Predicting Driver Takeover Performance and Designing Alert Systems in Conditionally Automated Driving

by

Na Du

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in The University of Michigan
2021

Doctoral Committee:

Assistant Professor X. Jessie Yang, Chair
Professor Yili Liu
Associate Professor Lionel Robert
Professor Nadine Sarter

Na Du

nadu@umich.edu

ORCID iD: 0000-0002-4383-2451

For all the people

ACKNOWLEDGEMENTS

There are many people for whom without their support, this dissertation would not be possible.

First and foremost, I would first like to thank my advisor, Dr. X. Jessie Yang, for her guidance, support, and encouragement throughout this dissertation. Thank you for leading me in this exploration journey, for helping me grow into an independent researcher, which has inspired me to be a good mentor and teacher to my future students. I am incredibly grateful for all the trust and patience you have in me.

I would also like to thank my committee members, Dr. Nadine Sarter, Dr. Yili Liu, and Dr. Lionel Robert, who provided their invaluable time and wisdom at various parts of this dissertation process. I truly appreciate the valuable feedback and support you offered, even during the pandemic. I have always been inspired by your thirst for knowledge and kindness to students, which further encouraged me to pursue an academic career.

I am grateful to work with extraordinary faculty collaborators - Dr. Dawn Tilbury, Dr. Lionel Robert, Dr. Feng Zhou, Dr. Anuj Pradhan, and Dr. Lisa Molnar. Your insights and knowledge into the subject matter steered me through the research. Without your help and wise guidance, my doctoral study would not have been the same. I am indebted to your dedication to my growth as an independent researcher. I also enjoyed working with fantastic student collaborators - Ruikun Luo, Jacob Haspiel, Jill Meyerson, Qiaoning Zhang, Kevin Yiwei Huang, Peifu Mo, Jackie Ayoub, Jinyong Kim, Shreyas Bhat, Julia Schachinger, Jose Cordova, and Yining Cao.

This dissertation would not have been possible without the contributions from the University of Michigan Transportation Institute and Industrial and Operations Engineering faculty and staff. Many thanks to Brian, Marina, Chris, Matt, Tina, Tracey, Teresa, Cathy, Rod, Sherri, Tom, Akosua, Wanda, Valerie, Liz, Mint, Sheryl, Cornelius, Edwin, James, John, Paul, and Mark. I am appreciative of your unlimited support and help that made everything on this journey turn out fine. Thanks are also to all the participants who volunteered in my experiments.

This work was financially supported by the University of Michigan Mcity, Industrial Operations Engineering Department, and the Rackham Graduate School.

I was also fortunate to work with industrial partners through the sponsored research projects and internship during my doctoral study. I would like to thank Xingwei Wu, Teruhisa Misu, Kumar Akash, and Sujitha Martin from Honda Research Institute, as well as Elizabeth Pulver from the State Farm Insurance Company. The support and guidance from you helped me to study the research questions from an industrial perspective.

I would like to thank my labmates in the ICRL lab – Ruikun, Yaohui, Taa and in the MAVRIC Lab - Qiaoning, Suresh, Hebert, Arsha, Conner, Huajing. Thank you, friends and cohorts in the program - Ke, Albert, Karanvir, Kevin, Yuzhi, Yidu, Kam, Yifan, Patrik, Justin, Sol, Yanxun, and Yadrianna. You brought wonderful memories and made my doctoral program fun and unforgettable experience. I would also like to thank my friends I met in Ann Arbor. Special thanks go to Tingting, Liqi, and Shuanghong for being my roommates and having fun with me during this journey.

Finally, I would like to take the time to thank my family for their unwavering and unconditional support, encouragement, and love.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

ABSTRACT

With the Society of Automotive Engineers Level 3 automation, drivers are no longer required to actively monitor driving environments, and can potentially engage in non-driving related tasks. Nevertheless, when the automation reaches its operational limits, drivers will have to take over control of vehicles at a moment's notice. Drivers have difficulty with takeover transitions, as they become increasingly decoupled from the operational level of driving. In response to the takeover difficulty, existing literature has investigated various factors affecting takeover performance. However, not all the factors were studied comprehensively, and the results of some factors were mixed. Meanwhile, there is a lack of research on the development of computational models that predict drivers' takeover performance using their physiological and driving environment data. Furthermore, current research on the design of in-vehicle alert systems suffers from methodological shortcomings and presents identical takeover warnings regardless of event criticality.

To address these shortcomings, the goals of this dissertation were to (1) examine the effects of drivers' cognitive load, emotions, traffic density, and takeover request lead time on their driving behavioral (takeover timeliness and quality) and psychophysiological responses (eye movements, galvanic skin responses, and heart rate activities) to takeover requests; (2) develop computational models to predict drivers' takeover performance using their physiological and driving environment data via machine learning algorithms; and (3) design in-vehicle alert systems with different

display modalities and information types and evaluate the displays in different event criticality conditions via human-subject experiments.

The results of three human-subject experiments showed that positive emotional valence led to smoother takeover behaviors. Only when drivers had low cognitive load, they had shorter takeover reaction time in high oncoming traffic conditions. High oncoming traffic led to higher collision risk. High speed led to higher collision risk and harsher takeover behaviors in lane changing scenarios, but engendered longer takeover reaction time and smoother takeover behaviors in lane keeping scenarios. Meanwhile, we developed a random forest model to predict drivers' takeover performance with an accuracy of 84.3% and an F1-score of 64.0%. Our model had finer granularity than and outperformed other machine learning models used in prior studies. The findings of alert system design studies showed that drivers had more anxiety with the *why* only information compared to the *why + what will* information when information was presented in the speech modality. They felt more prepared to take over control of the vehicle and had more preference for the combination of augmented reality and speech conditions than others when drivers were in high event criticality situations.

This dissertation can add to the knowledge base about takeover response investigation, takeover performance prediction, and in-vehicle alert system design. The results will enhance the understanding of how drivers' emotions, cognitive load, traffic density, and scenario type influence their takeover responses. The computational models for takeover performance prediction are underlying algorithms of in-vehicle monitoring systems in real-world applications. The findings will provide design recommendations to automated vehicle manufacturers on in-vehicle alert systems. This will ultimately enhance the interaction between drivers and automated vehicles and improve driving safety in intelligent transportation systems.

CHAPTER 1

Introduction

1.1 SAE Level 3 automated vehicles

Automated vehicles (AVs) have the potential to provide our society with more fuel-efficient driving, reduce driving-related injuries and deaths, and reshape transportation and logistics (*Robert*, 2019; *Du et al.*, 2019b). AVs are becoming a major focus for automotive manufacturers, such as General Motors, Tesla, and Honda. The Society of Automotive Engineers (SAE) defines driving automation from Level 0 to Level 5 (*Society of Automotive Engineers*, 2018). While people are still speculating if and when Level 5 full automation will be ready (*Sparrow and Howard*, 2017), Level 2 partial automation is already on the road (*Bishop*, 2019). Honda Sensing Elite, which is the first commercially available Level 3 system, was released in its domestic-market Legend sedan in March 2021 (*Gilboy*, 2021).

As shown in Table 1.1, in partially automated driving (Level 2), drivers complete the object and event detection and response task. They are asked to intervene with the automation when necessary. In conditionally automated driving (Level 3), object and event detection and response task is handled by the AV, and drivers resume responsibility when receiving a takeover request (TOR). In automated driving mode, drivers are allowed to potentially engage in other tasks, and therefore become increasingly out of the control loop. This transition of control represents the transfer

Table 1.1: Society of Automotive Engineers Levels of Automation (Adapted from SAE (2018))

Level	Description
0	No automation involved. The human driver controls all elements of the driving task and monitors the driving environment.
1	Automation controls either the steering or acceleration/braking of the vehicle, while the human controls all other elements of the driving task and monitors the driving environment.
2	Automation controls both the steering and acceleration/braking of the vehicle, while the human monitors the driving task and serves as an immediate fallback for the automation with little notice.
3	Automation controls both the steering and acceleration/braking of the vehicle and monitors the driving task, while the driver serves as a fallback for the automation. Transitions of control are guided by takeover requests, except during automation failures.
4	Automation executes all control and monitoring aspects within a specified operational design domain (ODD), and does not require the driver to serve as a fallback for the automation. Human drivers (if any) may assume control after exiting the ODD, but the system does not rely on drivers to do so.
5	Automation controls all aspects of the driving task under all roadway and environmental conditions. Input is never expected from a human driver.

of the longitudinal and lateral control responsibilities from the automated vehicle to the human driver, and usually involves the driver terminating NDRTs, moving their eyes/hands/feet back to the road/steering wheel/pedals, and resuming control of the vehicle. Research indicates that drivers have difficulty in takeover transitions as they become increasingly decoupled from the operational level of driving (*Gold et al.*, 2016; *Eriksson and Stanton*, 2017; *Petersen et al.*, 2019).

