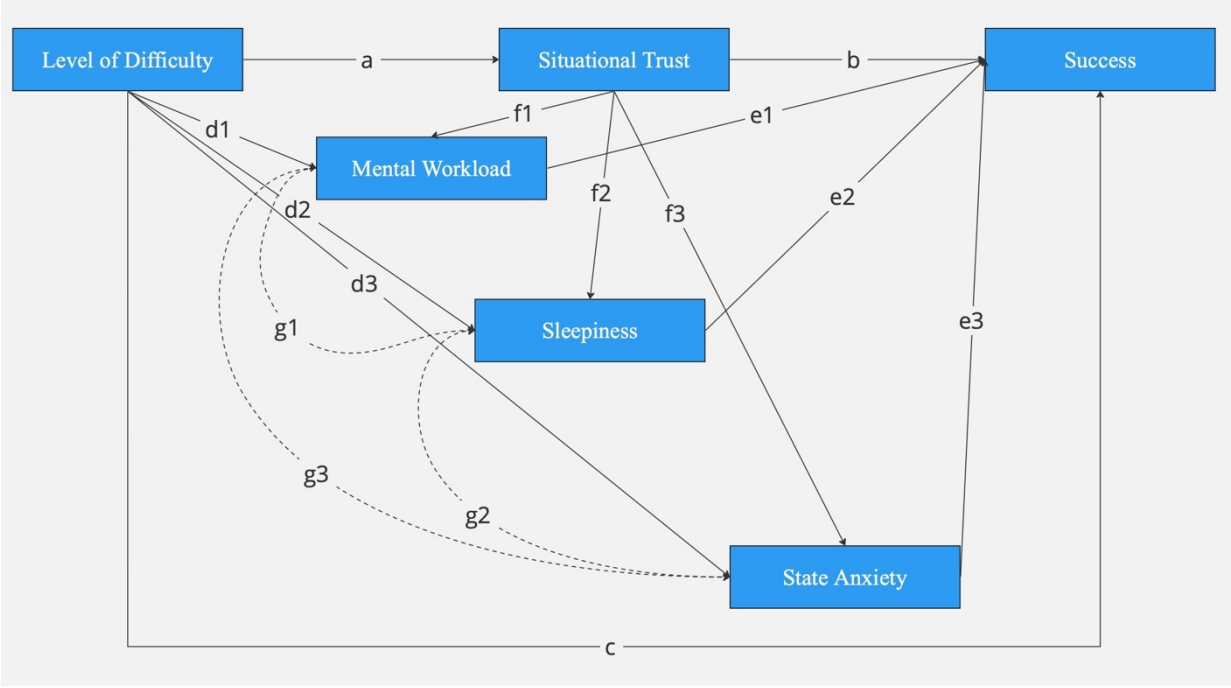
Furthermore, in virtual reality (VR) research, subjective ratings are tested by scales after the performance, or as suggested by Walker (2021) continuous measures could be designed to better track the changes in the driver experience. We explored the possibility of using a voice chatbot as a way of asking (after each main event) for changes in subjective assessment on multiple variables (situational trust, mental workload, sleepiness, and state anxiety) as opposed to only asking about these aspects after the scenarios.

We expect that (RQ1) the length of the window of opportunity to handover from AI to humans (i.e., the difficulty of taking over due to time constraints) will affect situational trust and the drivers' ability to successfully perform the requested action for taking over (success in takeover). Moreover, (RQ2) we expect that situational trust and level of handover difficulty will directly affect drivers' mental workload, sleepiness, and state anxiety and these mediators will affect drivers' success in takeover. Moreover, (RQ3) we expect to see a covariation between mental workload, sleepiness, and state anxiety. Finally, in an exploratory fashion, (RQ4) we will test if there is a significant difference between collecting drivers' subjective reactions (situational trust, mental workload, sleepiness, and state anxiety) after each scenario by forms or by a voice chatbot simulation.

The above research questions can be visually modeled as presented in Figure 1. Each relationship in the model can be formalized as a prediction as follows:

Figure 1. *Graphical presentation of expected relationships between experimental variables: Within-subject independent variable (Level of difficulty to takeover from AI), dependent variable (Success in takeover), and expected mediators (Situational trust, Workload, State anxiety, and Sleepiness). Each predicted relationship between variables (predictions) is represented by an arrow and a letter. Arrows indicate the expected direction of effect. Letters represent the following assumptions: a. Level of difficulty influences situational trust; b. Situational trust influences success in takeover; c. Level of difficulty influences success in takeover; d1. Level of difficulty influences mental workload; d2. Level of difficulty influences sleepiness; d3. Level of difficulty influences state anxiety; e1. Mental workload influences success in takeover; e2. Sleepiness influences success in takeover; e3. State anxiety influences success in takeover; f1. Situational trust influences mental workload; f2. Situational trust influences sleepiness; f3. Situational trust influences state anxiety; g1. Mental workload is*

correlated with sleepiness (exploratory); g2. Sleepiness is correlated with state anxiety (exploratory); g3. Mental workload is correlated with state anxiety (exploratory).



Note. Arrows represent the expected relationship and direction between variables. Double-sided arrows indicate covariances, while dashed arrows indicate exploratory predictions.

Our first research question is (RQ1): Does the length of the window of opportunity during handover from AI to humans (i.e., hand over time difficulty) affect drivers' situational trust, mental workload, sleepiness, state anxiety, and success in takeover?

Yousfi et al. (2021) found that handover duration has a positive impact on driver's situational trust. We expect that (Prediction 1). *The length of the window of opportunity (level of difficulty) that the automated vehicle provides during handover has a positive impact on driver's situational trust (Figure 1, relation a).* Moreover, mental workload increases when the task demand increases (Wilson & Sharples, 2015), and drivers have higher levels of mental workload during short handover scenarios (Yousfi et al., 2021). So, we predict that (Prediction 2). *The length of the window of opportunity (level of difficulty) that the automated vehicle provides during handover has a negative impact on mental workload (Figure 1, relation d1).* Moreover, it has been found that semi-automatic tasks that allow mind-wandering can cause

boredom and sleepiness (Einger, 1999), so we can predict that when drivers have a longer time to react during L3 automated driving, they will experience higher levels of sleepiness: (Prediction 3). *The length of the window of opportunity (level of difficulty) that the automated vehicle provides during handover has a positive impact on sleepiness (Figure 1, relation d2).* It has been known that state anxiety increases when one faces pressure about an event (Spielberger & Smith 1996; Lu et al., 2020), so we can expect that (Prediction 4). *The length of the window of opportunity (level of difficulty) that the automated vehicle provides during handover has a negative impact on state anxiety (Figure 1, relation d3).* Moreover, we know that the length of handover has a significant effect on compliance with the handover request (Eriksson, & Stanton, 2017; Dogan et al., 2021), so we expect (Prediction 5). *The length of the window of opportunity (level of difficulty) that the automated vehicle provides during handover has a positive impact on driving performance (Figure 1, relation c).*

Our second research question is (RQ2): Does situational trust affect drivers' mental workload, sleepiness, state anxiety, and success in takeover in scenarios of handover from AI to humans?

It has been suggested that trust has a significant relationship and a direct effect on sleepiness (Kundinger et al., 2019), mental workload (Du et al., 2019; Yousfi et al., 2021; Clement et al., 2022), and state anxiety (Koo et al., 2015). We expect that when drivers have higher levels of situational trust in AV, they will experience a sense of security and confidence, leading to lower levels of state anxiety. Furthermore, it has been suggested that over-trust in the automated system can result in an out-of-loop state in drivers (Kundinger et al., 2019), leading to poorer success in takeover. Regarding these findings, we expect the following predictions:

Prediction 6. *Situational trust in the automated vehicle has a significantly negative effect on mental workload (Figure 1, relation f1).*

Prediction 7. *Situational trust in the automated vehicle has a significantly positive effect on sleepiness (Figure 1, relation f2).*

Prediction 8. *Situational trust in the automated vehicle has a significantly negative effect on state anxiety (Figure 1, relation f3).*

Prediction 9. *Situational trust in the automated vehicle has a significantly negative effect on driving performance (Figure 1, relation b).*

Our third research question is (RQ3): Do mental workload, sleepiness, and state anxiety, mediated by handover difficulty and situational trust, affect success in takeover?

Concerning the mediating effects of handover duration and situational trust on mental workload, sleepiness, and state anxiety, we expect these factors to have an impact on driving performance:

Prediction 10: *Mental workload has a significant effect on driving performance during handover from AI to humans (Figure 1, relation e1).*

Prediction 11: *Sleepiness has a significant effect on driving performance during handover from AI to humans (Figure 1, relation e2).*

Prediction 12: *State anxiety has a significant effect on driving performance during handover from AI to humans (Figure 1, relation e3).*

Moreover, given that we are expecting mental workload, sleepiness, and state anxiety to be mediated by situational trust, we expect a correlation between these factors. Thus, we explore:

Prediction 13: *Mental workload has a significant correlation with sleepiness during handover from AI to humans (Figure 1, relation g1).*

Prediction 14: *Mental workload has a significant correlation with state anxiety during handover from AI to humans (Figure 1, relation g3).*

Prediction 15: *Sleepiness has a significant correlation with state anxiety during handover from AI to humans (Figure 1, relation g2).*

Outside of the model in Figure 1, the current study aimed to explore the usage of voice chatbot simulation as a way of collecting continuous subjective ratings compared to using a questionnaire. Thus, we explore the following research question:

The fourth research question is (RQ4): Can a voice chatbot simulation serve as an alternative to using a questionnaire for assessing subjective ratings of situational trust, mental workload, sleepiness, and state anxiety, yielding similar results?

Regarding the feasibility of having a voice chatbot as a way of continuous measurement of multiple variables in VR research, we predict that:

Prediction 16a: *There will be no significant difference between chatbot and form groups' situational trust.*

Prediction 16b: *There will be no significant difference between chatbot and form groups' mental workload.*

Prediction 16c: *There will be no significant difference between chatbot and form group's sleepiness.*

Prediction 16d: *There will be no significant difference between chatbot and form groups' state anxiety.*

With these predictions, we aimed to explore whether a voice chatbot can be utilized as a continuous measurement of subjective changes or not as an alternative to measuring subjective changes with a form.

Method

Participants and Sample Group

A total of 14 participants took part in the experiment. Two participants were selected for preliminary tests, and twelve participants were selected as the main study participants. The

participants received information about the context of the study, and they gave consent to take part in the study. The main study group consisted of 6 males (50%) and 6 females (50%), age range from 20 to 26 (M = 22.08, SD = 2.19). Participants were students at the University of Twente with a valid driving license, prior driving experiences ranging from 1 year to 8 years (M = 3.83, SD = 2.08).

The participants were randomly assigned into two groups and both groups had a balanced sex distribution:

- Form Group: Participants who used only forms to rate their subjective ratings of situational trust, mental workload, sleepiness, and state anxiety.
- Chatbot Group: Participants who used forms to rate their subjective ratings of situational trust, mental workload, sleepiness, and state anxiety after the end of each scenario, but also verbally rated the same variables after each takeover request via a simulated chatbot modality following a script read by a male researcher.

Only three participants (25%) were recruited through the University of Twente's online SONA system. These participants received 2 SONA credits after their participation. 9 participants (75%) were recruited through WhatsApp student groups. Participants recruited through WhatsApp did not receive any compensation for their participation.

The data collection for the main study started on 6 April 2023 and ended on 19 April 2023. The ethical approval of the study was granted by the Ethics Committee of the Faculty of Behavioral, Management and Social Sciences (BMS) of the University of Twente (project nr. 230068).

Design of the VR Simulation Setting

Simulation setting, task difficulty, and instructions

Instructions regarding handover requests were presented as an arrow (to the left or the right) when participants had to steer, or as a stop sign as the indication to break when

participants had to break. There were 3 different levels of difficulty: easy, moderate, and hard. Each level was repeated twice, and a Latin Square design was followed to counterbalance learning effects. The car accelerated from 0 km/h to a maximum speed of 80 km/h, with a torque of 20, during each scenario.

The level of difficulty was decided based on the literature. It was found that when drivers are not engaged in a secondary task, the time they take to comply with the handover from AI is presented ranges between 1.97 and 25.75 seconds (Median = 4.56) (Eriksson, & Stanton, 2017). Based on this finding, we purpose three levels of difficulty by changing the length of the window of opportunity for participants during the handover:

- Easy level: Participants have a 10-second handover duration.
- Moderate level: Participants have a 5-second handover duration.
- Hard level: Participants have a 2-second handover duration.

Preliminary Tests

The initial plan to inform participants about whether they successfully took over or not was to show them after each instruction in the virtual reality setting, with a green “Success” or a red “Miss”. However, in our first trial of the study, we observed that due to unknown reasons Unity was not registering the input from the steering wheel and the break. Thus, we decided to observe the participants and decide whether they took over or not manually. Moreover, the initial plan to show them their success or fail in virtual reality did not work, so we decided to inform the participants ourselves after each instruction.

In the second trial of the study, we observed that the participant successfully completed every instruction and rated similar easiness for each level of difficulty. Thus, we decided to reduce the window of opportunity during handover to 5 seconds for the easy level, 3 seconds for the moderate level, and 1 second for the hard level.

Materials and Apparatus

Hardware and Software

Driving Simulator. The driving simulator room at the BMS Lab offers a driving simulator and a VR headset (Varjo XR-3). The driving simulator includes a "Next Level Racing" chair and the Logitech G920 Driving Force which is composed of foot pedals and a steering wheel. Varjo XR-3 headset was used to track eye tracking. Due to issues in importing, eye tracking data was not used in this study.