1.2 Factors influencing takeover responses

In response to the takeover difficulty, research has been conducted to investigate factors affecting takeover performance, including drivers' states (e.g., workload, emotions, characteristics), driving environments (e.g., traffic situations, weather conditions, takeover scenarios), and vehicle capability (e.g., TOR lead time, automation level) (*Eriksson and Stanton*, 2017; *Gold et al.*, 2016; *Wan and Wu*, 2018; *Körber* *et al.*, 2016; *Helldin et al.*, 2013; *Gold et al.*, 2018a; *Radlmayr et al.*, 2014; *Eriksson et al.*, 2018).

1.2.1 Drivers' states

Drivers' states can be influenced by their non-driving-related tasks (NDRTs) in automated driving or depend on their characteristics. As reported in the literature, a wide range of NDRTs have been utilized in experimental studies, including both naturalistic tasks (e.g., text messaging) and artificial tasks (e.g., N-back memory task) (*Baek et al.*, 2019; *Wandtner et al.*, 2018; *Wan and Wu*, 2018; *Radlmayr et al.*, 2014). Drivers' characteristics include their age, AV training, and experience, among others. Literature has shown that drivers' workload and emotions induced by NDRTs, as well as their characteristics, influenced their takeover responses.

1.2.1.1 Workload

Prior research compared drivers' takeover performance when performing versus not performing an NDRT, and revealed that NDRTs decreased takeover quality, resulting in more crashes in high-traffic situations (*Radlmayr et al.*, 2014), shorter minimum TTC (*Gold et al.*, 2016; *Körber et al.*, 2016), larger lateral acceleration (*Louw et al.*, 2015), and larger standard deviation of lane position (*Zeeb et al.*, 2016). In addition, several studies examined the effects of performing different types of NDRTs. For example, *Wandtner et al.* (2018) compared NDRTs in visual (i.e., Surrogate Reference Task) and auditory modalities (i.e., N-back task). They found that operation of the visual task with handheld devices degraded takeover performance and led to a higher collision rate, while the auditory task led to comparable performance to a baseline without any task. Moreover, *Zeeb et al.* (2016) and *Wan and Wu* (2018) examined drivers' takeover performance while they were typing, reading, watching a video clip, playing a game, or taking a nap. Results showed that different NDRTs led to few differences in takeover reaction time, but that watching a video and taking a nap resulted in a worse takeover quality, as they occupied more sensory modalities or induced a very low arousal level.

The above-mentioned studies shed some light on the influence of performing NDRTs on takeover performance. However, most of the studies did not directly manipulate the workload of NDRTs. One exception is the study of *Zeeb et al.* (2017), wherein the drivers' cognitive task load was manipulated via the difficulty of NDRTs (i.e., reading vs. proofreading a text) and manual task load via the tablet location (i.e., handheld vs. mounted). Results revealed that high manual task load increased reaction times and deteriorated takeover quality, while the effect of cognitive task load on takeover ability was dependent on the type of driver intervention. High cognitive load lengthened the reaction times and deteriorated takeover quality in steering maneuvers but not braking maneuvers. In contrast, *Bueno et al.* (2016) manipulated workload by asking drivers to identify and type out the association between three pictures, and found a non-significant effect of workload on takeover time and quality. Further research is needed to elucidate these mixed results and to examine the impact of cognitive load.

1.2.1.2 Emotions

According to *Russell* (1980), emotion has at least two dimensions. The first dimension is valence, or how negative or positive a stimulus is. For example, watching a baby smiling is more positive than seeing a patient dying. The second dimension is arousal, or how sleepy-inducing or exciting a stimulus is. For example, listening to rock bands is associated with higher arousal than listening to meditation music. The two dimensions can be mapped in a two-dimensional space, and the combinations of different values of valence and arousal are associated with different discrete emotions (See Figure 1.1). For example, the upper left corner of the two-dimensional space

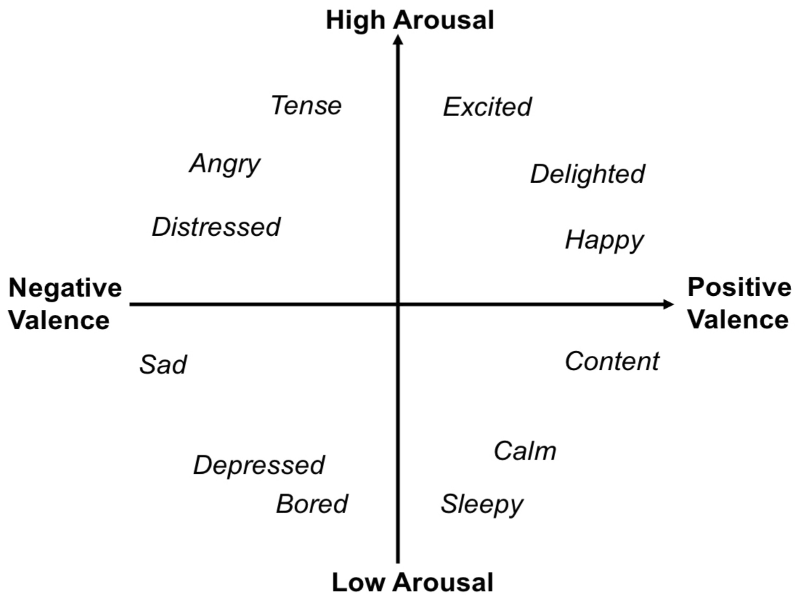


Figure 1.1: The circumplex model of emotion (*Russell*, 1980)

represents emotions of negative valence and high arousal, such as anger; the lower right corner of the space represents emotions of positive valence and low arousal, such as calmness. While it may be more straightforward for individuals to report discrete emotions they are experiencing, the dimensional view of emotion provides a more fundamental explanation of the relationships between emotions and behaviors (*Barrett*, 1998). The dimensional view is also supported by studies in neuroscience. With evidence of neural activities in the brain, valence and arousal appear to influence cognitive processes and human behaviors via distinct mechanisms (*Dolcos et al.*, 2004; *Kensinger*, 2004).

Manual driving is a complex task involving attention, information processing, and action-based judgment. Drivers can become emotional on the road when they interact with external environments and other road users, which may lead to enormous consequences (*Jeon*, 2017). Some studies placed emphasis on the effects of specific emotions on manual driving, in particular, the effects of anger. Anger, as one of the most commonly experienced emotions during driving, has received a substantial amount of research attention. A recent survey study by the AAA Foundation for

Traffic Safety found that nearly 80 percent of drivers had expressed anger, aggression, or road rage, which are significant contributors to fatal crashes, at least once in the previous year (*AAA Foundation for Traffic Safety*, 2016). An analysis of naturalistic driving data showed that drivers in elevated emotional states, including anger, sadness, crying, and/or emotional agitation, have an increased risk of a crash by 9.8 times (*Dingus et al.*, 2016). Moreover, experimental studies indicate that anger leads to risky and aggressive behaviors, such as speeding and traffic rule violation (*Abdu et al.*, 2012; *Jeon et al.*, 2014; *Underwood et al.*, 1999; *Deffenbacher et al.*, 2003; *Hu et al.*, 2018). For example, *Abdu et al.* (2012) conducted a driving simulator study with 15 licensed drivers, and found that angry drivers crossed more yellow traffic lights and tended to drive faster. Similarly, *Jeon et al.* (2014) found that anger led to a significantly lower perceived safety and degraded driving performance (i.e., larger deviations from the center line and more violations of traffic rules).

Moreover, researchers went beyond specific emotions and systematically explored the effects of positive/negative valence and high/low arousal on manual driving performance (*Hancock et al.*, 2012; *Trick et al.*, 2012; *Chan and Singhal*, 2013; *Ünal et al.*, 2013; *Pêcher et al.*, 2009). *Chan and Singhal* (2013) investigated the effects of emotional valence on driving. In their study, participants were responsible for longitudinal and lateral control of the vehicle. At the same time, they were asked to view words of positive, negative, and neutral valence on roadside billboards, and later to recall as many words as possible. Results revealed that drivers recalled more negative words than positive words, suggesting that negative stimuli distracted drivers' attention more severely. In another study, participants drove and viewed emotional images concurrently. Viewing positive images led to better lateral control but also slower speeds when compared to negative images (*Hancock et al.*, 2012). The positive association between better vehicle control and positive valence was also reported in the studies by *Trick et al.* (2012) and *Groeger* (2013). Interestingly, using another emotion induction method, *Pêcher et al.* (2009) asked drivers to listen to music, and found that happy music (positive valence) resulted in an unexpected large decrement of speed and a deteriorated lateral control in comparison with sad music (negative valence). The reason for the inconsistent findings could be due to the differences in participants' emotion induction methods (*Steinhauser et al.*, 2018) and participants' base emotions and personal experience.

In addition, *Trick et al.* (2012) manipulated both emotional valence and arousal in an experiment where participants were exposed to a variety of images that were either positive or negative in valence and either high or low in arousal. After viewing the images, participants needed to brake in reaction to the sudden deceleration of a lead vehicle. Results showed that higher arousal led to faster hazard response if the hazard was presented shortly after viewing an image. Similarly, *Navarro et al.* (2018) and *Ünal et al.* (2013) conducted experiments to manipulate drivers' emotional arousal using musical tempo. Results showed that in a car-following task, an arousing musical background improved drivers' responsiveness to the speed changes of the followed vehicle compared to relaxing music.