Microsoft Surface Pro. A Microsoft Surface Pro was booked from the BMS Lab and was used as a digital form.

Varjo Base. Varjo Base is the software for the Varjo XR-3 headset. In this study, it was used to record eye tracking.

Unity. The autonomous driving simulation was created with Unity. The development of the simulation is reported in Appendix B.

Qualtrics. Qualtrics was used to collect online data for demographic information, cybersickness, perceived easiness of scenarios, situational trust, mental workload, sleepiness, state anxiety, and additional comments. The Qualtrics form can be accessed via this link: https://utwentebs.eu.qualtrics.com/jfe/form/SV_6VXAKTLJBZGMiA6.

Measurements

Our study contained several subjective measurements to measure cybersickness, sleepiness, workload, situational trust, trait anxiety, and state anxiety. The measurement questions for these constructs were adapted from the prior literature (see Appendix C). Cybersickness was assessed using the CyberSickness in Virtual Reality Questionnaire (CSQ-VR) (Kourtesis et al., 2023). Sleepiness was measured using the Karolinska Sleepiness Scale (KSS) (Åkerstedt & Gillberg (1990). Mental workload was evaluated using the Rating Scale

Mental Effort (RSME) (Zijlstra, 1993). Situational trust, trait anxiety, and state anxiety measurements were adapted from Lu et al. (2022).

Procedure

Before the start of the study, each participant received information verbally about the context of the study (Appendix D) and an explanation of the instructions they are going to receive (Appendix E). After giving their consent to participate in the study, the participants received questions regarding demographic information (age, sex, previous driving experience, trait anxiety, and cybersickness before the start of the experiment). Then, after the calibration of the VR headset, the training scenario started.

The initial training was around five minutes long and it included two instructions (stop and take the exit to the right). Each instruction had a reaction time of 25 seconds in the training phase. After the training phase, each participant followed a special sequence of scenarios that was decided by using a Latin Square.

Between each scenario, participants removed the VR headset. They were asked if they are feeling well or not, and if they need water or anything else to ensure their well-being. The participants filled in the CSQ-VR questionnaire and then, they rated the perceived easiness of the scenario, situational trust, mental workload, sleepiness, and state anxiety.

We adopted a mixed-design approach. Participants performed all the same tasks and scenarios with a counterbalanced level of difficulty in easy, moderate and hard levels (i.e., time of instruction before action). However, when it comes to filling out the subjective assessments, the participants were randomly associated with two conditions: Form-assessment after each scenario (form group) or verbal continuous assessment (chatbot group).

Participants in the chatbot condition received and answered verbally to the situational trust, mental workload, sleepiness, and state anxiety items after each task is missed or successfully completed. The participants did not remove their VR, rather they answered these

items as part of a voice chatbot experience (simulated by a male researcher). After each instruction, the voice reminded each item of the previous score and asked the participant if the score decreased or increased and how much (Appendix F). After each scenario (easy, moderate, and hard) the participants assigned to this group also received the questionnaire in written form. Participants received this questionnaire after removing their VR headset. However, this time the participants were not reminded of their previous ratings.

The form group participants received the measurement items after each scenario. The participants received the questionnaire after removing their VR headset. The participants were not reminded of their previous answers before they filled in the questionnaire.

Data Analysis

For the data cleaning and analysis, we used Microsoft Excel Version 16.72 (Microsoft, 2023) and RStudio Version 2023.03.0+386 (Rstudio Team, 2023). We used Rstudio for all the analyses. The verbal answers of the chatbot group were collected by two researchers via Microsoft Teams Version 1.6.00.1159's Excel function. Additionally, a separate Excel sheet was created to focus solely on the final verbal answers of the chat group, allowing for a comparison with the average verbal answers. There were no missing data or outliers in our sample group. Descriptive statistics were performed on participants' demographics and to provide an overview of their performances in the test.

Moreover, by a Multivariate Analysis of Variance (MANOVA) we checked for differences between the two group conditions (Form and Chatbot) in terms of age, sex, previous driving experience, trait anxiety, and cybersickness reported before the start of the experiment. As cybersickness due to the usage of VR may significantly affect performance, we checked for the changes in cybersickness. Additionally, we controlled if different levels of difficulties induced significant levels of cybersickness with a linear regression analysis.

Furthermore, we examined whether participants' perceived easiness of the scenario was influenced by the level of difficulty by conducting a linear regression analysis. The difference between chatbot group's verbal and form answers in terms of situational trust, mental workload, sleepiness, and state anxiety was checked by a MANOVA. The differences between chatbot group and form group in terms of situational trust, mental workload, sleepiness, and state anxiety were visualized by violin plots.

To test out the predicted relationships summarized in the theoretical model in Figure 1 we used the Structural Equation Modelling (SEM) approach. Given the complexity of the theoretical construct model (Figure 1) and the low number of participants involved in the experiments a bootstrapped approach, known as resampling with replacement (Awang et al., 2015), was initially used. The bootstrapping approach is recognized for its advantages in (a) transparency and ease of use, (b) non-restrictive approach, (c) wide applicability (Streukens, & Leroi-Werelds, 2016).

Moreover, we used a Bayesian approach to SEM (BSEM) because this approach provides more information regarding model fit and parameter estimates compared to classic approaches, and it increases the accuracy of the analysis with small samples sizes (Muthén & Asparouhov, 2012). Bayesian approaches to model assessment and selection, and estimations using credible intervals are considered highly valuable (Colvin, 2013). We decided to take a model selection perspective, intended as an exploratory and comparative approach that aims to identify the best model to fit the data (Preacher & Yaremych, 2022).

For our BSEM analysis we used three Markov chain Monte Carlo (MCMC) chains with the *"stan"* pre-compiled marginal approach, as set as default in *blavaan* package in R (Merkle & Rosseel, 2018; Merkle et al., 2021) to produce Bayesian estimation of the posterior distribution. "The idea behind MCMC is that the conditional distribution of one set of parameters given other sets can be used to make random draws of parameters values, ultimately

resulting in an approximation of the joint distribution of all parameters" (Muthén & Asparouhov, 2012, p. 334). To our knowledge, previous research did not explore the relationship among situational trust, engagement, sleepiness, situation awareness, state anxiety, and mental workload over time within the context of handover from AI to humans, so we used non-informative priors.

For the specific case of users of the Chatbot Group, there were three possible ways to model their subjective answers use the data from Chatbot Group. So, we created three datasets to feed the model in Figure 1:

1. Last verbal answers and average of form answers (Dataset 1),
2. Average of verbal answers and average of form answers (Dataset 2),
3. Only the form answers (Dataset 3).

We compared three different models using posterior predictive *p*-value (PPP), root mean square error of approximation (RMSEA), Watanabe-Akaike (or 'widely applicable') information criterion (WAIC) and leave-one-out information criterion (LOOIC). A PPP score around 0.5 indicates a perfect fit (Muthén & Asparouhov, 2012). While an RMSEA score below 0.08 indicates an acceptable fit, an RMSEA score lower than 0.05 indicates good fit (Salarzadeh Jenatabadi et al., 2017). WAIC and LOOIC are interpreted by comparing the scores of different models. Lower WAIC and LOOIC scores imply higher predictive accuracy (Brouwer, 2021).

In line with the rationale of the model selection approach (Preacher & Yaremych, 2022), we extended our analysis by proposing and testing an alternative model and an alternative dataset. The alternative model (Appendix G) considered the relationship between situational trust, mental workload, sleepiness, and state anxiety as covariational. On our alternative dataset approach, we combined easy and moderate levels together and compared them to hard level. By comparison of PPP, RMSEA, WAIC, and LOOIC scores, a model selection approach was performed.

Results

Descriptive analysis

The demographic characteristics of age, years of driving experience and averaged results of trait anxiety questionnaire are displayed in Table 1.

Table 1. *Demographic characteristics.*

Variable	Min	Max	Mean	SD
Age	20	26	22.08	2.19
Experience	1	8	3.833	2.08
Trait anxiety	1.75	4	3.125	2.81

Note. SD = Standard Deviation.

The differences between Chatbot Group and Form Group in terms of age, years of driving experience, trait anxiety, and cybersickness were tested by using a MANOVA analysis. There was no significant difference between the two groups regarding any of the demographic characteristics (Appendix H).

Additional checks

Manipulation Check

We conducted a linear regression analysis to examine the impact of our manipulation of the level of difficulty on participants' perception of scenario easiness. The results revealed that the perceived difficulty of the hard level was significantly higher compared to the easy level ($\beta = -2.2500$, $p < 0.05$). However, there was no significant difference in perceived easiness between the moderate level and the easy level ($\beta = -0.1250$, $p = 0.756$). These findings indicate that the manipulation during the moderate level was not successful in effectively increasing the perceived difficulty. In contrast, the hard level demonstrated a significant increase in perceived

difficulty compared to the easy level, suggesting that our manipulation during the hard level was successful in making it more challenging.

Changes in Cybersickness

To determine its potential as a confounding variable, the changes in cybersickness throughout the experiment were checked. The results indicate that the changes in the level of cybersickness remained low and stable with a mean of 1.5 and a standard deviation of 0.5. Thus, we can eliminate the possibility of cybersickness affecting the study as a confounding variable (See Appendix I for the visualization of changes in cybersickness).

To analyze the changes in cybersickness between different levels of difficulty further, a linear regression analysis was used. There was no significant difference between easy and moderate levels in terms of cybersickness (β= 0.1389, p = 0.3351). Similarly, there was also no significant difference in cybersickness between easy and hard levels (β = 0.2639, p = 0.0694). However, it might be worth noting that the moderate and hard levels showed a trend toward increased cybersickness.

Effect of Trait Anxiety on State Anxiety

To explore if trait anxiety (personality characteristic) can be considered as a confounding variable that affects state anxiety (a transient emotion), linear regression analyses were conducted on the data obtained from form and verbal answers. In both cases, there is no significant effect of trait anxiety on state anxiety (β_{form}= -0.02458, p_{form} = 0.674; β_{verbal}= -0.02458, p_{verbal} = 0.674). Thus, we can eliminate trait anxiety as a confounding variable for state anxiety.

Difference Between Chatbot Group's Form and Verbal Answers

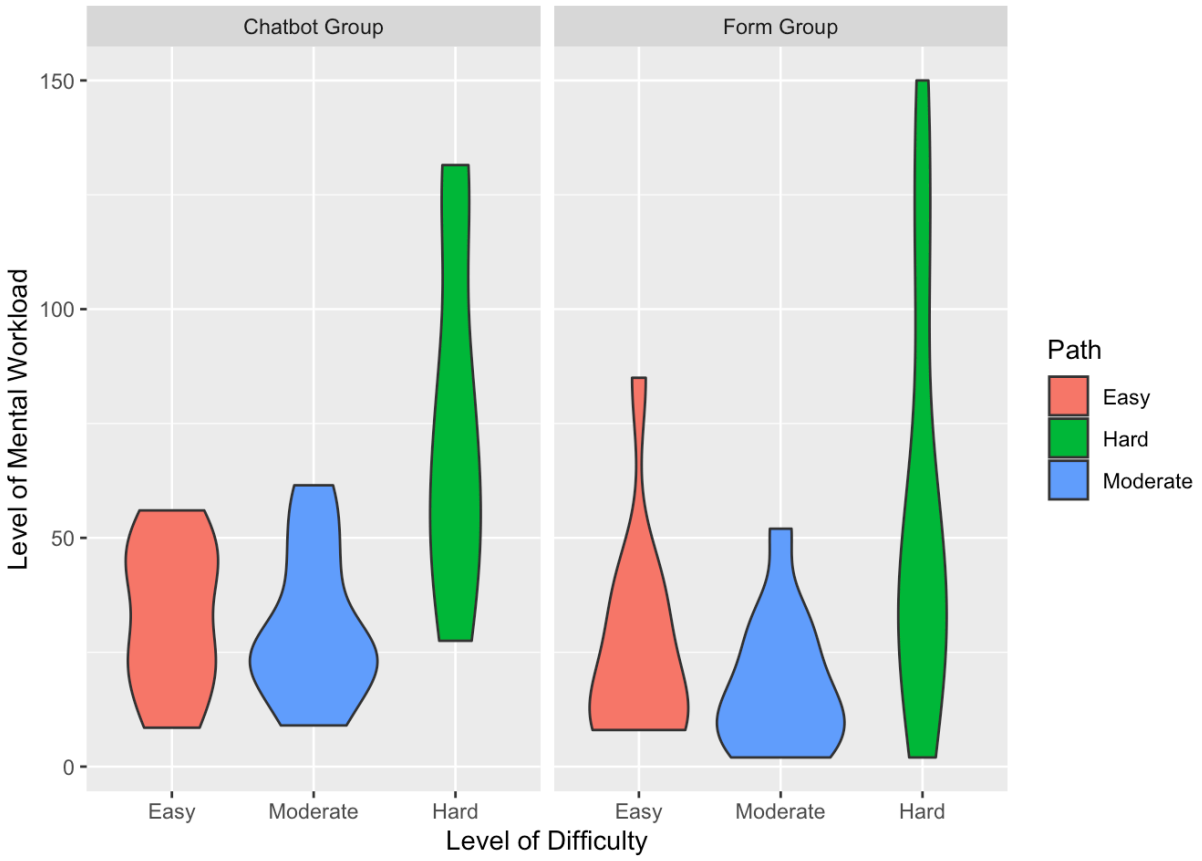
To test if the Chatbot Group evaluated their experience differently when answering the questions by a form or verbally in a simulated chatbot modality, we performed a MANOVA to compare the differences in situational trust, mental workload, sleepiness, and state anxiety

ratings collected in different modalities. The results indicate that there was no significant difference between Chatbot Group's form and verbal answers in terms of situational trust ($p = 0.6823$), mental workload ($p = 0.913$), sleepiness ($p = 0.9882$), and state anxiety ($p = 0.737$).