Despite the large amount of research on manual driving (*Abdu et al.*, 2012; *Jeon*, 2017; *Chan and Singhal*, 2013; *Pêcher et al.*, 2009; *Trick et al.*, 2012), few studies have examined how emotions influence driving performance in conditionally automated driving. In addition, those few studies have primarily been focused on algorithm development to automatically detect drivers' emotional states by analyzing their physiological data. For example, *Izquierdo-Reyes et al.* (2018) developed an algorithm using drivers' faces and electroencephalogram (EEG) data as features for classifier training. The results showed that a k-Nearest Neighbors algorithm was able to recognize nine different emotions (neutral, anger, disgust, fear, joy, sadness, surprise, amusement, and anxiety) with an accuracy of approximately 97%. Thus, it is necessary to study how emotions influence drivers' takeover responses in conditionally automated driving.

1.2.1.3 Drivers' characteristics

Drivers' individual characteristics play an important role in determining takeover performance. Previous studies have focused on factors such as takeover training, cultural background, and age (*Hergeth et al.*, 2017; *Clark and Feng*, 2017; *Gold et al.*, 2018a). For example, *Clark and Feng* (2017) studied younger (ages 18–35) and older drivers (ages 62–81) in the takeover of vehicle control, and found that older drivers deviated less from the road centerline and drove at a lower speed, but applied more pressure on the brake pedal compared with young drivers. *Li et al.* (2018) found that older adults had larger takeover time compared to younger adults. With regard to AV experience, *Zeeb et al.* (2015) found that repetitive exposure to takeover transitions reduced drivers' hands-on-wheel and eyes-on-road reaction time.

1.2.2 Driving environments

Researchers have investigated the impacts of driving environments, including traffic density (*Gold et al.*, 2016; *Körber et al.*, 2016), takeover scenarios (*Eriksson et al.*, 2018; *Naujoks et al.*, 2017), and weather (*Li et al.*, 2018; *Louw et al.*, 2015), on drivers' takeover performance.

1.2.2.1 Traffic density

Traffic density indicates the average number of vehicles (in the same or opposite direction as the ego vehicle) within a certain distance (e.g., per kilometer). Research showed that heavy traffic density in the same direction of the AV led to longer takeover time (*Gold et al.*, 2016; *Körber et al.*, 2016), more braking rather than steering (*Eriksson and Stanton*, 2017), lower minimum time to collision (*Gold et al.*, 2016; *Körber et al.*, 2016), more collisions (*Gold et al.*, 2016; *Körber et al.*, 2016), and larger maximum accelerations (*Gold et al.*, 2016; *Li et al.*, 2018; *Körber et al.*, 2016). In these studies, high traffic density was coupled with fewer escape paths. With a high traffic density, there were more vehicles in the same or neighboring lanes of the automated ego vehicle, restricting the drivers' action opportunities. There is a need to decouple traffic density and the availability of escape paths.

1.2.2.2 Takeover scenarios

A variety of takeover scenarios have been designed in previous studies in order to examine their effects on takeover performance. For example, *Naujoks et al.* (2014) examined the effects of TOR modalities (i.e., visual vs. visual plus auditory) under three takeover scenarios: missing lane markings, temporary lines because of a work zone, and high road curvature. The results of their study showed that drivers' lateral control was better with visual-auditory TORs, and such advantages were especially pronounced in the high road curvature scenario. Later, *Naujoks et al.* (2017) used the same three takeover scenarios and manipulated automation level (hands-on vs. hands-off vs. manual) and NDRT (with vs. without NDRT) in another experiment. They found that only in the temporary lines because of a work zone condition, engaging NDRT increased the self-reported situation criticality. Only in the high curvature scenario, high automation level increased variability of the lateral vehicle position and the self-reported situation criticality.

More recently, *Dogan et al.* (2019) investigated the effects of NDRTs on takeover performance in two takeover scenarios (i.e., missing lane markings and obstacle ahead). Results showed that regardless of NDRT type, drivers had shorter TOR reaction time and lower mental workload in the obstacle avoidance scenario than the missing lane scenario. Similarly, *Wu et al.* (2019) analyzed drivers' takeover performance under different scenarios and NDRTs. Scenarios included obstacle on the left, obstacle ahead, and obstacle ahead with lead vehicle. NDRTs included a 1-back memory task and a letter-game task. They found that drivers had the shortest steering reaction time in the obstacle ahead scenario, and longer minimum TTC in the obstacle ahead scenario than the obstacle ahead with lead vehicle scenario. Yet, drivers' workload, braking reaction time, and maximum deceleration were similar among all scenarios.

One possible reason for the mixed results across different scenarios is that these scenarios were not categorized based on their underlying dimensions. To our knowledge, only two studies mentioned below tried to extract underlying influential dimensions of takeover scenarios and study their effects systematically.

Eriksson et al. (2018) is one exception that attempted to categorize scenarios into lane changing and braking scenarios, corresponding to different augmented visual interfaces to support drivers in making a correct lane changing or braking reaction in takeover transitions. Although the study did not compare the two scenario types directly, it did emphasize the fundamental differences between the two types and used different augmented feedback to provide recommendations. *Zeeb et al.* (2017) designed two scenarios based on the types of takeover reactions in a simulated driving study. In the longitudinal scenario, drivers were required to intervene with a cutting-in and hard-braking vehicle. In the lateral scenario, drivers were required to intervene with the drifting vehicle on the curve induced by a wind gust from the left side. They found that increasing manual task load had detrimental effects on reaction time and takeover quality in both scenarios, but the effects were more pronounced in the lateral scenario. High cognitive load decreased reaction time and takeover quality in the lateral maneuver, but not in the longitudinal maneuver. Notably, researchers excluded ten drivers who reacted with a lane change in the longitudinal scenario from analysis because their behavior was not comparable to the other drivers' reactions. However, it is common for drivers to brake and change lanes simultaneously in critical situations. This suggests that another criterion is needed to categorize takeover scenarios.

1.2.2.3 Weather

With regard to weather, *Louw et al.* (2017) found that less available visual information (i.e., fog) was linked to shorter minimum distance headway and minimum TTC. Similarly, *Li et al.* (2018) showed that drivers in adverse weather conditions (i.e., snow, rain, and fog) had longer TOR reaction time, shorter minimum TTC, larger resulting acceleration, and steering wheel angles.

1.2.3 Vehicle capability

Vehicle capability indicates its configuration, such as automation level, TOR lead time, and human-machine interfaces. Literature has shown that vehicle automation level and TOR lead time influence drivers' takeover responses (*Mok et al.*, 2015; *van den Beukel and van der Voort*, 2013; *Wan and Wu*, 2018; *Walch et al.*, 2015; *Gold et al.*, 2013; *Madigan et al.*, 2018). The literature on the effects of human-machine interfaces is reviewed in Section 1.5.

1.2.3.1 Automation level

As shown in Table 1.1, AVs' automation levels range from Level 0 to Level 5. Compared to continuous manual driving, driving emergencies that requested takeovers from automated driving led to longer reaction time and worse driving quality (*Gold et al.*, 2013; *Radlmayr et al.*, 2014). Furthermore, some researchers compared drivers' takeover responses at different automation levels of AVs (*Madigan et al.*, 2018; *Shen and Neyens*, 2014). For example, *Madigan et al.* (2018) compared takeover responses in Level 2 and Level 3 AVs, and found that drivers had later overtaking manoeuvres in Level 2 than Level 3 AVs. This dissertation only focuses on Level 3 AVs.

1.2.3.2 TOR lead time

One of the most important factors that influences takeover performance is the TOR lead time. Lead time refers to the time to collision at the time of the TOR (*McDonald et al.*, 2019b). According to previous meta-analyses (*Gold et al.*, 2018b), an increase of 1 s in TOR lead time generally led to a 0.2 to 0.3 s increase in takeover time. Yet, such effects may disappear when the lead time is short. It may be because the short lead time triggers reflexive and quick responses. Research demonstrated that shorter TOR lead time degraded takeover quality in the form of shorter minimum TTC, higher crash rates, greater maximum accelerations, and greater standard deviation of steering wheel angle (*Mok et al.*, 2015; *van den Beukel and van der Voort*, 2013; *Wan and Wu*, 2018; *Walch et al.*, 2015). It would be interesting to investigate whether there are any interaction effects between TOR lead time and other factors.