Differences in Situational Trust, Mental Workload, Sleepiness, and State Anxiety Across Difficulties

To visualize the changes in situational trust, mental workload, sleepiness, and state anxiety across different levels of difficulties, we generated violin plots that were grouped by Chatbot and Form Groups (Appendix J). Among the variables examined, mental workload displayed substantial variations between various levels of difficulty (see Figure 2), whereas the remaining variables demonstrated similar patterns across different difficulty levels.

Figure 2. *Changes in mental workload across different levels of difficulties grouped by Form and Chatbot group.*



Model Comparison and Prediction Testing

When it comes to the Chatbot group these participants collected data in a different way and throughout time (by multiple verbal answers to the scales, and by answering a form at the end of the scenario), compared to the form group participants that only answered to a final form after each scenario. As we explained above, albeit we formalized one theoretical construct (e.g., model, Figure 1), we can feed this model with different datasets when it comes to the chatbot group, as data were collected both verbally and by forms. Therefore, we can consider, for assessing the experience of the chatbot users the data of:

1. The last verbal answers and average of form answers (Dataset 1),
2. The average of verbal answers and average of form answers (Dataset 2),
3. Only the form answers i.e., like in the form group (Dataset 3).

Considering these three possibilities, we decided to first compare the performance in terms of fitness (PPP, RMSEA, LOOIC, WAIC) of the model when fed by these different datasets. This comparative analysis will help us select the dataset to us for the prediction testing.

Model Comparison

The performances of the three datasets in terms of model fitting are summarized in Table 2. All three models had a good fitness (PPP) with Dataset 3 resulted to be the closest to the perfect fit (PPP=0.5), concurrently all three models showed very good performances in terms of residuals (RMSEA < 0.05). Finally, Dataset 1 showed lower WAIC and LOOIC scores compared to Dataset 2 and Dataset 3. This indicates that Dataset 1 has better predictive accuracy compared to the other datasets (Vehtari et al., 2017; Brouwer, 2021). Therefore, to test our predictions, we will use the last verbal answers of the Chatbot Group and we will compare it with the form group condition.

Table 2. *Model comparison between Dataset 1 (uses the last verbal answer merged with the average of form answers for Chatbot Group's data), Dataset 2 (uses the average of verbal answers merged with an average of form answers for Chatbot Group's data), and Dataset 3 (uses form answers of Chatbot Group).*

Models	PPP	RMSEA	WAIC	LOOIC
Dataset 1	0.557	0.035	827.026	827.356
Dataset 2	0.549	0.036	827.531	828.101
Dataset 3	0.535	0.035	849.839	850.103

Note. PPP = posterior predictive *p*-value; RMSEA = root mean square error of approximation; WAIC = widely applicable information criterion; LOOIC = leave-one-out information criterion.

Prediction Testing

Looking at the key indexes of model fit, including the RMSEA, comparative fit index (CFI, Bentler, 1990), and Tucker-Lewis index (TLI, Tucker & Lewis, 1973; Bentler & Bonett, 1980), both the SEM and bootstrapped SEM analyses showed perfect fit of the model to the data (CFI = 1.000, TLI = 1.261, RMSEA = 0.000). Due to our small sample size, we interpret this as overfitting. and therefore, we decided to mainly rely for our analysis on a BSEM analysis for our prediction testing.

Table 3.a and Table 3.b summarizes the findings for Predictions 1-15. We did not find enough evidence to support our Predictions 1, 2, 3, 4, 5, 7, 9, 11, 12, 13, and 15. However, we found evidence to support the following predictions: Prediction 6 (*Situational trust in the automated vehicle has a significantly negative effect on mental workload*), 8 (*Situational trust in the automated vehicle has a significantly negative effect on state anxiety*), 10 (*Mental workload has a significant effect on driving performance during handover from AI to humans*), and 14 (*Mental workload has a significant correlation with state anxiety during handover from*

AI to humans). The model with only the significant relationships is depicted in Figure 3. It is crucial to note that these findings are only of 12 participants, so we cannot draw definitive conclusions or reject the predictions we did not find enough evidence to support.

Table 3.a. *Results for predictions 1-12. The arrows indicate the direction of the effect between variables.*

Predicted Relationships	**Coefficient Value**	**Standard Deviation**	**95% Lower Bound**	**95% Upper Bound**	**Is It Supported?**
1. Situational Trust ← Level of Difficulty	0.123	0.143	-0.163	0.396	No
2. Mental Workload ← Level of Difficulty	-0.103	0.142	-0.373	0.180	No
3. Sleepiness ← Level of Difficulty	-0.169	0.143	-0.464	0.105	No
4. State Anxiety ← Level of Difficulty	0.066	0.125	-0.177	0.311	No
5. Success in takeover ← Level of Difficulty	-0.067	0.051	-0.168	0.032	No
6. Mental Workload ← Situational Trust	-0.268	0.120	-0.498	-0.033	Yes
7. Sleepiness ← Situational Trust	0.173	0.121	-0.066	0.409	No
8. State Anxiety ← Situational Trust	-0.509	0.107	-0.720	-0.292	Yes

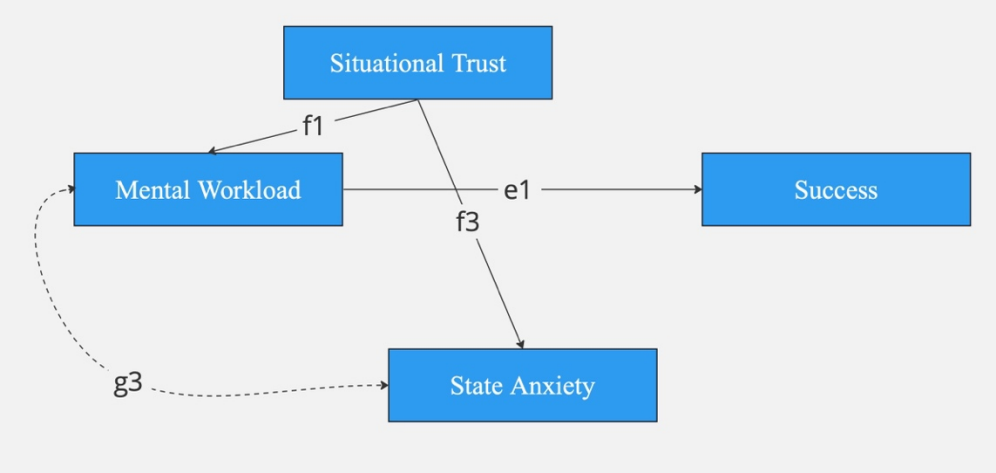
9. Success in takeover ← Situational Trust	-0.049	-0.050	-0.147	0.052	No
10. Success in takeover ← Mental Workload	-0.326	0.049	-0.423	-0.231	Yes
11. Success in takeover ← Sleepiness	-0.048	0.045	-0.137	0.039	No
12. Success in takeover ← State Anxiety	0.023	0.053	-0.081	0.125	No

Table 3.b. *Results for predictions 13-15. The double-headed arrows indicate a covariation between variables.*

Predicted Relationships	**Coefficient Value**	**Standard Deviation**	**95% Lower Bound**	**95% Upper Bound**	**Is It Supported?**
13. Mental Workload ←→ Sleepiness	0.162	0.117	-0.057	0.413	No
14. Mental Workload ←→ State Anxiety	0.314	0.107	0.123	0.542	Yes
15. Sleepiness ←→ State Anxiety	0.167	0.103	-0.028	0.376	No

Figure 3. *Graphical presentation of significant relationships between experimental variables: Each relationship between variables is represented by an arrow and a letter. Arrows indicate the direction of effect. Letters represent the following significant relationships: e1. Mental workload influences success in takeover; f1. Situational trust influences mental workload; f3.*

Situational trust influences state anxiety; g3. Mental workload is correlated with state anxiety (exploratory).



Note. Arrows represent the relationship and direction between variables. Double-sided arrow indicate covariance, and dashed arrow indicate exploratory relationship.

Furthermore, we also checked the concurrent effects of multiple variables on success in takeover (see Appendix K) in our BSEM model. This analysis indicates that:

- Mental workload (mediated by the level of difficulty and situational trust) has a significant effect on success in takeover (β = -0.298, 95%CI [-0.424, -0.172]).
- There is a significant effect of combination of mental workload, sleepiness, and state anxiety (all are mediated by the level of difficulty and situational trust) on success in takeover (β = -0.011, 95%CI [-0.610, -0.162]).
- The overall model has a significant effect (combined effects of level of difficulty, situational trust, mental workload, sleepiness, and state anxiety) on success in takeover (β = -0.386, 95%CI [-0.709, -0.208]).

When it comes to using a chatbot or a form to measure subjective changes in multiple variables in an autonomous driving context, the results indicate that there is no significant effect of administration type (chatbot or form) on measuring multiple variables. From the results, we

can accept Predictions 16a (*There will be no significant difference between chatbot and form groups' situational trust*) (β = -0.006, 95%CI [-0.441, 0.431]), 16b (*There will be no significant difference between chatbot and form groups' mental workload*) (β = 0.000, 95%CI [-0.455, 0.453]), 16c (*There will be no significant difference between chatbot and form group's sleepiness*) (β = -0.002, 95%CI [-0.480, 0.454]), 16d (*There will be no significant difference between chatbot and form groups' state anxiety*) (β = -0.000, 95%CI [-0.393, 0.394]). These findings show that there is no difference between a verbal chatbot and a questionnaire in terms of collecting subjective measurements.

Exploration of Alternative Models

To enlarge the analysis and better inform the selection model for future experiments we tested two potential alternative constructs to the one proposed in Figure 1.

Testing the Model with Covariations Between Situational Trust, Mental Workload, Sleepiness, and State Anxiety

We wanted to examine an alternative model (Appendix G) where the relationships between situational trust, mental workload, sleepiness, and state anxiety are all represented as covariations. The difference between our main model and the alternative model is that the alternative model does not propose an effect of situational trust on mental workload, sleepiness, and state anxiety but argues that there is a covariation between these variables. In Table 4 we reported only the results for the covariations of situational trust with mental workload, sleepiness, and state anxiety because all the other relationships are the same as our main model.

Table 4. *Results for covariations of situational trust with mental workload, sleepiness, and state anxiety. The arrows indicate a covariation between variables.*

Covariations	Coefficient Value	Standard Deviation	95% Lower Bound	95% Upper Bound

Situational Trust Mental Workload	←→	-0.241	0.116	-0.495	-0.022
Situational Trust Sleepiness	←→	0.166	0.122	-0.061	0.418
Situational Trust State Anxiety	←→	-0.472	0.135	-0.765	-0.237

The findings suggest that situational trust has a significant correlation with mental workload (β = -0.241, 95%CI [-0.495, -0.022]), and state anxiety (β = -0.472, 95%CI [-0.765, -0.237]). However, there is no significant correlation between situational trust and sleepiness (β = 0.166, 95%CI [-0.061, 0.418]). Moreover, we tested the effect of the overall alternative model on success in takeover. The results indicated that the alternative model does not have a significant explanation of the overall effect on success in takeover (β = -0.060, 95%CI [-0.193, 0.074]). This suggests that the alternative model does not provide a comprehensive explanation for the relationship between the variables and success in takeover. Furthermore, we compared the two models based on their model fitness (Table 5).

Table 5. *Model fitness comparison between the main model (model with the predicted effect of situational trust on mental workload, sleepiness, and state anxiety) and the alternative model (covariations between situational trust, mental workload, and state anxiety).*

Models	**PPP**	**RMSEA**	**WAIC**	**LOOIC**
Main Model	0.557	0.035	827.026	827.356
Alternative Model	0.547	0.036	827.220	828.574

Note. PPP = posterior predictive *p*-value; RMSEA = root mean square error of approximation; WAIC = widely applicable information criterion; LOOIC = leave-one-out information criterion.

The results show that the alternative model has an overall good fitness to our data (RMSEA < 0.05, PPP = 0.547). However, when we compare two models based on their WAIC and LOOIC scores it is apparent that the main model has a better predictive accuracy compared to the alternative model because it has lower WAIC and LOOIC scores. Thus, we can say that our main model does a better job of explaining the overall effect on success in takeover and it is a better fit to our data.

Testing the Model with 2 Levels of Difficulty

Earlier in this section (manipulation check) we found that the moderate level of difficulty did not have a significant difference from the easy level. This suggest that the medium level of difficulty it was not perceived difficult enough by our population. Thus, in an exploratory fashion, we wanted to check what changes if we combine easy and moderate levels together. Table 6 summarizes the predictions that were not supported by 3 levels of difficulty (Easy, moderate, and hard) but now are supported by 2 levels of difficulty (Easy + moderate and hard).

Table 6. *Predictions that were not supported with 3 levels of difficulty but are supported with 2 levels of difficulty. The arrows indicate the direction of the effect between variables.*

Predicted Relationships	**Coefficient Value**	**Standard Deviation**	**95% Lower Bound**	**95% Upper Bound**
2. Mental Workload ← Level of Difficulty	1.326	0.176	0.979	1.672
5. Success in takeover ← Level of Difficulty	-0.876	0.032	-0.939	-0.814

The model fitness analyses are summarized in Table 7. Even though the model with 2 levels of difficulty provides a good fitness in terms of PPP (PPP = 0.551), the RMSEA result indicates that the model is not a good fit because it is higher than 0.08 (RMSEA = 0.103). In contrast, the WAIC and LOOIC scores of both models suggest that the model with 2 levels of difficulty is a better fit to our data compared to the model with 3 levels of difficulty.

Table 7. *Model fitness comparison between model with 3 levels of difficulty and model with 2 levels of difficulty.*

Models	**PPP**	**RMSEA**	**WAIC**	**LOOIC**
Model with 3 Levels of Difficulty	0.557	0.035	827.026	827.356
Model with 2 Levels of Difficulty	0.551	0.103	609.398	609.890

Note. PPP = posterior predictive *p*-value; RMSEA = root mean square error of approximation; WAIC = widely applicable information criterion; LOOIC = leave-one-out information criterion.

Discussion

The study had two objectives. Firstly, it aimed to explore the relationship between handover duration, situational trust, mental workload, sleepiness, state anxiety, and success in takeover across varying levels of complexity within handover time windows. To achieve this, we developed and tested an exploratory theoretical model, and compared it to alternative models. Secondly, the research aimed to assess the relevancy of utilizing a voice chatbot simulation as a continuous measurement tool.

Our first research question addressed the effect of the length of handover on situational trust, mental workload, sleepiness, state anxiety, and success in takeover. Our results did not find any significant effect of handover duration on any of the variables, and we did not find

enough evidence to support predictions 1 – 5. These findings contradict previous studies suggesting a significant influence of handover length on situational trust (Yousfi et al., 2021), mental workload (Wilson & Sharples, 2015; Yousfi et al., 2021), and handover (Eriksson & Stanton, 2017; Dogan et al., 2021). Additionally, we anticipated an impact of handover length on sleepiness and state anxiety concerning the established findings on them in other domains (i.e., Spielberger & Smith, 1996; Einger, 1999; Lu et alç, 2020). However, it should be noted that our data analysis revealed our moderate level manipulation was not sufficiently challenging, making it indistinguishable from the easy level. Combining the easy and moderate levels demonstrated significant effects of handover duration on mental workload and success in takeover. Consequently, we suggest future studies either shorten the handover duration for moderate difficulty or eliminate the moderate level, focusing solely on two difficulty levels (easy and hard).

The second research question explored the impact of situational trust on mental workload, sleepiness, state anxiety, and success in takeover. We did not find a significant effect of situational trust on sleepiness and success in takeover. This is contrary to what was found in previous studies regarding the effect of situational trust on sleepiness, and success in takeover (Kundinger et al., 2019). However, it is crucial to acknowledge that our study is a pilot with limited sample size, thus lacking sufficient power to confidently reject predictions 7 and 9.

Nevertheless, our study found a significant effect of situational trust on mental workload during moments of handover from AI to humans in L3 automated vehicles (Table 3.a). The results indicate that situational trust has a negative effect on mental workload. This finding confirms the negative effect of situational trust on mental workload during the handover from AI to humans as found in previous studies (i.e., Du et al., 2019; Stephenson et al. 2020; Yousfi et al., 2021; Clement et al., 2022). Moreover, we found a negative effect of situational trust on state anxiety (Table 3.a). These results further build upon the previous findings in the literature

(i.e., Loo et al., 2015; Du et al., 2019). These results support the notion that exploring the connection between emotional states and trust in emotion can provide valuable insights into the psychological factors underlying trust formation and calibration (Hoff & Bashir, 2015).

Our third research question aimed to examine the effects of mental workload, sleepiness, and state anxiety on success in takeover. Our results did not find enough evidence that supports the effects of sleepiness and state anxiety on success in takeover. However, we found a significant negative effect of mental workload on success in takeover (Table 3.a), in line with previous research (Yoon & Ji, 2019; Kim et al., 2020). Notably, Kim et al. found that increased mental workload leads to increased time to handover from AI. Similarly, our study indicates that increased mental workload leads to poorer success in takeover, highlighting the negative impact of mental workload on handover time and reinforcing previous literature. These findings emphasize the impact of mental workload as a crucial factor affecting the success of handovers, and it highlights the necessity for the future of L3 automated vehicle designs to focus on reducing drivers' mental workload to a minimal level.

Moreover, our exploratory analysis found a significant covariation between mental workload and state anxiety (Table 3.b). To our knowledge, the current study is the first study to explore the relationship between mental workload and state anxiety during the handover from AI to humans in automated vehicles. This finding suggests that the relationship between mental workload and state anxiety during handovers should be further investigated in future research.

Additionally, we examined the concurrent effects of multiple variables on success in takeover. The results of our study revealed the following:

- Mental workload, mediated by handover length and situational trust, significantly influences success in takeover, demonstrating a negative impact.

- The combination of mental workload, sleepiness, and state anxiety, moderated by handover length and situational trust, also exerts a significant influence on success in takeover.
- Importantly, the overall model, which incorporates handover length, situational trust, mental workload, sleepiness, and state anxiety, significantly contributes to explaining success in takeover.

These findings suggest that our model has a significant explanation of the overall effect on success in takeover. More importantly, it shows the complexity of the underlying psychological mechanisms during handovers from AI to humans in automated vehicles. As mentioned earlier, this study is the first study that represents the first attempt to examine handover moments as a complex interaction between humans and AI and explores a theoretical model. Thus, the overall success of the model in explaining success in takeover should inspire future researchers to further investigate our model with bigger sample sizes. Moreover, concurrent effects should be considered in future research.

Through model selection, we compared different models that can be utilized for understanding the relationship between handover length, situational trust, mental workload, sleepiness, state anxiety, and success in takeover. The alternative model considered the relationship between situational trust, mental workload, sleepiness, and state anxiety as covariations (See Appendix G). Our analysis revealed that our main model (Figure 1) exhibited a better fit compared to the alternative model (Appendix G), requiring fewer information criteria to explain the relationships. Thus, the model we developed based on our predictions, represents a promising tool for future researchers.

Lastly, our fourth research question aimed to find out if a voice chatbot simulation can be used as a potential continuous measurement tool of subjective variables in automated vehicles. Our results show that there is no difference between using a verbal chatbot and a form

to collect data. This result, if confirmed with a larger group of participants, might be a crucial contribution to VR studies especially those focused on measuring repeated changes on trust and other subjective variables due to events during the interaction with automated vehicles. Certainly, as suggested by Walker (2012) questionnaires cannot provide real-time changes in variables, nevertheless we found that (simulated) verbal chatbots can be utilized as a way to continuously measure without asking participants to exit the simulation or interrupt the simulation. With this approach, we were able to ask the participants about their subjective ratings several times in one scenario, opening the possibility for researchers to collect multiple assessments.

Secondly, it is crucial to consider the Regulation (EU) 2019/2144 of the European Parliament and of the Council of 27 November 2019 (The European Parliament and of the Council, 2019) on the development of autonomous vehicles. The regulations stipulate that autonomous vehicles should be able to detect driver drowsiness and warn the driver. A voice chatbot can be a feasible solution for the autonomous vehicle industry. However, it should be noted that there can be a gap between what drivers report and what they experience (Walker, 2021). Therefore, the disparity between objective measurements and subjective data collected by voice chatbot should be investigated in future research.

Limitations and Recommendations for Future Studies

The present pilot study suffered from eight main limitations. First the sample size was small, consisting of only 12 participants, which limited the generalizability of the findings. The model developed in the study was found to be overfitting to the data, highlighting the need for larger sample sizes in future studies. Even though we employed a Bayesian approach to mitigate the drawbacks of small sample size, it has been suggested that a Bayesian SEM analysis should at least have a sample size of 200 participants (Liang et al., 2020). It might be more feasible to focus on the relationships that were found to be significant in this study. Moreover, in this study

we did not investigate the effect of distrust in AI, which can be a feasible variable for the development of this model. Additionally, investigating the linearity of predicted relationships should be considered in future studies to ensure thorough analysis of all relationships.

Second, during the experiment, we did not randomize the order of questions in forms and verbal cues. Thus, an order effect can be a possible confounding variable that we did not consider in this study. We encourage future studies to randomize the question orders and thus account for a possible order effect.

Third, our plan was to also collect objective metrics about the user experience with the car using for instance eye tracking, skin conductance, and heart rate variability, however, this was not possible due to technical issues. And we suggest future studies to complement data collection with objective data. Moreover, we aimed to measure engagement and situation awareness with eye-tracking, this can be considered in future research.

An additional limitation due to technical issues originated from the fact that the detection of success in takeover was done by two researchers manually. Future studies should collect this data objectively from the system itself as 0 and 1 codes respond respectively to fail and pass. Because of this manual detection, there might be instances where the researchers made a faulty observation by a millisecond. Moreover, a male researcher informed the participants whether they were successful in the handover or not. We advise future studies to also do this automatically after each event.

During the study, participants reported shakiness of the simulation and some even wanted to take longer breaks in between scenarios because of the dizziness caused by the shakiness. Future studies should fix the shakiness in the simulation. Furthermore, there was a slightly noticeable rectangle in the simulation that followed the gaze of the participants. This was also reported by some participants. Future researchers should also fix this issue because it can be distracting for some participants.

Finally, in our simulation, we did not use a traffic system (i.e., other cars on the road, and pedestrians). Some participants reported that their ratings of situational trust would change drastically if there were other cars on the road. Thus, future studies should consider traffic as an important factor and add it to the simulation.

Conclusion

This research contributes to examining potential underlying psychological factors that might affect human-AI interaction during moments of handover from AI to humans in L3 automated vehicles. We investigated the relationship between handover length, situational trust, mental workload, sleepiness, state anxiety, and success in taking over. We found evidence that highlights the impact of mental workload and state anxiety on people's success in taking over. Situational trust was found not to have an impact on taking over; however, it was found to moderate the impact of mental workload and state anxiety.

The results of the present pilot cannot be generalized due to a small sample size; however, it should be noted that by the mean of a model selection approach, we identified a promising model for future studies within this field.

Moreover, our preliminary data suggest that utilizing voice chatbot as a continuous measurement tool is feasible, and that collecting multiple data about people's subjective reactions to each event by voice brings to a better model fit compared to use summative measures (i.e., questionnaire) at the end of the test. This finding has implications for the future of VR studies, but also for the design of autonomous vehicles. Future researchers, as well as designers, should consider voice chatbots as a possible solution for enabling continuous measurement of subjective variables during autonomous driving.

References

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. Machine Learning with Applications, 2, 100006. https://doi.org/10.1016/j.mlwa.2020.100006

Åkerstedt, T., Anund, A., Axelsson, J., & Kecklund, G. (2014). Subjective sleepiness is a sensitive indicator of insufficient sleep and impaired waking function. *Journal of Sleep Research*, *23*(3), 242–254. https://doi.org/10.1111/jsr.12158

Åkerstedt, T., & Gillberg, M. (1990). Subjective and Objective Sleepiness in the Active Individual. *International Journal of Neuroscience*, *52*(1–2), 29–37. https://doi.org/10.3109/00207459008994241

Ansari, S., Naghdy, F., & Du, H. (2022). Human-Machine Shared Driving: Challenges and Future Directions. *IEEE Transactions on Intelligent Vehicles*, *7*(3), 499–519. https://doi.org/10.1109/TIV.2022.3154426

Autocrypt (2023, January 13). The State of Level 3 Autonomous Driving in 2023: Ready for the Mass Market? https://autocrypt.io/the-state-of-level-3-autonomous-driving-in-2023/

Avetisyan, L., Ayoub, J., & Zhou, F. (2022). Investigating explanations in conditional and highly automated driving: The effects of situation awareness and modality. *Transportation Research Part F: Traffic Psychology and Behaviour*, *89*, 456–466. https://doi.org/10.1016/j.trf.2022.07.010

Awang, Z., Wan Afthanorhan, W. M. A., & Asri, M. A. M. (2015). Parametric and Non Parametric Approach in Structural Equation Modeling (SEM): The Application of Bootstrapping. *Modern Applied Science*, *9*(9). https://doi.org/10.5539/mas.v9n9p58

Axelsson, J., Ingre, M., Kecklund, G., Lekander, M., Wright, K. P., & Sundelin, T. (2020). Sleepiness as motivation: a potential mechanism for how sleep deprivation affects behavior. *Sleep*, *43*(6). https://doi.org/10.1093/sleep/zsz291

Ayoub, J., Yang, X. J., & Zhou, F. (2021). Modeling dispositional and initial learned trust in automated vehicles with predictability and explainability. *Transportation Research Part F: Traffic Psychology and Behaviour*, *77*, 102–116. https://doi.org/10.1016/j.trf.2020.12.015

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Bolton, M., Biltekoff, E., & Humphrey, L. (2022). The Level of Measurement of Subjective Situation Awareness and Its Dimensions in the Situation Awareness Rating Technique (SART). *IEEE Transactions on Human-Machine Systems*, *52*(6), 1147–1154. https://doi.org/10.1109/THMS.2021.3121960

Borowsky, A., Zangi, N., & Oron-Gilad, T. (2022). Interruption Management in the Context of Take-Over-Requests in Conditional Driving Automation. *IEEE Transactions on Human-Machine Systems*, *52*(5), 1015–1024. https://doi.org/10.1109/THMS.2022.3194006

Borsci, S., Malizia, A., Schmettow, M., van der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2022). The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents. *Personal and Ubiquitous Computing*, *26*(1), 95–119. https://doi.org/10.1007/s00779-021-01582-9

Brouwer, A. (2021). Bayesian Structural Equation Modeling: Explained and Applied to Educational Science.