1.3 Drivers' responses in driving

1.3.1 Driving behavioral responses

To quantify how different factors influence takeover transitions, existing studies have mainly focused on driving behaviors after TORs. In takeover performance measurements, driving behaviors are categorized into two aspects, namely, takeover timeliness and takeover quality. Takeover timeliness means how quickly drivers respond to TORs, and is measured as the time between the TOR and the first indicator of a takeover maneuver. Takeover quality consists of a wide range of metrics, including speed, acceleration and jerk statistics, time/distance to collision statistics, steering angle and pedal statistics, lane deviation statistics, and crash rate. For example, *Gold et al.* (2016) measured drivers' minimum TTC and crash numbers, and illustrated that heavy traffic density led to worse takeover quality demonstrated by shorter minimum TTC and more crashes. More recently, *Du et al.* (2020d) used smaller maximum resulting acceleration and maximum resulting jerk as indicators of good takeover quality to show the advantages of positive emotional valence for takeovers during automated driving.

While these driving metrics quantify drivers' vehicle control after TORs and provide insight into the prominent effects of factors on takeover performance, they have the following limitations. First, driving metrics capture drivers' behaviors at a specific moment (e.g., minimum TTC) or overall level (e.g., standard deviation of lane positions), but cannot account for the entire takeover process in a consecutive time-series way. Second, although drivers sometimes do not show observable varieties at the performance level, their cognitive and emotional states might be significantly influenced, and should be used to measure their overall takeover experience. Self-reported subjective measures can also assess drivers' internal states. Yet, self-reporting internal states significantly interferes with the real-time task at hand, and could be difficult for drivers during the takeover transitions (*Schmidt et al.*, 2009). Therefore, it is necessary to collect drivers' psychophysiological signals to examine their workload, emotions, attention, and situation awareness, promptly and continuously.

1.3.2 Psychophysiological responses

With the development of low-cost and non-invasive wearable sensors, it is achievable to collect drivers' psychophysiological signals to reflect their cognitive and emotional states as affected by NDRTs, vehicle configurations, and driving environments. Commonly used measurements in vehicle-related research include eye movements, heart rate (HR) activities, galvanic skin responses (GSRs), facial expressions, and brain signals.

1.3.2.1 Gaze behaviors

Gaze behaviors, such as gaze dispersion and blink number, have been widely used in driving studies to reflect drivers' cognitive load, attention, and situational awareness (*Wang et al.*, 2014; *Luo et al.*, 2019; *Young et al.*, 2013; *Lemercier et al.*, 2014). Researchers have shown that increases in drivers' cognitive load induced by NDRTs and environments are linked to increases in pupil diameter and decreases in horizontal gaze dispersion and blink number (*Wang et al.*, 2014; *Gold et al.*, 2016; *Merat et al.*, 2012; *Luo et al.*, 2019). For example, *Merat et al.* (2012) compared drivers' states when they were in different scenarios (with vs. without critical incident), NDRTs (with vs. without Twenty Questions Task), and drive (manual vs. automated). They found that blink frequency was generally suppressed during high workload conditions, where drivers experienced critical incidents and Twenty Questions Task. *Gold et al.* (2016) found that horizontal gaze dispersion was the most sensitive measure of drivers' cognitive demand in NDRTs during conditionally automated driving. From the attention perspective, *Louw et al.* (2015) investigated driver attention in automated driving and measured drivers' gaze dispersion with four manipulations: (1) no manipulation, (2) light fog, (3) heavy fog, and (4) heavy fog with a visual NDRT. They found that drivers had wider gaze dispersion when the driving scene was completely in the heavy fog conditions, but became more concentrated if a visual NDRT existed. Although gaze dispersion and eyes-on-road time percentage are traditionally treated as distraction indicators in manual driving, wider gaze dispersion and larger eyes-on-road time percentage imply high situation awareness in automated driving (*Young et al.*, 2013; *Molnar*, 2017).

1.3.2.2 Heart rate activities

Heart rate and heart rate variability (HRV) have the sensitivity to assess drivers' workload and detect workload changes before the presence of observable effects in driving performance (*Mehler et al.*, 2012, 2009; *Bashiri and D Mann*, 2014; *Lohani et al.*, 2019; *Hidalgo-Muñoz et al.*, 2019). For instance, *Hidalgo-Muñoz et al.* (2019) conducted a driving simulator study with 18 subjects, and found that decreases in HRV were associated with increases in cognitive load during manual driving. More importantly, HRV reflected such variations in attention and cognitive load levels before differences in driving performance were evident. Although some researchers have argued that cardiac responses remain open for attention interpretation, it is widely established that heart rate acceleration and deceleration are associated with defense and orienting responses, respectively. Specifically, Lacey and Sokolov proposed that heart rate acceleration occurred in situations involving stimulus ignorance and environmental rejection, while heart rate deceleration indexed the intake and enhancement of environmental stimuli (*Libby Jr et al.*, 1973; *Lacey*, 1970; *Sokolov*, 1963; *Sokolov and Paramonova*, 1961). Take the driving context as an example, *Reimer et al.* (2011) found that younger drivers had heart rate acceleration in response to the phone conversation task in simulated manual driving. This pattern indicated that drivers selectively ignored or rejected disruptive input, which was the phone task in this setting. However, late middle-aged drivers did not demonstrate such a pattern, possibly due to individual differences in attentional focuses.

1.3.2.3 Galvanic skin responses

Galvanic skin responses measure skin conductance controlled by changes in the sympathetic nervous system. Raw GSR signals comprise two components, phasic activation (rapid changes to a specific stimulus) and tonic activation (slower responses at a background level of the activity) (*Boucsein*, 2012). GSRs have been found to be associated with drivers' cognitive load, stress, and emotional arousal (*Collet et al.*, 2009; *Mehler et al.*, 2012; *Wintersberger et al.*, 2018). For example, *Mehler et al.* (2012) conducted an on-road study where 108 drivers across three age groups performed an auditory working memory task with three difficulty levels during manual driving. Results showed that drivers had increased heart rate and skin conductance with a high level of cognitive demand. In the context of automated driving, *Wintersberger et al.* (2018) measured drivers' GSRs after TORs in a simulated driving study. They found that GSR phasic activation, as an indicator of drivers' arousal and stress, became higher when TORs were presented during an NDRT than between NDRTs.

1.3.2.4 Facial expressions

Facial expressions have been used to recognize drivers' and passengers' emotional states in driving (*Wintersberger et al.*, 2016; *Gao et al.*, 2014; *Izquierdo-Reyes et al.*, 2018). For example, *Wintersberger et al.* (2016) made use of passengers' facial expressions to estimate their emotional responses (in pleasure and arousal dimensions) when they were in a vehicle driven by an automated driving system, a male, or a female driver. Furthermore, *Izquierdo-Reyes et al.* (2018) developed a k-Nearest Neighbors algorithm to classify drivers' emotions (e.g., anger, sadness, joy, anxiety) in automated driving using facial expressions and reached an accuracy level of approximately 97%. Such models can potentially be used to understand drivers' emotional states, and the vehicle might respond in real time to improve drivers' user experience and reduce possible aggressive behaviors (e.g., when in anger).

1.3.2.5 Brain signals

Brain signals, such as electroencephalogram (EEG) and functional near-infrared spectroscopy (fNIRS) signals, have been used to develop computational models to detect drivers' workload and emotions in real time (*Izquierdo-Reyes et al.*, 2018; *Thirunavukkarasu et al.*, 2016; *Sibi et al.*, 2016). For example, *Thirunavukkarasu et al.* (2016) utilized drivers' EEG signals and basic vehicle information to develop an emotion detection system. The system categorized drivers' emotions into happy, sad, relaxed, and angry using a support vector machine algorithm, and helped improve driver performance and safety. Similarly, *Sibi et al.* (2016) collected drivers' fNIRS signals in a simulated partially automated vehicle and found that compared to the automated driving mode, fNIRS signals from the prefrontal cortex implied drivers' additional cognitive load in manual driving. However, the ways to collect these brain metrics tend to be intrusive compared to previously mentioned psychophysiological metrics.

1.4 Existing models for takeover performance prediction

Most of the existing literature has focused on the effects of certain variables on takeover performance, providing valuable yet largely relational insights. However, knowing the relationships between certain factors and takeover performance is not enough to accurately predict a driver's takeover performance in the real world, because many influential factors could interact with one another. Computational models capable of predicting drivers' takeover performance under various takeover conditions in real time are needed. Although a substantial amount of research has identified factors that influence drivers' takeover performance, there is a lack of research on the development of computational models for predicting drivers' takeover performance, with few exceptions (*Gold et al.*, 2018b; *Braunagel et al.*, 2017).

To predict takeover performance, *Gold et al.* (2018b) analyzed 753 takeover events using data from six driving simulator experiments, and developed regression models. Their study modeled takeover performance measures (e.g., takeover time, minimum TTC, brake application, and crash probability) as a function of the time-budget, traffic density, non-driving-related task, repetition, the current lane, and driver's age. The models were validated using 729 takeover events from five additional experiments. The validation results showed that the regression models accurately predicted takeover time, time-to-collision, and crash probability, and moderately predicted the brake application.