Clark, H., McLaughlin, A. C., & Feng, J. (2017). Situational Awareness and Time to Takeover: Exploring an Alternative Method to Measure Engagement with High-Level Automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 1452–1456. https://doi.org/10.1177/1541931213601848

Clement, P., Veledar, O., Könczöl, C., Danzinger, H., Posch, M., Eichberger, A., & Macher, G. (2022). Enhancing Acceptance and Trust in Automated Driving trough Virtual Experience on a Driving Simulator. *Energies*, *15*(3), 781. https://doi.org/10.3390/en15030781

Colvin, K. F. (2013). Kruschke, J. K. (2011). Doing Bayesian Data Analysis: A Tutorial with R and BUGS. Burlington, MA: Academic Press. *Journal of Educational Measurement*, *50*(4), 469–471. https://doi.org/10.1111/jedm.12029

Detjen, H., Salini, M., Kronenberger, J., Geisler, S., & Schneegass, S. (2021). Towards Transparent Behavior of Automated Vehicles. *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, 1–12. https://doi.org/10.1145/3447526.3472041

DINGES, D. F. (1995). An overview of sleepiness and accidents. *Journal of Sleep Research*, *4*, 4–14. https://doi.org/10.1111/j.1365-2869.1995.tb00220.x

Dogan, E., Yousfi, E., Bellet, T., Tijus, C., & Guillaume, A. (2021). Manual takeover after highly automated driving. *European Conference on Cognitive Ergonomics 2021*, 1–6. https://doi.org/10.1145/3452853.3452880

Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A. K., Yang, X. J., & Robert, L. P. (2019). Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation Research Part C: Emerging Technologies*, *104*, 428–442. https://doi.org/10.1016/j.trc.2019.05.025

Endsley, M. R. (2019). *Situation Awareness in Future Autonomous Vehicles: Beware of the Unexpected* (pp. 303–309). https://doi.org/10.1007/978-3-319-96071-5_32

Eriksson, A., & Stanton, N. A. (2017). Takeover Time in Highly Automated Vehicles: Noncritical Transitions to and From Manual Control. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*(4), 689–705. https://doi.org/10.1177/0018720816685832

The European Parliament and of the Council. (2019). Regulation (EU) 2019/2144 of the European Parliament and of the Council of 27 November 2019 on type-approval requirements for motor

vehicles and their trailers, and systems, components and separate technical units intended for such vehicles, as regards their general safety and the protection of vehicle occupants and vulnerable road users, amending Regulation (EU) 2018/858 of the European Parliament and of the Council and repealing Regulations (EC) No 78/2009, (EC) No 79/2009 and (EC) No 661/2009 of the European Parliament and of the Council and Commission Regulations (EC) No 631/2009, (EU) No 406/2010, (EU) No 672/2010, (EU) No 1003/2010, (EU) No 1005/2010, (EU) No 1008/2010, (EU) No 1009/2010, (EU) No 19/2011, (EU) No 109/2011, (EU) No 458/2011, (EU) No 65/2012, (EU) No 130/2012, (EU) No 347/2012, (EU) No 351/2012, (EU) No 1230/2012 and (EU) 2015/166.

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2016). Keep Your Scanners Peeled. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 509–519. https://doi.org/10.1177/0018720815625744

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

Holthausen, B. E., Wintersberger, P., Walker, B. N., & Riener, A. (2020). Situational Trust Scale for Automated Driving (STS-AD): Development and Initial Validation. *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 40–47. https://doi.org/10.1145/3409120.3410637

Hopkins, D., & Schwanen, T. (2021). Talking about automated vehicles: What do levels of automation do? *Technology in Society*, *64*, 101488. https://doi.org/10.1016/j.techsoc.2020.101488

Kim, J., Kim, H.-S., Kim, W., Lee, S.-J., & Yoon, D. (2020). Investigation on the Effect of Mental Workload on the Time-related Take-over Performance. *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, 1912–1917. https://doi.org/10.1109/ICTC49870.2020.9289513

Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., & Nass, C. (2015). Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, *9*(4), 269–275. https://doi.org/10.1007/s12008-014-0227-2

Körber, M., Baseler, E., & Bengler, K. (2018). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics*, *66*, 18–31. https://doi.org/10.1016/j.apergo.2017.07.006

Kourtesis, P., Linnell, J., Amir, R., Argelaguet, F., & MacPherson, S. E. (2023). Cybersickness in Virtual Reality Questionnaire (CSQ-VR): A Validation and Comparison against SSQ and VRSQ. *Virtual Worlds*, *2*(1), 16–35. https://doi.org/10.3390/virtualworlds2010002

Kundinger, T., Wintersberger, P., & Riener, A. (2019). (Over)Trust in Automated Driving. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3290607.3312869

Liang, X., Yang, Y., & Cao, C. (2020). The Performance of ESEM and BSEM in Structural Equation Models with Ordinal Indicators. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(6), 874–887. https://doi.org/10.1080/10705511.2020.1716770

Lu, Y., Yi, B., Song, X., Zhao, S., Wang, J., & Cao, H. (2022). Can we adapt to highly automated vehicles as passengers? The mediating effect of trust and situational awareness on role adaption moderated by automated driving style. *Transportation Research Part F: Traffic Psychology and Behaviour*, *90*, 269–286. https://doi.org/10.1016/j.trf.2022.08.011

Merkle, E. C., Fitzsimmons, E., Uanhoro, J., & Goodrich, B. (2021). Efficient Bayesian Structural Equation Modeling in *Stan*. *Journal of Statistical Software*, *100*(6). https://doi.org/10.18637/jss.v100.i06

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian Structural Equation Models via Parameter Expansion. *Journal of Statistical Software*, *85*(4). https://doi.org/10.18637/jss.v085.i04

Merriman, S. E., Plant, K. L., Revell, K. M. A., & Stanton, N. A. (2021). Challenges for automated vehicle driver training: A thematic analysis from manual and automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *76*, 238–268. https://doi.org/10.1016/j.trf.2020.10.011

Merritt, S. M. (2011). Affective Processes in Human–Automation Interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(4), 356–370. https://doi.org/10.1177/0018720811411912

Meyer, G., Dokic, J., & Müller, B. (2015). *Elements of a European Roadmap on Smart Systems for Automated Driving* (pp. 153–159). https://doi.org/10.1007/978-3-319-19078-5_13

Microsoft. (2022). Microsoft Excel. https://www.microsoft.com/nl-nl/microsoft-365/excel/

Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., & Ju, W. (2016). Behavioral Measurement of Trust in Automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 1849–1853. https://doi.org/10.1177/1541931213601422

Miller, D., Sun, A., & Ju, W. (2014). Situation awareness with different levels of automation. *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 688–693. https://doi.org/10.1109/SMC.2014.6973989

Moore, K., & Gugerty, L. (2010). Development of a Novel Measure of Situation Awareness: The Case for Eye Movement Analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(19), 1650–1654. https://doi.org/10.1177/154193121005401961

Muslim, H., Kiu Leung, C., & Itoh, M. (2022). Design and evaluation of cooperative human–machine interface for changing lanes in conditional driving automation. *Accident Analysis & Prevention*, *174*, 106719. https://doi.org/10.1016/j.aap.2022.106719

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. https://doi.org/10.1037/a0026802

Novakazi, F., Orlovska, J., Bligård, L.-O., & Wickman, C. (2020). Stepping over the threshold linking understanding and usage of Automated Driver Assistance Systems (ADAS). *Transportation Research Interdisciplinary Perspectives*, *8*, 100252. https://doi.org/10.1016/j.trip.2020.100252

Petersen, L., Robert, L., Yang, J., & Tilbury, D. (2019). Situational Awareness, Driver's Trust in Automated Driving Systems and Secondary Task Performance. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3345543

Preacher, K. J., & Yaremych, H. E. (2022). Model selection in structural Equation Modeling. Handbook of Structural Equation Modeling, 206.

R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.r-project.org/

SAE. (2021). Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. SAE International in United States, J3016_202104.

Salarzadeh Jenatabadi, H., Moghavvemi, S., Wan Mohamed Radzi, C. W. J. B., Babashamsi, P., & Arashi, M. (2017). Testing students' e-learning via Facebook through Bayesian structural equation modeling. *PLOS ONE*, *12*(9), e0182311. https://doi.org/10.1371/journal.pone.0182311

Sanders, M. S., & McCormick, E. J. (1993). Human factors in engineering and design. New York: McGraw-Hill.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379–399. https://doi.org/10.1037/h0046234

Schroeter, R., & Steinberger, F. (2016). Pokémon DRIVE. *Proceedings of the 28th Australian Conference on Computer-Human Interaction - OzCHI '16*, 25–29. https://doi.org/10.1145/3010915.3010973

Sheng, S., Pakdamanian, E., Han, K., Kim, B., Tiwari, P., Kim, I., & Feng, L. (2019). A Case Study of Trust on Autonomous Driving. *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 4368–4373. https://doi.org/10.1109/ITSC.2019.8917251

Spielberger, C. D., Gonzalez-Reigosa, F., Martinez-Urrutia, A., Natalicio, L. F., & Natalicio, D. S. (1971). The state-trait anxiety inventory. Revista Interamericana de Psicologia/Interamerican Journal of Psychology, 5(3 & 4).

Steiger, J. H. (1990). Structural Model Evaluation and Modification: An Interval Estimation Approach. *Multivariate Behavioral Research*, *25*(2), 173–180. https://doi.org/10.1207/s15327906mbr2502_4

Steiger, J. H., & Lind, J. C. (1980). Statistically based tests for the number of common factors. Paper presented at the Annual Meeting of the Psychometric Society, Iowa City, IA.

Stephenson, A. C., Eimontaite, I., Caleb-Solly, P., Morgan, P. L., Khatun, T., Davis, J., & Alford, C. (2020). Effects of an Unexpected and Expected Event on Older Adults' Autonomic Arousal and Eye Fixations During Autonomous Driving. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.571961

Streukens, S., & Leroi-Werelds, S. (2016). Bootstrapping and PLS-SEM: A step-by-step guide to get more out of your bootstrap results. *European Management Journal*, *34*(6), 618–632. https://doi.org/10.1016/j.emj.2016.06.003

Thill, S., Hemeren, P. E., & Nilsson, M. (2014). The apparent intelligence of a system as a factor in situation awareness. *2014 IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 52–58. https://doi.org/10.1109/CogSIMA.2014.6816540

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10. https://doi.org/10.1007/BF02291170

Van Brummelen, J., O'Brien, M., Gruyer, D., & Najjaran, H. (2018). Autonomous vehicle perception: The technology of today and tomorrow. *Transportation Research Part C: Emerging Technologies*, *89*, 384–406. https://doi.org/10.1016/j.trc.2018.02.012

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Victor, T. W., Tivesten, E., Gustavsson, P., Johansson, J., Sangberg, F., & Ljung Aust, M. (2018). Automation Expectation Mismatch: Incorrect Prediction Despite Eyes on Threat and Hands on Wheel. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *60*(8), 1095–1116. https://doi.org/10.1177/0018720818788164

Vlakveld, W., van Nes, N., de Bruin, J., Vissers, L., & van der Kroft, M. (2018). Situation awareness increases when drivers have more time to take over the wheel in a Level 3 automated car: A simulator study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *58*, 917–929. https://doi.org/10.1016/j.trf.2018.07.025

Walker, F. (2021). *To trust or not to trust?: assessment and calibration of driver trust in automated vehicles* [University of Twente]. https://doi.org/10.3990/1.9789055842766

Walker, F., Wang, J., Martens, M. H., & Verwey, W. B. (2019). Gaze behaviour and electrodermal activity: Objective measures of drivers' trust in automated vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour*, *64*, 401–412. https://doi.org/10.1016/j.trf.2019.05.021

Weinger, M. B. (1999). Vigilance, boredom, and sleepiness. *Journal of Clinical Monitoring and Computing*, *15*(7/8), 549–552. https://doi.org/10.1023/A:1009993614060

Wilson, J. R., & Sharples, S. (Eds.). (2015). *Evaluation of Human Work*. CRC Press. https://doi.org/10.1201/b18362

Wilson, K. M., Yang, S., Roady, T., Kuo, J., & Lenné, M. G. (2020). Driver trust & mode confusion in an on-road study of level-2 automated vehicle technology. *Safety Science*, *130*, 104845. https://doi.org/10.1016/j.ssci.2020.104845

Woide, M., Colley, M., Damm, N., & Baumann, M. (2022). Effect of System Capability Verification on Conflict, Trust, and Behavior in Automated Vehicles. *Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 119–130. https://doi.org/10.1145/3543174.3545253

Yoon, S. H., & Ji, Y. G. (2019). Non-driving-related tasks, workload, and takeover performance in highly automated driving contexts. *Transportation Research Part F: Traffic Psychology and Behaviour*, *60*, 620–631. https://doi.org/10.1016/j.trf.2018.11.015

Yousfi, E., Malin, S., Halit, L., Roger, S., & Dogan, E. (2021). Driver experience and safety during manual intervention in a simulated automated vehicle. *European Conference on Cognitive Ergonomics 2021*, 1–7. https://doi.org/10.1145/3452853.3452888

Zhang, Y., Ma, J., Pan, C., & Chang, R. (2021). Effects of automation trust in drivers' visual distraction during automation. *PLOS ONE*, *16*(9), e0257201. https://doi.org/10.1371/journal.pone.0257201

Zijlstra, F. R. H. (1993). Efficiency in work behaviour. *A design approach for modern tools*.

Appendix A – Systematic Literature Review

Introduction

Automated vehicles have emerged as a new and promising approach to transportation. However, the development of fully automated vehicles is still facing challenges, and technical limitations remain in addressing challenging situations (e.g., bad weather) (Van Brummelen et al., 2018). As a result, takeover requests (TORs) remain an important aspect of automated vehicle design, even if fully automated vehicles become available (Woide et al., 2022). This makes investigating the role of human-automation interaction crucial.

Automated vehicles are classified into six levels of automation by the Society of Automotive Engineers (SAE, 2021). L0, L1, and L2 level of automation is referred to as “driver support features” and L3, L4, and L5 are referred to as “automated driving features” (SAE, 2021).

Previous studies have identified that trust in automated vehicles is an important variable that can be linked to driving experience, especially during takeover requests (TORs) (e.g., Kundinger et al., 2019; Walker et al., 2019; Novakazi, 2020; Detjen et al., 2021). For example, if drivers over-trust the system, they may not be aware of the situation and show signs of sleepiness (Kundinger et al., 2019), while a lack of trust can lead to reduced utilization of the automated vehicle's full potential (Walker et al., 2019).

Thus, the interaction between trust in AI and driving experience is our area of interest. This systematic literature review aims to identify previous works that investigated the relationship between trust in AI and driver engagement, sleepiness, situation awareness, state anxiety, and mental workload.

Method

Study design

This systematic literature review followed the PRISMA guidelines (Page et al., 2021) to identify the journal articles investigating the relationship between trust and several variables (engagement, sleepiness, situation awareness, state anxiety, and mental workload) during automated driving from 1985 until 2023.

Research questions

To identify the effect on trust in automation on driving experience during TOR, this systematic literature review focused on (1) driver engagement, (2) sleepiness, (3) situation awareness, (4) state anxiety, and (5) mental workload. Thus, this review sought to answer the following research questions:

- RQ1 – Is there a relationship between trust in AI and engagement on automated driving during TOR?
- RQ2 – Is there a relationship between trust in AI and sleepiness on automated driving during TOR?
- RQ3 – Is there a relationship between trust in AI and situation awareness on automated driving during TOR?
- RQ4 – Is there a relationship between trust in AI and state anxiety on automated driving during TOR?
- RQ5 – Is there a relationship between trust in AI and mental workload on automated driving during TOR?

Eligibility criteria

The inclusion criteria for this review included articles that:

- focused on the relationship between trust in AI and various variables (engagement, sleepiness, situation awareness, state anxiety, and mental workload) during automated driving in automated cars.

In the review, the excluded articles were that of:

1. not focusing on the relationship between trust and various variables but rather reporting them as separate variables.
2. using different types of automated vehicles than of cars (e.g., automated busses, marines, aircrafts).
3. literature reviews.
4. not using empirical research methods.
5. focusing on situation awareness of automated driving system, not the driver.
6. focuses on a new prototype.

Search strategy

The search for this review included four databases SCOPUS, IEEE XPLORE, PsycInfo, and Web of Science. The Boolean search strings were:

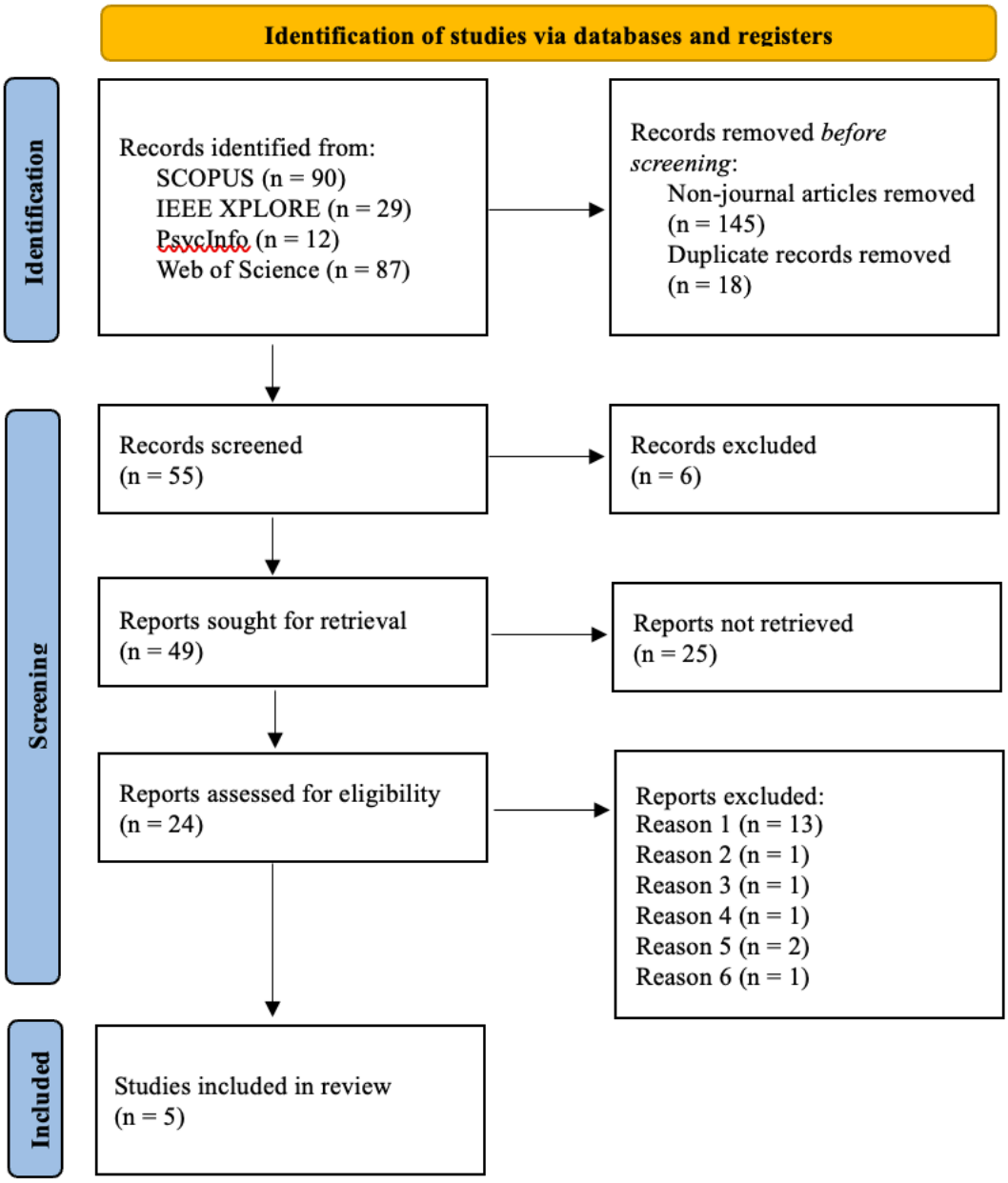
- Engagement:
    a. “autonomous vehicle” AND “trust” AND “engagement”
- Sleepiness:
    a. “autonomous vehicle” AND “trust” AND “sleepiness”
    b. “autonomous vehicle” AND “trust” AND “drowsiness”
- Situation Awareness:
    a. “autonomous vehicle” AND “trust” AND “situation awareness”
- State Anxiety:
    a. “autonomous vehicle” AND “trust” AND “state anxiety”
- Mental Workload:
    a. “autonomous vehicle” AND “trust” AND “mental workload”
    b. “autonomous vehicle” AND “trust” AND “cognitive load”
    c. “autonomous vehicle” AND “trust” AND “cognitive workload”

Results

The exclusion process of this reviews is as shown in Figure 4. A total of 218 records were identified at the first phase. 73 journal articles were identified by using the filters on the databases. Then, 18 duplicates were removed. After the removal of duplicates, the remaining articles' titles were scanned and six articles that mentioned different automated vehicles in their titles (e.g., automated busses, marines, aircrafts) were excluded. The abstracts of the remaining articles were examined, and 25 articles were removed after the abstract scan. Finally, after the full text scan 19 articles were rejected and only five articles were accepted for the review. The number of articles used in the review per variable were:

- Trust in AI and engagement: n = 0
- Trust in AI and sleepiness: n = 0
- Trust in AI and situation awareness: n = 3
- Trust in AI and state anxiety: n = 0
- Trust in AI and mental workload: n = 2

Figure 4. *The systematic literature review process in line with the flowchart of the PRISMA guidelines (Page et al., 2021).*



Discussion

Situation Awareness (n=3)

Borowsky et al. (2022), found that trust can affect the takeover strategy drivers employ. In their study, they tested four take over strategies in automated vehicles while the drivers were engaged with a secondary task (Simon game): (1) take control and continue the secondary task, (2) take control and abandon the secondary task, (3) postpone taking over, finish secondary task and then take control, and (4) reject TOR, continue the secondary task and wait for automatic deactivation of the vehicle. They found that drivers who report high level of trust in AI employ

strategy 4 more than they employ any other strategy. They argue that with high trust drivers feel more confident in AI so that they reject TOR and continue secondary task. They also report that some participants who employed strategy 4 not only rejected TOR but they also rejected engagement with secondary task. It is argued that these participants showed more situation awareness, and they were curious of how would the automation behave. However, this study shows weak connection specifically between trust and situation awareness in the context of automated driving.

Vlakveld et al. (2018) tested how situation awareness affects driver behavior during TOR in L3 automated vehicles. They found that driver characteristics (age, driving experience, sensation seeking, and trust) do not have a significant effect on situation awareness. Thus, they suggest that trust in AI does not have a significant impact on drivers' situation awareness during TOR in L3 automated vehicles. However, they defined situation awareness as the extent in which the drivers are aware of potential hazards on the road in the driving simulation. In their simulation design, TOR was simulated as a red text on the right side of the instrument cluster and as a sound. Thus, their definition did not include the drivers' awareness of the TOR itself but rather of hazardous items/events on the road. A more in depth approach to drivers' situation awareness of TOR and how trust affects this relationship is needed.

Another study that investigated the relationship between trust and situation awareness Victor et al. (2018) suggests that overtrust can affect driver performance during TOR. They found that even though some drivers had their hands on the steering wheel and eyes on potential threat (signaling high level of situation awareness) they still failed to react in time and crashed. The researchers suggest that this was because they expected and trusted the AI to act. They found that overtrust predicts poor TOR response. Moreover, the researchers underline the difference between *initial learned trust* (trust in automation before interacting with the system) and *dynamic learned trust* (trust in automation during interaction with the system) (Hoff and

Bashir, 2015), and suggest that dynamic learned trust can override initial learned trust. They suggest further research into the relationship between dynamic learned trust and situation awareness during TOR in automated driving.

Mental Workload (n=2)

When it comes to the effect of trust in AI on drivers' mental workload during TOR in automated driving, it was found that drivers who have higher levels of initial learned trust in AI have lower levels of mental workload and they pay more attention to non-driving-related tasks (though the authors do not mention any relation to situation awareness in their paper) (Zhang et al., 2021).

Moreover, Stephenson et al. (2020), employed eye tracking, skin conductance and heart rate to investigate the effect of trust in AI in L5 automated driving simulation. They found that during unexpected events (i.e., unexpected stop) participants had increased skin conductance. However, these findings are difficult to interpret in terms of mental workload because the authors did not investigate attentional capacity, thus it can not be directly linked to mental workload. Additionally, this study was focused on elderly population, and it was done with a L5 automated driving simulation, which are outside of the context of this research project.

Conclusion

This systematic literature review only found three journal articles that investigated the relationship between trust in AI and situation awareness in automated driving. Furthermore, we found only two articles related to relationship between trust in AI and mental workload during automated driving. Suggesting that more research in this area is also needed. When it comes to the effect of trust on driver engagement, sleepiness, and state anxiety, no journal articles were found. Further implying the significance of the Master's Thesis continuing this internship.

Appendix B – Development Report

The simulation was developed within close cooperation with The BMS Lab and another Master's Human Factors and Engineering Psychology student İlkyaz Çağgül Armağan Arslankaya. My responsibilities as a part of the Development Team were:

- Designing the experiment with the internal supervisor and within the agreement with APPLUS IDIADA. This was done by 8 meetings with internal supervisor and 2 meeting with the representative from APPLUS IDIADA. There were also several meetings with other members of The BMS Lab regarding assessing the methodology of the proposed project and equipment use.
- Getting started with Unity; creating the terrains for the simulation and altering the terrains with a team member for a more realistic environment that blocks certain parts of the road so that the participants do not understand that they are driving in a loop.

Figure 5. *Simulation in Unity.*

Figure 6. *Simulation map in Unity.*

- Regular follow-up meetings with the Development Team of The BMS Lab to create a schedule for certain steps in the project and updating each other on the progress.
- For data tracking, the initial plan was to connect Unity with iMotions. To achieve that, I looked up for manuals and instructions online but there was not much information available at that time. So, I contacted with iMotions and Unity teams. Then I worked with the Development Team at The BMS Lab. However, to do that 2020 Version of Unity was necessary but the simulation was in 2021 Unity Version. To combat this issue, I connected Varjo Base with iMotions, and later learned how to put VR recording from Varjo Base into iMotions so that I can create AOIs and analyze eye-tracking data.
- When testing the Varjo Base, we found out that the PC at the driving simulator room was not able to handle the CPU requirement of Varjo XR-3. To combat this, I tested out

a laptop and another PC at The BMS Lab to find the proper computer to run the experiment on. This also included setting up Varjo Base and connecting Varjo XR-3 on several devices.