Braunagel et al. (2017) used machine learning algorithms to predict drivers' takeover quality (named as "takeover readiness" in the article). The study categorized takeover quality into low and high levels by analyzing driving parameters such as lane deviations. Data were collected from a driving simulator study with 81 participants. The first feature input was situation complexity, with three levels decided by raters; the second set of features was the type of NDRTs performed by drivers; and the third set of features was drivers' gazes at the road. Using machine learning algorithms including k-nearest neighbors (kNN), support vector machine (SVM) with radial basis function (RBF) and linear kernel, Naive Bayes, and linear discriminant, they predicted takeover quality with an accuracy of 79% and F1-score of 77%.

However, the above-mentioned models were developed and tested when drivers were engaged in different types of NDRTs (e.g., monitoring vs. reading), where apparent contextual cues existed to discriminate drivers' states. In daily life, even with a specific type of NDRTs such as writing an email, drivers' states can be rather different depending on the importance of the email. Also, some factors deliberately manipulated in the experiment settings, such as emotions, are not easily accessible in the real world. Although advanced wearable technology has made it convenient to collect drivers' physiological signals to reflect their cognitive and emotional states, only gaze behaviors were used in previous studies.

Physiological data can be used to understand drivers' cognitive and emotional states by applying machine learning models to continuously monitored physiological data. The data captured via non-intrusive sensors can be used to build models that estimate drivers' states and their interactions with the driving environments. Drivers' physiological signals combined with environmental factors are promising indicators to predict takeover performance in conditionally automated driving in real time (*Braunagel et al.*, 2017).

1.5 In-vehicle alert systems in automated driving

Human-machine interfaces (HMIs) have been designed and implemented widely in automated vehicles to facilitate the interactions between drivers and vehicles. When the vehicle is driving itself, the in-vehicle HMIs provide drivers with information to conduct non-driving-related tasks (NDRTs) and feedback on the vehicle status and driving environments to develop mental representations of the vehicle. For example, *Schartmüller et al.* (2019) demonstrated that using head-up displays (HUDs) as the NDRT user interface improved productivity and reduced workload and attractiveness in comparison to auditory displays. *Wintersberger et al.* (2019) augmented the traffic objects and upcoming vehicle maneuver via color-coded triangles and arrows, respectively in the automated driving mode. They found that the application of augmented reality could increase automation transparency, foster user acceptance, and improve user trust in fully automated vehicles.

1.5.1 Takeover requests

In Level 3 automated vehicles, drivers are required to take over control of the vehicle when the automation fails. Within a limited period of time, drivers are likely not to be well prepared in critical situations because of their decreased situational awareness in the automation mode. To help drivers take over control of the vehicles, existing literature has extensively studied the design of takeover requests. The most frequent takeover request is the combination of auditory and visual alerts (*McDonald et al.*, 2019a). The visual alert can be a symbol of the wheel or the pedal on the dashboard or the HUD (*Bueno et al.*, 2016; *Du et al.*, 2020d). It indicates that drivers need to take over control of the vehicle by pressing the pedals or rotating the wheel. Text such as “take over” was shown as part of the visual interface in some studies (*Walch et al.*, 2015; *Forster et al.*, 2017). The auditory alert can be speech such as “take over” (*Du et al.*, 2020d; *Ma et al.*, 2020) or a generic tone such as a beep

(*Bazilinskyy et al.*, 2018; *Naujoks et al.*, 2014). Some studies also used vibrotactile alerts as takeover requests (*Eriksson et al.*, 2018; *Petermeijer et al.*, 2017).

In addition to the takeover requests, some in-vehicle interfaces provide event-related takeover information to support drivers in taking over control of the vehicle (*Damböck et al.*, 2012; *Forster et al.*, 2017; *Richardson et al.*, 2018). For example, *Forster et al.* (2017) added semantic speech output alongside a generic warning tone for an upcoming takeover request, and found that the speech output explaining the takeover scenarios shortened the reaction time and was rated as superior for its trust, anthropomorphism, and usability. Similarly, *Damböck et al.* (2012) implemented contact-analogue HUDs in highly automated driving. Road signs recognized by the automation were marked by the frame, and the vehicle ahead and its trajectory were highlighted. The results showed that such displays increased drivers' comprehension of the automation actions and improved driver-automation-cooperation by significantly decreasing reaction times in takeovers.

While these studies showed the promise of providing extra event-related information, other literature went one step further and proposed different types of event-related information and modalities to see how they influenced drivers' takeovers.

1.5.2 Display modality

Like takeover requests, event-related information can be presented in visual, auditory, and vibrotactile modalities. Auditory displays can use abstract sounds (e.g., tones) or voice messages (e.g., speech) to convey event-related information. *Nees et al.* (2016) examined the effects of different types of auditory displays during a video-simulated self-driving car ride with a visual secondary task. Speech alerts, auditory icons, and a visual control condition were employed to give information about the vehicle's operating status and the driving scenarios (i.e., Level 1 situational awareness). The findings showed that speech alerts led to better memory for alerted events and were promising to increase Level 1 situational awareness of routine scenarios. Although both auditory display types resulted in less perceived effort, they did lead to greater perceived annoyance.

HUD has been widely used in automated vehicle environments. Augmented reality (AR) HUD allows direct superimposition of the information onto the real spatiotemporal scene, whereas traditional HUD presents warning icons on the windshield (*Langlois et al.*, 2016; *Langlois and Soualmi*, 2016). *Langlois and Soualmi* (2016) compared classical HUD with AR HUD when taking over from automated driving. The vehicle in front was highlighted via a rectangle in the AR HUD. Drivers were expected to change the lane in the slow traffic scenarios. The findings showed that AR HUD helped drivers better anticipate the lane change maneuver (i.e., increased the distance to the maneuver limit point) and improve driving comfort (i.e., reduced drivers' resulting acceleration and maximum speed).

The vibrotactile modality was studied in the prior literature, but researchers found that it was not effective in conveying complex information (*Meng and Spence*, 2015). For example, *Petermeijer et al.* (2017) compared the effects of takeover modality and directionality on drivers' takeover responses. Participants needed to avoid the stationary vehicle ahead by changing to the left or right lane. The results showed that directional vibrotactile stimuli did not evoke a directional response in uninstructed drivers.

Therefore, the AR HUD and speech are appropriate approaches to provide event-related information in takeover transitions.

1.5.3 Display information

Existing literature has proposed different types of event-related information to potentially support takeovers. For example, *Koo et al.* (2015) categorized speech information about AV action into three types: *how* the car is acting, *why* the car is acting, and *why* + *how* the car is acting. They found that both the *why* only and *why* + *how* information increased AV acceptance compared to *how* only information. Furthermore, the *why* only information led to the highest positive emotional valence while *why* + *how* led to the least positive emotional valence.

Wright et al. (2018, 2017) separated information into environment cues (e.g., crosswalk ahead), threat cues (e.g., scan for pedestrians), and combination cues (e.g., crosswalk ahead, scan for pedestrians). All the cues were in the auditory modality. The results showed that, compared with threat cues, environment cues not only increased the detection of latent hazards, but also prepared drivers to safely mitigate these hazards during takeover transitions. The combination cues showed a similar pattern to the environment cues but were not the most beneficial.

Eriksson et al. (2018) used augmented reality displays to show information and decision suggestions for takeover actions. The objects in the scenarios were highlighted using spheres (information acquisition), lane availabilities were indicated using carpets with different colors (information analysis), and action recommendations were symbolized by arrows with different colors (decision selection). Results showed that the carpet and arrow conditions supported drivers in making a correct lane changing or braking reaction in takeover transitions. Solely highlighting an obstacle using a sphere did not improve decision making, but rather increased unnecessary braking.

Although existing studies employed different criteria to categorize the event-related information, they inherently followed the common rules, such as the information processing theory. Given that the results regarding the effects of different types of display information were mixed, it is necessary to determine the most appropriate display information in different conditions.

1.5.4 Alert system subjective evaluation

Drivers' acceptance of the display technology is one of the deterministic factors in whether the technology will be implemented and used in the vehicle. Researchers have developed models to predict user acceptance. One of the most frequently used models is the Technology Acceptance Model (TAM). TAM posits that Perceived Usefulness (PU) and Perceived Ease of Use (PEOU) are significant predictors of Behavioral Intention (*Davis*, 1989). PU measures the degree to which the technology is perceived as helpful and enhances job performance. PEOU measures the degree to which using a system reduces effort. The TAM has been widely used to evaluate drivers' acceptance of in-vehicle technology, such as navigation systems (*Bolon-Canedo et al.*, 2011; *Chen and Chen*, 2011), on-board monitoring systems (*Ghazizadeh et al.*, 2012), and assistance systems (*Larue et al.*, 2015; *Rahman et al.*, 2017).

Similarly, the Unified Theory of Acceptance and Use of Technology (UTAUT) proposes four components of Behavioral Intention and Actual Behavior: Performance Expectancy, Effort Expectancy, Social Influence, and Facilitating Conditions (*Venkatesh et al.*, 2003). Performance Expectancy is defined as the degree to which an individual believes that using the system will help him or her to attain gains in job performance. Effort Expectancy is the degree of ease associated with the use of the system. Social Influence is defined as the degree to which an individual perceives that important others believe he or she should use the new system. Facilitating Conditions is the degree to which an individual believes that an organizational and technical infrastructure exists to support the use of the system. UTAUT asserts that Performance Expectancy, Effort Expectancy, and Social Influence positively influence Behavioral Intention, and Behavioral Intention and Facilitating Conditions positively influence Actual Behavior. Literature has used UTAUT to model drivers' acceptance towards the advanced driver assistance systems, such as the automated road transport systems (*Madigan et al.*, 2016, 2017).

Many of the studies that have focused on drivers' acceptance of AVs also examined their preference for AV and anxiety (*Koo et al.*, 2015, 2016b; *Nass et al.*, 2005; *Molnar et al.*, 2018; *Shabanpour et al.*, 2018; *Abraham et al.*, 2017). Anxiety is often defined as a feeling of fear, worry, apprehension, or concern. The more effective AV explanations are, the less anxiety someone should have about using the AV (*Koo et al.*, 2015, 2016b). Thus, anxiety should be an outcome used to assess the effectiveness of in-vehicle technology. Preference is defined as the degree to which someone likes or has a fondness for a particular AV technology. Preference is important because all things being equal, individuals may prefer one AV technology over others for reasons that are not always understood (*Abraham et al.*, 2017). If this were true, it would be important to capture preference along with other measures. In all, these outcomes represent either barriers to adoption or positive attitudes associated with a successful adoption.

1.6 Research aim

In response to takeover difficulty in conditionally automated driving, a substantial amount of research has been conducted to investigate drivers' responses to takeover requests and to develop in-vehicle alert systems to facilitate takeover performance. According to the literature review above, there are some research gaps.

First, not all the factors that influence takeover responses were studied comprehensively, and the results of some factors were mixed. Meanwhile, existing studies on drivers' responses to TORs mainly focused on their takeover performance. Little is known about drivers' cognitive load, attention styles, and emotional states amid takeover transitions, which can be reflected through psychophysiological measurements.

Second, it is difficult to estimate drivers' takeover performance in real time only with the identified quantitative relationships. Computational models that are capable of predicting drivers' takeover performance in real time are needed.

Third, without considering drivers' dynamic states and the complexity of driving environments, the existing in-vehicle alert systems issue identical takeover warnings with certain formats all the time. Such design may induce heavy information load or otherwise fail to trigger enough alertness in the face of emergencies.

To fill these research gaps, the following three specific aims are addressed in this dissertation:

(1) We will examine the effects of drivers' cognitive load, emotions, traffic density, and TOR lead time on their driving behavioral (takeover timeliness and quality) and psychophysiological (eye movements, galvanic skin responses, and heart rate activities) responses to takeover requests.

(2) We will develop computational models to predict drivers' takeover performance using drivers' physiological data and driving environment data via machine learning algorithms.

(3) We will design in-vehicle alert systems with different display modalities and information types. To evaluate the displays, we will use human-subject experiments to measure drivers' acceptance of and preference for displays, as well as their anxiety and preparedness in different takeover scenarios.

1.7 Dissertation structure

This dissertation is presented in five chapters. Chapter one provides an introduction to the problem, related work, and aims for this work. Chapter two details three experiments that investigated drivers' driving behavioral and psychophysiological responses to takeover requests in different conditions. Chapter three presents the computational modeling of takeover performance using drivers' physiological data and environment data via machine learning algorithms. Chapter four illustrates the design and evaluation of in-vehicle alert systems for takeovers. Chapter five is an integration of the findings from the previous chapters, and presents the broad impact and intellectual merit of the work, as well as the direction of future work.

CHAPTER 2

Examining Drivers' Behavioral and Physiological Responses to Takeover Requests

2.1 Introduction

To address takeover difficulty, researchers have investigated different factors that influence drivers' takeover responses, including their states, individual characteristics, driving environment, and vehicle capacity (*Eriksson and Stanton*, 2017; *Gold et al.*, 2016; *Wan and Wu*, 2018; *Clark and Feng*, 2017). However, not all the factors were studied comprehensively, and the results of some factors were mixed because of the lack of strict experimental control. Meanwhile, existing studies on drivers' responses to TORs mainly focused on their takeover performance. Little is known about drivers' cognitive load, attention styles, and emotional states during takeover transitions, which can be reflected through psychophysiological measurements. Therefore, we conducted three experiments to examine the effects of different factors on drivers' driving behavioral and physiological responses to takeover requests. These will provide a comprehensive understanding of factors that influence driver takeover in conditional automated driving.

Despite the important role emotions play in human-machine interaction and in manual driving, little is known about how emotions influence drivers' takeover performance. **Experiment 1** aimed to complement the existing literature by examining the effects of drivers' emotional valence and arousal on their takeover performance.

Existing studies on the effects of cognitive load on drivers' takeover performance show mixed findings, as few studies strictly manipulated driver's cognitive load, regardless of whether they explored its interaction effects with other road situation and vehicle status variables. Meanwhile, high traffic density was coupled with fewer escape paths in current literature. With a high traffic density, there were more vehicles in the same or neighboring lanes of the automated ego vehicle, hence restricting the drivers' action opportunities. There is a need to decouple traffic density and the availability of escape paths. Thus, **Experiment 2** aimed to study the influence of drivers' cognitive load, oncoming traffic, and TOR lead time on their behavioral and psychophysiological responses to takeover requests.

Despite existing investigations, the impact of takeover scenarios has not been thoroughly examined. The few studies investigating the effects of takeover scenarios were based on only a few scenarios (*Naujoks et al.*, 2017; *Dogan et al.*, 2019), which were not representative of real-world situations. Furthermore, the existing scenario categorizations that separate braking and steering scenarios may not be ideal, because braking and steering could happen at the same time (*Eriksson et al.*, 2018). To fill the research gap, **Experiment 3** aimed to systematically examine the effect of takeover scenarios and vehicle speed on drivers' behavioral and physiological responses to takeovers.

In the experiments, participants drove a simulated vehicle with Level 3 automation. Each of them experienced takeover events in different conditions, and their takeover performance and psychophysiological responses were recorded and analyzed. The methods and results of each experiment are discussed and summarized in detail below.

2.2 Experiment 1

Experiment 1 aimed to complement the existing literature by examining the effects of drivers' emotional valence and arousal on their takeover performance.

2.2.1 Method

This research complied with the American Psychological Association code of ethics and was approved by the Institutional Review Board of the University of Michigan.

2.2.1.1 Participants

A total of 32 university students (average age = 21.4 years, SD = 2.9; 17 females and 15 males) with normal or corrected-to-normal vision participated in the experiment. Participants were screened for valid US driver license status and susceptibility to simulator sickness. The study lasted 60 to 75 minutes, and each participant was compensated with $30 upon completion of the experiment.

2.2.1.2 Apparatus and stimuli

The study was conducted using a fixed-based desktop driving simulator. The simulator ran the SimCreator driving simulation engine from Realtime Technologies Inc. (RTI, Michigan, USA) (Figure 2.1). To present the driving environment to participants, forward road scenes were displayed on a 32-inch computer monitor about 2.5 feet in front of the driver. A rear-view image was also displayed in a separate window on the forward screen. The simulated vehicle was controlled by a Logitech steering wheel and pedal system connected via a USB interface to the SimCreator components.

In this study, the automation features of the driving simulator were programmed to simulate an SAE Level 3 AV, wherein the AV performed the longitudinal and lateral vehicle control, navigated, and responded to traffic control devices and other traffic

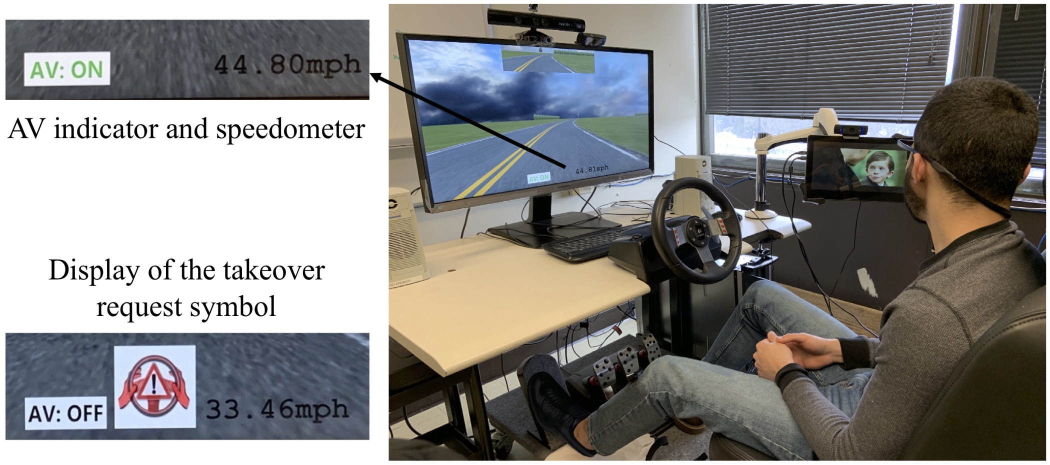


Figure 2.1: Illustration of driving simulator in the experiment

elements, and the driver was not required to actively monitor the driving environment. However, there were unexpected events that the AV could not handle, and it would request the driver to take over control of the vehicle.