- Moreover, I tried to connect Shimmer3 GSR+ with iMotions, unfortunately due to issues with The BMS Lab my access to Shimmer module in iMotions license was postponed until the end of February. As an alternative, I started working with Empatica E4 wristband tracker. Right now, I am currently working on connecting Empatica E4 with iMotions.
- I also created the paths that the autonomous car will follow in the simulation. I created event points for instructions, reaction times, and UI elements in the scenario (Arrows for taking an exit, “Success”/ “Miss” after each event, etc.).
- The future activities include connecting Empatica E4/Shimmer3 GSR+ with iMotions and testing whether GSR and heart rate data can be synced reliably with the eye-tracking data or not, finalizing the simulation and pilot testing the project.
- Moreover, a manual explaining how to connect iMotions with Shimmer and Varjo Base and how to perform eye-tracking, GSR, and heart rate variability analysis on iMotions will be created during the Master’s Thesis and will be provided to The BMS Lab.

Figure 7. *Autonomous driving path in Unity.*

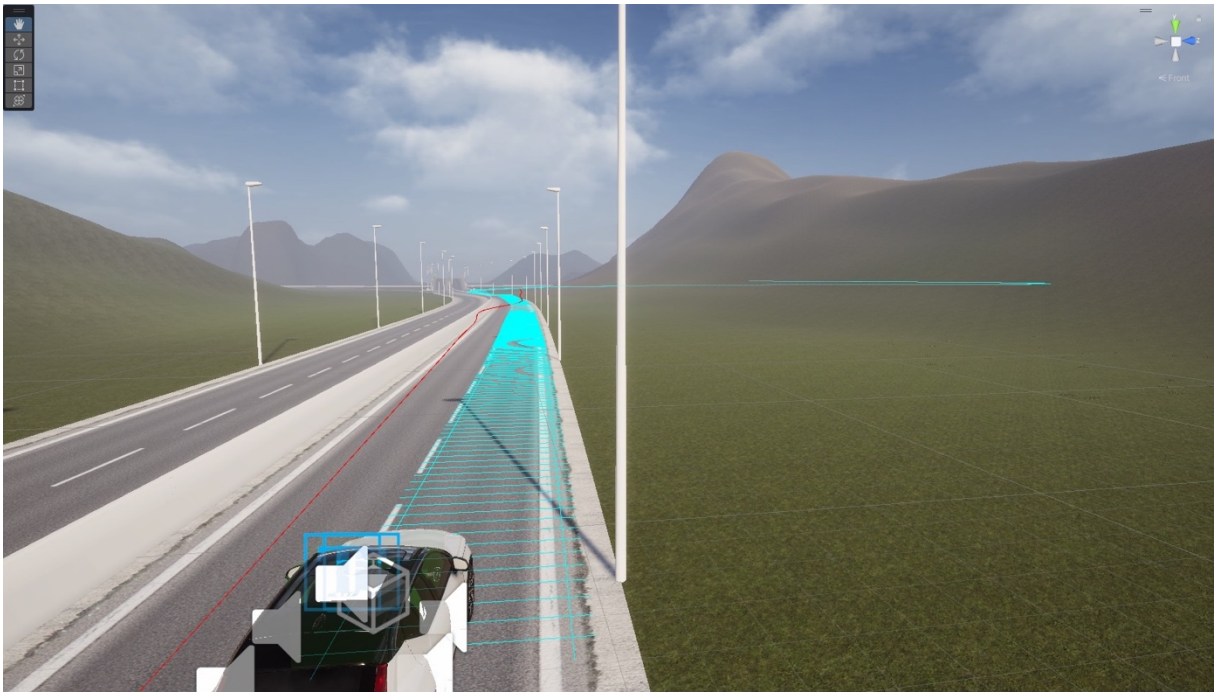
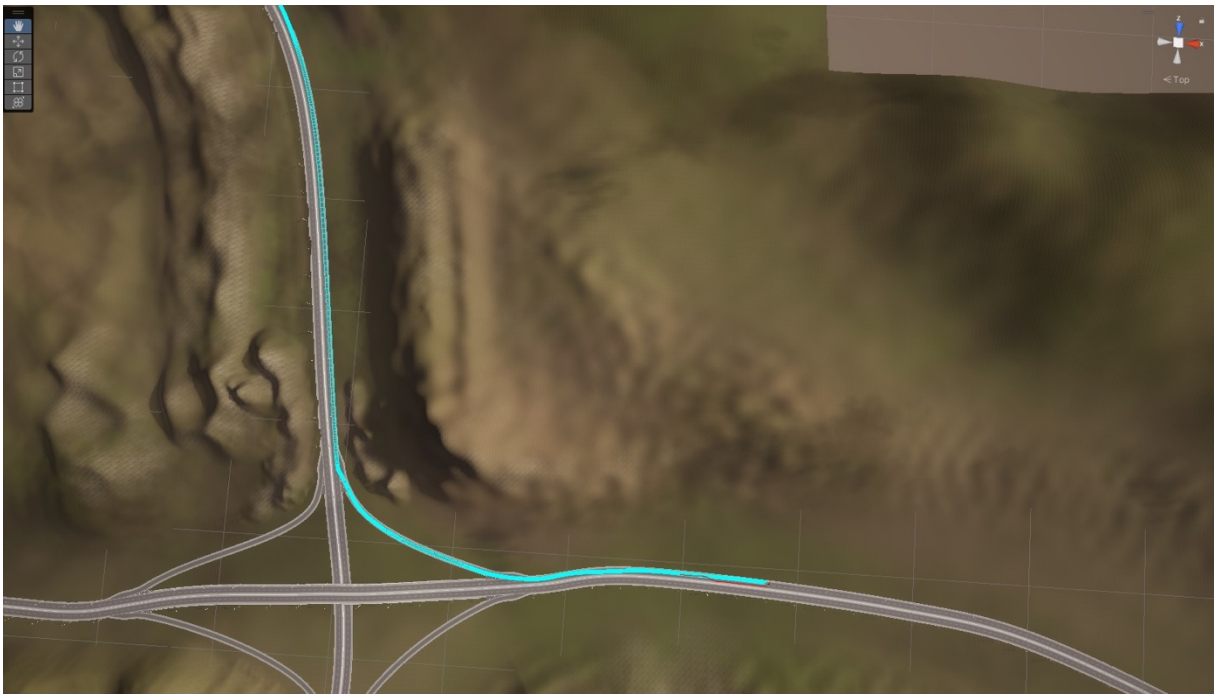


Figure 8. *Autonomous driving path map in Unity.*



Moreover, after the end of internship, I had to spend additional time on the development of the simulation. The development process ended on 6th of April 2023. From the end of internship to 6th of April I had to do the following things for the development:

- After the autonomous car module was available on Unity, I realized that the car did not follow the paths exactly. Thus, I adjust the paths and created new paths. More scenarios were created.
- Creating even points throughout the autonomous driving paths in different scenarios to distinguish between UI and general events.
- Using Unity's node editor to distinguish different events (taking an exit, changing lane, or stopping) on different event points. Added experiment requirements (length of handover durations, steering of the wheel, pressing the pedal) into the nodes and UI events in every event per scenario.

Appendix C – Measurement Items

Table 8. *Measurement items.*

Factor	Description of measurement items	Origin
Cybersickness	CyberSickness in Virtual Reality Questionnaire (CSQ-VR)	Kourtesis et al., 2023
Sleepiness	Karolinska Sleepiness Scale (KSS)	Åkerstedt & Gillberg (1990)
Trust	I trust the automation in this situation.	Adapted from Lu et al. (2022)
Mental Workload	Rating Scale Mental Effort (RSME)	Adapted from Zijlstra (1993)
Trait Anxiety	I tire quickly.	Spielberger et al. (1971) as cited in Lu et al. (2022)
	I worry too much over something that really doesn't matter.	
	Some unimportant thought runs through my mind and bothers me.	
	I am a steady person.	
State Anxiety	I feel calm.	Lu et al. (2022)
	I feel nervous.	
	I am tense.	

Appendix D - Verbal Information on Study

Dear participant, thank you for participating this study on assessing the importance of trust in autonomous vehicles. The whole experiment will take approximately 90 minutes. (ONLY TO SONA PARTICIPANTS) This study is worth 2 SONA credits, you will receive your SONA credit after you have completed the study.

This study involves the use of a VR headset together with a simulator in which you might be experiencing cybersickness (e.g., dizziness, feeling to vomit), please immediately report to us if you experience any discomfort before, during and after the experiment.

You will have the right to withdraw this experiment at any moment without any reason, your data will also be removed. If you wish to have a copy of the informed consent, please inform us. If you have any questions up to this point, please let us know.

In this study, you will be experiencing a Level 3 autonomous vehicle, with by simple definition the vehicle will mainly be controlled by the automation system and you as a driver are expected to take over when needed. So, in this study, you will be asked to respond to the task displayed on the screen on your right-hand side. You can respond by turning the steering wheel in the direction of left or right. You can step in the middle paddle to stop. The entire experiment contains 7 scenarios. After each scenario you will be filling in a questionnaire. Do you have any questions for now?

Before we began, we would like to address a few things:

Please kindly put your phone on silence mode and place it away from your pocket, so it will not hinder you during the experiment.

You can adjust the sitting position that best suits you by pulling the bar underneath the chair.

Please relax and sit back during the experiment.

Please kindly place your dominant feet on the middle paddle to stop the car.

Please make sure you are in a sitting position + VR position where you can clearly and fully see the steering wheel and the monitor at your right-hand side in VR environment, you can also request us to adjust if these visualisations are unclear for you.

Please note that there will be a certain level of shakiness in the vision due to technical difficulties we encounter, please don't let in concern you.

During the experiment you will be seeing a light grey square, it was used to perform eye tracking, please don't let it concern you.

At the end of the experiment, the vehicle will continue to run, it may crash, or go off the road due to technical difficulties, please do not let it concern you as well.

The program might pause or glitch due to the unity program, this will not affect the experiment, so don't let it concern you.

Once again, if you felt uncomfortable during the experiment, please report to us immediately. We don't wish that participant to feel sick during the experiment, therefore you are free to withdraw anytime.

You don't need to steer the steering to the max or to the hard end but make sure your action was obvious and visible to the researchers, you can relax your arm and place it on your lap or other places, please do not place it on the steering wheel.

We will verbally notify you whether you fail or successfully complete the task. (ONLY TO CHATBOT GROUP) During the scenarios, after each instruction we will verbally ask you to rate several variables.

Appendix E – Instructions

INSTRUCTIONS

Society of Automative Engineers (SAE) defines Level 3 automation as conditional autonomy. The vehicle can operate independently. Steering functions, braking and acceleration are automated but the driver must be prepared to intervene. As an SAE Level 3 feature, the autonomous vehicle expects the fallback-ready user seated in the driver's seat to resume driving when requested to do so.

In this experiment, you will be asked to takeover from the automation during random points. Please do not touch the steering wheel or the pedals if you are not instructed to do so. The instructions are:

Steer the wheel to the right to move to the right lane

Steer the wheel to the right to take the right exit

Steer the wheel to the left to move to the left lane

Press on the break to stop the car

Break

(To stop the car when instructed to do so)

Steering wheel

(To exit or change lane when instructed to do so)

- If you experience nausea, dizziness, disorientation, postural instability, visually induced fatigue, and/or visually induced discomfort at any point during the experiment please inform the researchers.

Appendix F – Verbal Prompts

Prompts for the first instruction in each scenario:

Please indicate how much effort it took you to complete the task from 0 to 150, 0 is absolutely no effort, 57 is rather much effort, and 150 is extreme effort.

Now, please indicate how sleepy you are from 1 to 9, 1 is extremely alert, 5 is neither alert nor sleepy and 9 is very sleepy.

Please rate how much you agree with the following statements.

I trust the automation in this situation from 1 to 7, 1 is strongly disagree and 7 is strongly agree.

I feel calm from 1 to 7, 1 is strongly disagree and 7 is strongly agree.

I feel nervous from 1 to 7, 1 is strongly disagree and 7 is strongly agree.

I am tense from 1 to 7, 1 is strongly disagree and 7 is strongly agree.

Prompts for the second and third instructions in each scenario:

You rated your previous effort it took you to complete the task as (X) from 0 absolutely no effort to 150 extreme effort. How much would you rate it now?

You rated your sleepiness as (X) from 1 extremely alert to 9 very sleepy. How much would you rate it now?

You rated your trust in the automation as (X) from 1 lowest trust to 7 is highest trust. How much would you rate it now?

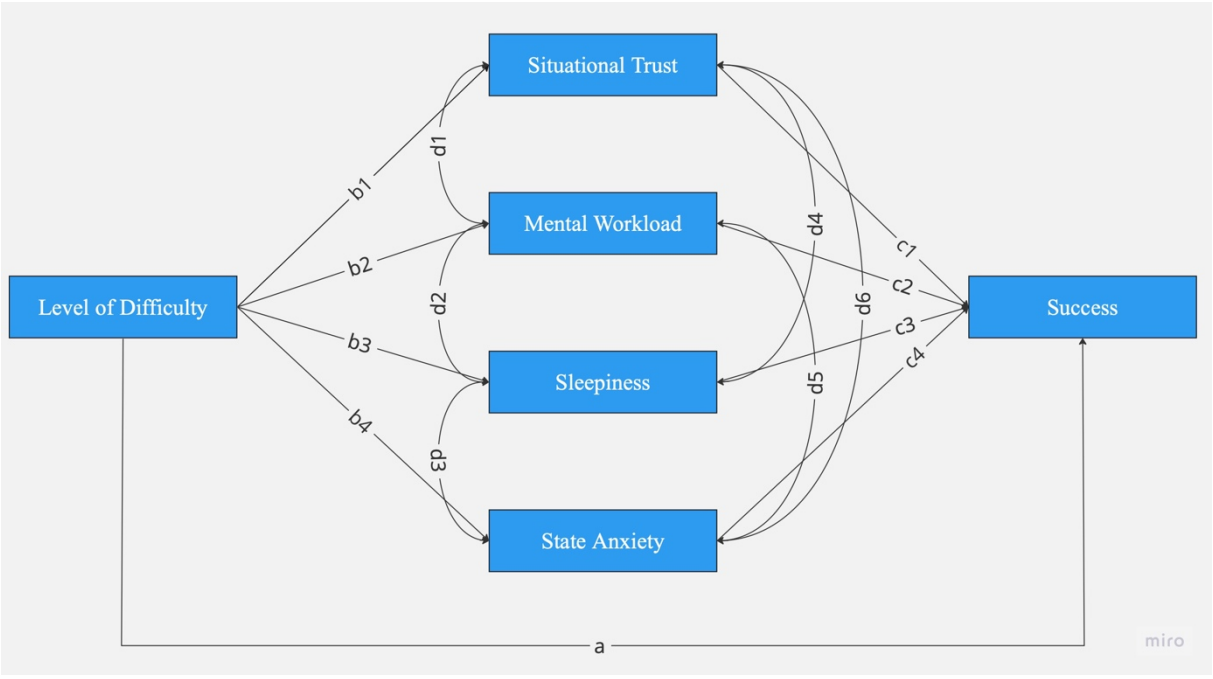
You rated your calmness as (X) from 1 lowest calmness to 7 highest calmness. How much would you rate it now?

You rated your nervousness as (X) from 1 lowest nervousness to 7 highest nervousness. How much would you rate it now?

You rated your tensity as (X) from 1 lowest tensity to 7 highest tensity. How much would you rate it now?

Appendix G – Alternative Model

Figure 9. *Graphical presentation of expected relationships between experimental variables: Within-subject independent variable (Level of difficulty), dependent variable (Success in takeover), and expected mediators (Situational trust, Workload, State anxiety, and Sleepiness). Each predicted relationship between variables (Predictions) is represented by an arrow and a letter. Arrows indicate the expected direction of effect. Letters represent the following assumptions: a. Level of difficulty influences success in takeover; b1. Level of difficulty influences situational trust; b2. Level of difficulty influences mental workload; b3. Level of difficulty influences sleepiness; b4. Level of difficulty influences state anxiety. c1. Situational trust influences success in takeover; c2 mental workload influences success in takeover; c3. sleepiness influences success in takeover; c4. state anxiety influences success in takeover; d1. situational trust has a covariation with mental workload; d2. mental workload has a covariation with sleepiness; d3. sleepiness has a covariation with state anxiety; d4. situational trust has a covariation with sleepiness; d5. mental workload has a covariation with state anxiety; d6. situational trust has a covariation with state anxiety.*



Appendix H – Comparison of Demographic Characteristics Between Groups

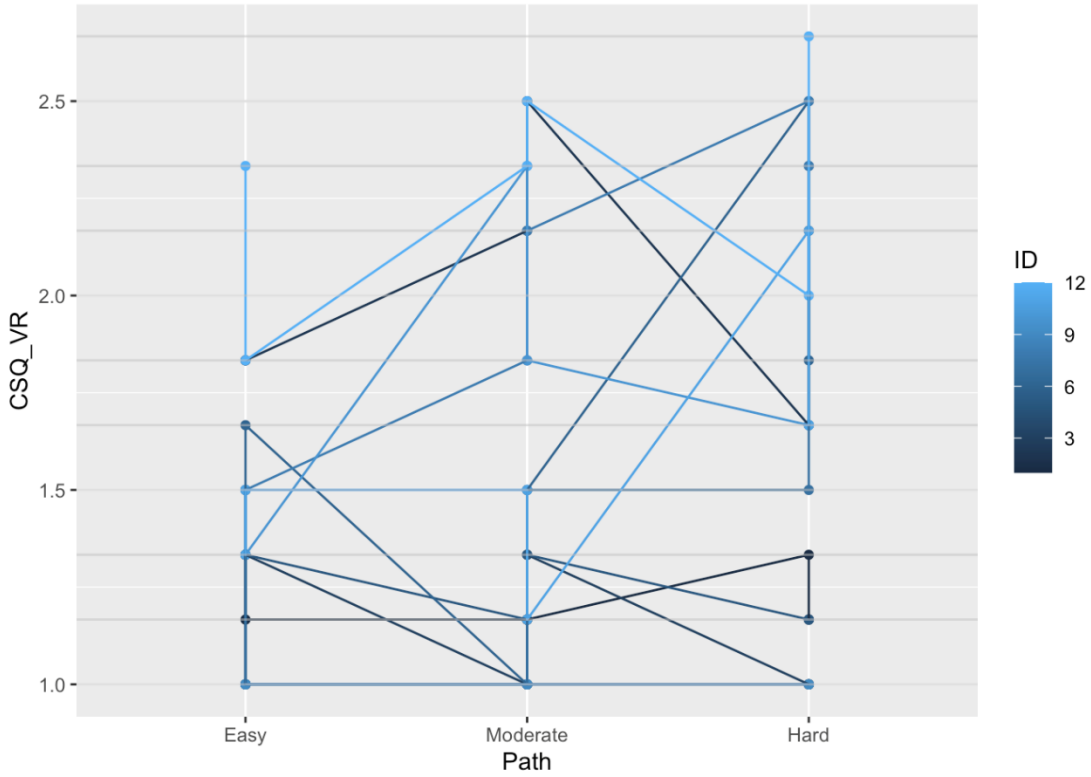
Table 9. *Comparison of demographic characteristics between form and chatbot groups (MANOVA).*

Variable	Df	Sum Sq	Mean Sq	F-value	p-value
Age	1	10.083	10.0833	2.3541	0.156
Experience	1	3	3	0.6716	0.4316
Trait anxiety	1	8.333	8.333	8.333	0.3276
Cybersickness	1	0.1875	0.1875	1.063	0.3268

Note. Df = Degrees of Freedom; Sum Sq = Sum of Squares; Mean Sq = Mean of Squares.

Appendix I – Changes in Cybersickness

Figure 10. Changes in the level of cybersickness reported by the participants. CSQ-VR represents cybersickness level, whereas "Path" represents the level of difficulty.



Appendix J – Violin Plots

Figure 11. *Changes in situational trust across different levels of difficulties grouped by Form and Chatbot Group.*

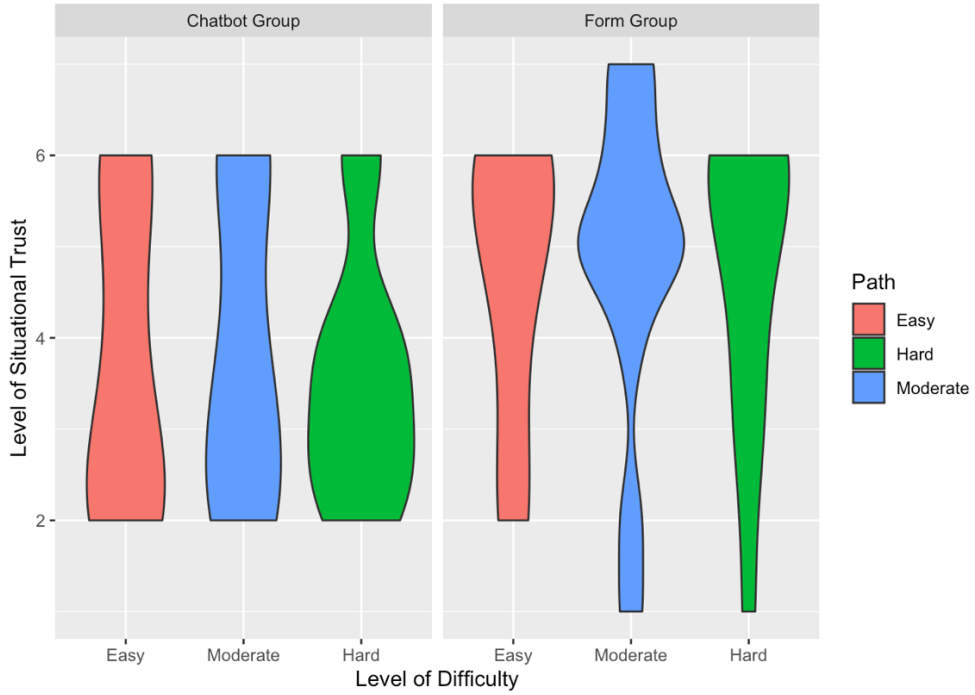


Figure 12. *Changes in sleepiness across different levels of difficulties grouped by Form and Chatbot group.*

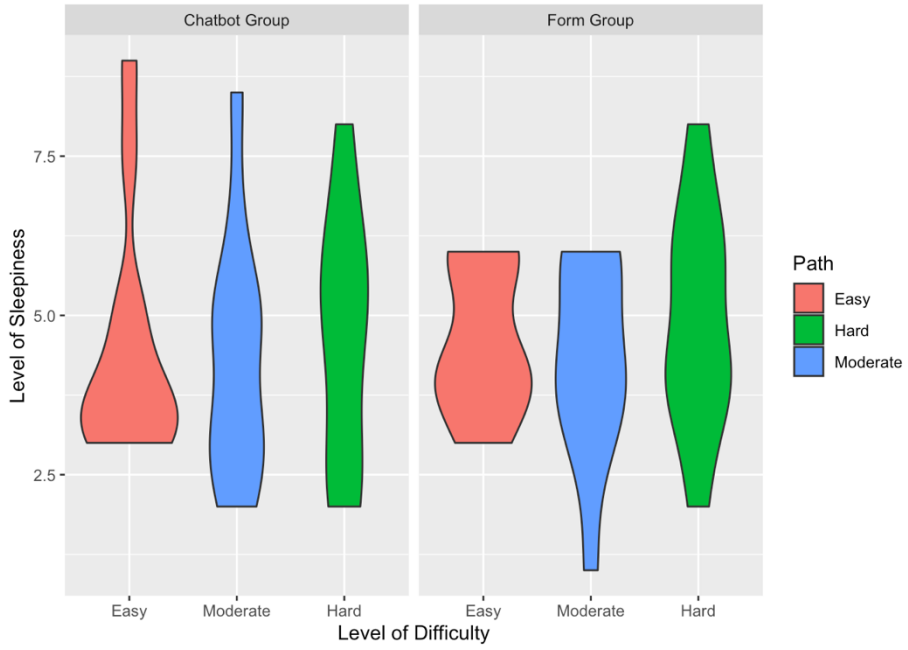
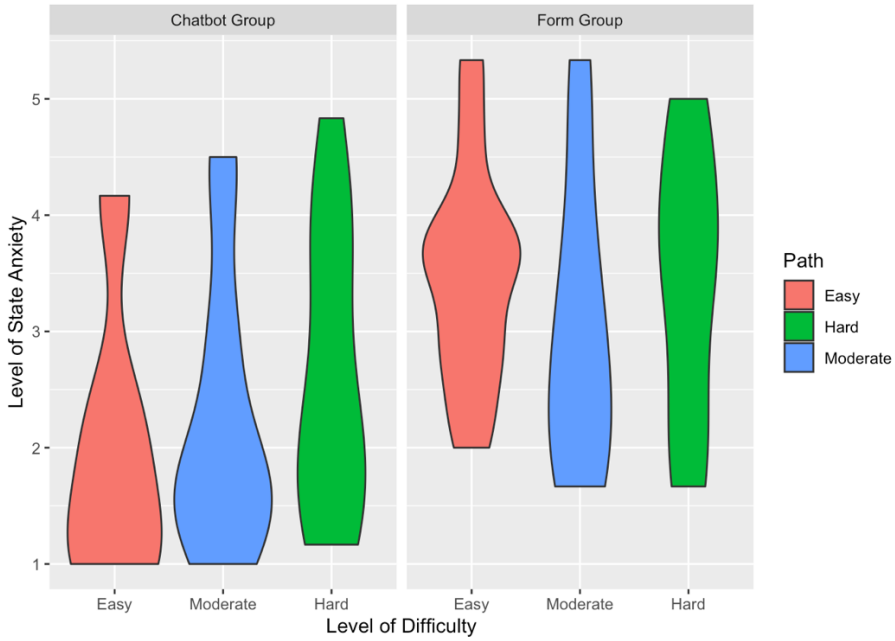


Figure 13. *Changes in state anxiety across different levels of difficulties grouped by Form and Chatbot Group.*



Appendix K – Analysis of Concurrent Effects

Table 10. *Combined effects of variables on success in takeover. The arrows indicate the direction of effect between variables. The plus signs indicate combined effects.*

Combined Effects	Coefficient Value	Standard Deviation	95% Lower Bound	95% Upper Bound
Success in takeover ← Level of Difficulty + Situational Trust	-0.073	0.052	-0.174	0.028
Success in takeover ← Mental Workload (Mediated by level of difficultly and situational trust)	-0.298	0.064	-0.424	-0.172
Success in takeover ← Sleepiness (Mediated by level of difficultly and situational trust)	-0.077	0.060	-0.195	0.041
Success in takeover ← State Anxiety (Mediated by level of difficultly and situational trust)	-0.011	0.086	-0.179	0.157
Success in takeover ← Mental Workload + Sleepiness + State Anxiety (Mediated by level of difficultly and situational trust)	-0.386	0.114	-0.610	-0.162
Success in takeover ← Level of Difficulty + Situational Trust + Mental Workload + Sleepiness + State Anxiety	-0.459	0.128	-0.709	-0.208

Appendix L – R Code

The R code can be accessed via the link:

https://drive.google.com/file/d/1NMl4bluFHrtYh0uoSno8rTWFou99rwFC/view?usp=sharing