2.2.1.3 Experimental design

The experiment used a within-subjects design in order to minimize effects of extraneous variables and to increase statistical power. We induced different values of emotional valence and arousal within each participant, covering the four quadrants of the valence-arousal space (Figure 1.1).

As shown in Table 2.1, eight 4-minute movie clips were selected for emotion induction based on prior literature (*Lisetti and Nasoz*, 2004; *Gross and Levenson*, 1997; *Uhrig et al.*, 2016; *Ekman et al.*, 1980). To minimize the ordering effect, the sequence of the emotion induction conditions was counterbalanced using a Latin square design. Among all the participants, 10 participants had previously watched 1 out of the 8 movies, and 5 participants watched 2 out of the 8 movies before the study. Based on participants' comments in the post-experiment debriefing sessions, they were actively engaged in watching the movie clips.

Drivers were assured that there was no need to monitor the driving environment

Table 2.1: Descriptions of movie clips used for emotion induction

Emotion	Movie	Scene	Citation
Sad	The Champ	Death of the Champ	*Lisetti and Nasoz* (2004)
	Finding Neverland	Death of the boy's mother	*Uhrig et al.* (2016)
Angry	Schindler's list	Woman engineer being shot	*Lisetti and Nasoz* (2004)
	Cry Freedom	Innocent people being shot	*Gross and Levenson* (1997)
Happy	Bruce Almighty	Man getting power from God	*Uhrig et al.* (2016)
	Big Fish	Boy expressing love to the girl	*Uhrig et al.* (2016)
Calm	ScreenPeace screensaver with city scenes		*Gross and Levenson* (1997)
	Beautiful trees and flowers in the world		*Ekman et al.* (1980)

Table 2.2: Descriptions of takeover events

Environment	Event descriptions
Urban	Construction zone ahead
Urban	Bicyclist in the lane ahead
Rural	Police vehicle on shoulder
Rural	Swerving vehicle ahead

when the AV was in the automation mode, and they could focus on watching the videos until a TOR was issued. The TOR was in both auditory and visual format: an audible spoken phrase "takeover" and an icon representing a pair of red hands on a red steering wheel (*Kuehn et al.*, 2017) (Figure 2.1). The sound volume and visual intensity were the same for all the participants throughout the experiment. Each participant confirmed in the practice drives that the TOR could be heard clearly while the participant was watching the movie clips. The speedometer, the AV mode indicator, and the TOR symbol were displayed in real time at the lower center of the screen.

Each participant went through four takeover events. The takeover events were designed based on prior literature (*Rezvani et al.*, 2016; *Koo et al.*, 2016b; *Miller et al.*, 2016) (Table 2.2). In the AV mode, the vehicle always drove in the right lane. The TOR was issued when the AV was around 4 seconds away from the construction zone/bicyclist/police vehicle/swerving vehicle. Participants were expected to change to the left lane during the takeover transitions.

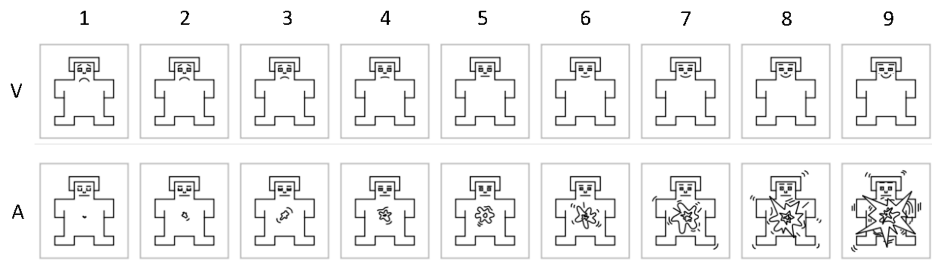


Figure 2.2: The Self-Assessment Manikin (SAM) (*Bradley and Lang*, 1994)

2.2.1.4 Measures

We measured participants' subjective ratings of emotional valence and arousal, as well as objective measures of their takeover performance.

The Self-Assessment Manikin (SAM) instrument (See Figure 2.2) was used to assess participants' emotional valence (1 = extremely negative, 9 = extremely positive) and arousal (1 = lowest arousal, 9 = highest arousal) based on their emotions induced by the movie clips (*Bradley and Lang*, 1994).

Takeover performance was assessed in the timing and quality aspects. Takeover time was calculated as the time between the TOR and the start of the maneuver. The start of the maneuver is defined as a 2-degree change in steering wheel angle or a 10% depress of pedals, whichever is first (*Gold et al.*, 2016). Takeover quality was assessed by three representative driving measures in the obstacles ahead scenarios: maximum resulting acceleration, maximum resulting jerk, and minimum TTC within the time window between the TOR and the end of the lane changing behavior (i.e., the center of the vehicle reached the boundary of the other lane). Consistent with prior research (*Hergeth et al.*, 2017), maximum resulting acceleration was calculated as $max\ acceleration_{resulting} = \max_t \sqrt{acceleration^2_{longitudinal} + acceleration^2_{lateral}}$. A smaller acceleration represents a smoother and safer reaction to TORs. In addition, we calculated the maximum resulting jerk as $max\ jerk_{resulting} = \max_t \sqrt{jerk^2_{longitudinal} + jerk^2_{lateral}}$. Jerk is the derivative of acceleration and has been utilized to evaluate shift quality, ride comfort (*Huang and Wang*, 2004) and driving aggressiveness (*Bagdadi and Várhelyi*, 2011; *Feng et al.*, 2017; *Bagdadi and Várhelyi*, 2013). Similarly, a smaller jerk represents higher takeover quality. TTC is a time-based safety indicator for detecting rear-end collision risk, and is defined as the time taken for two objects to collide if they maintain their present speeds and trajectories (*Hayward*, 1972).

Five crashes occurred in the study. Under such situations, minimum TTC was treated as "not applicable", and other driving dynamic variables were calculated using the time window between the TOR and the time when drivers re-engaged the automation mode.

2.2.1.5 Experimental procedure

Once participants arrived at the lab, they signed an informed consent and filled a demographic form. Next, participants received a 5-minute training, where they practiced how to change lanes and engage/disengage the automated driving mode by pressing a button on the steering wheel. They were asked to comply with all the traffic laws when they drove manually, and the speed limit was 35 mph. Participants also experienced an unexpected takeover event in the practice while watching a 1-minute movie clip of Zen Garden. The movie clip was played on a tablet located on the right side of the driver's seat. The takeover event was the scenario where the traffic lights at an intersection did not work, and required the driver to observe the surroundings and drive manually. Participants were asked to re-engage the AV once they had negotiated the drive.

After the training session, participants completed two drive courses, each containing two takeover events. As shown in Figure 2.3, each course began with the command to activate the automated driving mode. Then there was an emotion induction phase when participants were asked to watch two 4-minute movie clips aimed at inducing the same emotion. Close to the end of the movie clips, a takeover request was issued,

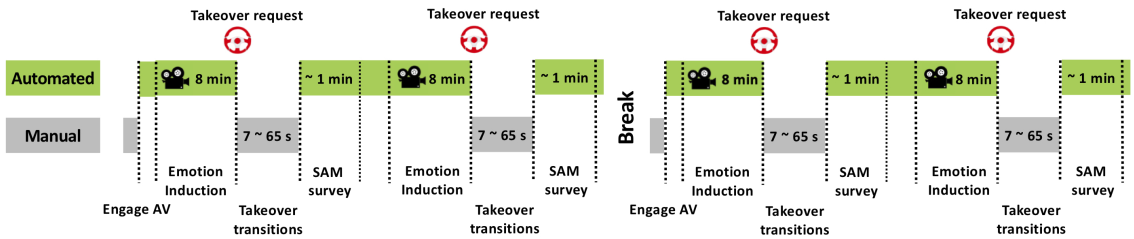


Figure 2.3: Sequence of takeover events in the experiment

and participants were required to take over control of the vehicle immediately. Once participants negotiated the drive, they could hand over the control back to the AV. After participant re-engaged the AV mode, they were asked to recall the scenes in the movie clips and complete the SAM survey to indicate their emotional valence and arousal when watching the movie clips.

2.2.1.6 Data analysis

Data from one participant was excluded from analysis as the participant did not follow the instructions from the experimenter. The SAM Likert scales from 1 to 9 (low arousal: 1-4, high arousal: 6-9; negative valence: 1-4; positive valence: 6-9) were normalized to a scale from -1 to 1 (*Miranda Correa et al.*, 2018). Data points with 0 valence or 0 arousal were deleted. Figure 2.4 shows the distribution of valence and arousal values in the four quadrants: positive valence high arousal, negative valence high arousal, negative valence low arousal, and positive valence low arousal, respectively.

We used a mixed linear model to analyze the relationship between valence, arousal and takeover performance (timeliness and quality). Results are reported as significant for α less than .05.

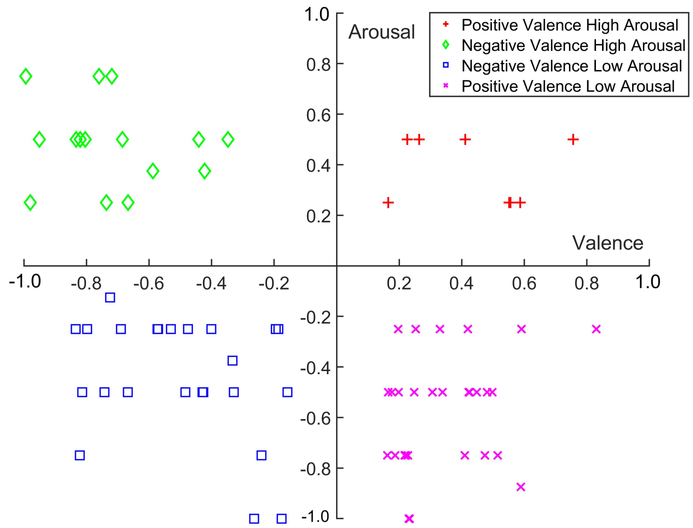


Figure 2.4: Raw data of subjective ratings in the valence-arousal plane

Table 2.3: Mean and standard error values of dependent measures

	Negative Valence		Positive Valance	
	Low arousal	High arousal	Low arousal	High arousal
Takeover time (s)	$1.88 \pm .09$	$1.69 \pm .07$	$1.79 \pm .07$	$1.88 \pm .16$
Max resulting acceleration (m/s^2)	$6.56 \pm .75$	$6.14 \pm .81$	$5.85 \pm .62$	$3.95 \pm .79$
Max resulting jerk (m/s^3)	113 ± 18	115 ± 21	94 ± 17	42 ± 16
Min TTC (s)	$.98 \pm .15$	$.77 \pm .15$	$.72 \pm .12$	$.67 \pm .15$

2.2.2 Results

Table 3 summarizes the mean and standard error (SE) values of the dependent measures.

2.2.2.1 Takeover time

No significant effect was found for either valence ($F(1, 57) = .04, p = .84$) or arousal ($F(1, 76) = .32, p = .57$), and the interaction effect was not significant ($F(1, 63) = .47, p = .50$).

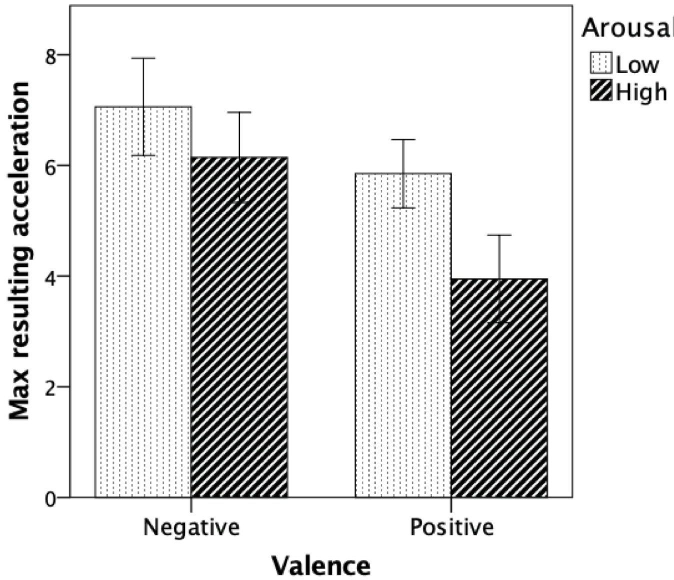


Figure 2.5: Maximum resulting acceleration (m/s^2)

2.2.2.2 Takeover quality

With regard to the maximum resulting acceleration, there was a significant effect of valence ($F(1,56) = 4.26, p = .04$). Positive valence led to a smaller maximum resulting acceleration (Figure 2.5). In addition, there was a trend that high arousal led to a smaller maximum resulting acceleration ($F(1,77) = 3.24, p = .08$). The interaction effect of valence and arousal on maximum resulting acceleration was not significant ($F(1,64) = .79, p = .38$).

There was a significant effect of valence on maximum resulting jerk ($F(1,55) = 6.47, p = .01$), with positive valence leading to a smaller maximum resulting jerk (Figure 2.6). The main effect of arousal ($F(1,73) = 1.84, p = .18$) and the interaction effect were non-significant ($F(1,61) = 1.71, p = .20$).

There were no significant effects of valence ($F(1,57) = 1.19, p = .28$) and arousal ($F(1,75) = .67, p = .42$) on TTC_{min}. The interaction effect was not significant ($F(1,65) = .31, p = .58$).

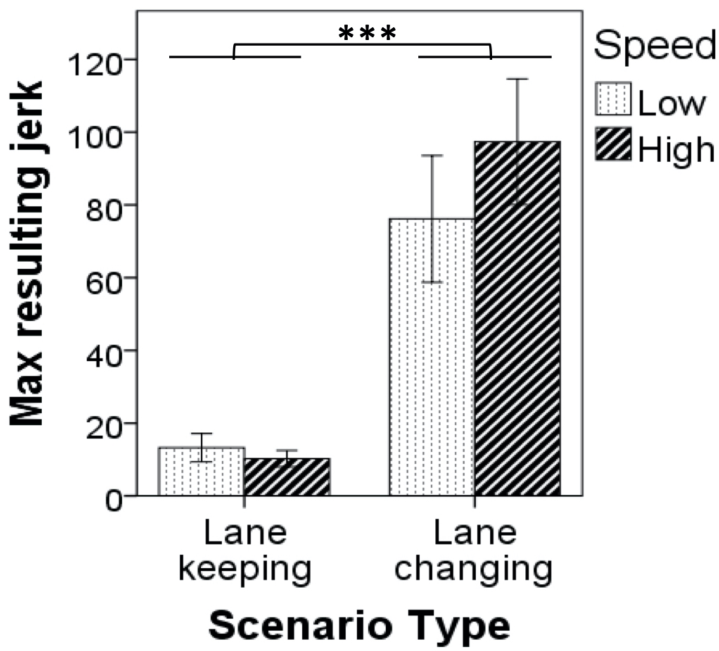


Figure 2.6: Maximum resulting acceleration (m/s^3)

2.2.3 Discussion

This study aimed to examine how emotions affected drivers' takeover performance in conditionally automated driving. Drivers' tasks in manual driving and conditionally automated driving (Level 3 AV) are fundamentally different. In manual driving, drivers continuously perform lateral and longitudinal control. Therefore, prior studies in manual driving mainly treated the NDRTs as distractions. With conditionally automated driving, however, drivers largely perform a single task (i.e., either the driving task or the NDRT) at any one time. When the automation mode is on, drivers can perform any NDRT at their own discretion. After receiving a TOR, drivers are expected to relinquish the NDRT and resume the driving task immediately. The distinction between manual driving and conditionally automated driving suggests findings in manual driving cannot be directly applied to takeover transitions.

2.2.3.1 Effects of valence

Our findings showed that positive valence led to better takeover performance, reflected by a smaller maximum resulting acceleration and a smaller maximum resulting jerk. Smaller maximum resulting acceleration and maximum resulting jerk are associated with higher levels of safety (*Hergeth et al.*, 2017), shift quality, and ride comfort (*Huang and Wang*, 2004), and lower driving aggressiveness (*Bagdadi and Várhelyi*, 2011; *Feng et al.*, 2017; *Bagdadi and Várhelyi*, 2013). The findings can be explained by the "broad-and-build" theory (*Fredrickson and Branigan*, 2005; *Rowe et al.*, 2007), which states that positive emotions prompt individuals to broaden their focus of attention and their thought-action repositories. The broader span of attention enables drivers to perceive and process different stimuli in the traffic situation and avoid tunnel vision. Larger thought-action repositories allow drivers to identify a more appropriate action given a specific traffic situation. In takeover transitions, the broadened attention and thought-action repositories aided drivers in traffic situation comprehension and action selection, and hence enhanced the takeover quality. The results are in line with some studies examining the effects of valence in manual driving, where positive valence led to better vehicle control (*Chan and Singhal*, 2013; *Hancock et al.*, 2012; *Trick et al.*, 2012; *Groeger*, 2013), suggesting that the benefits of positive valence can be carried over from manual driving to automated driving. However, we failed to find any difference in minimum TTC, and the reason could be explained as follows. Time to collision represents the time taken for two objects to collide with each other, and is an indicator of collision risk. With a negative emotion, drivers' attentional focus and thought-action repositories are narrowed. Therefore, they might employ immediate survival-oriented behaviors and brake abruptly, leading to a potentially larger minimum TTC.

In the present study, we adopted three metrics aimed at assessing driving smoothness, ride comfort and collision risk. However, we notice the wide range of metrics used to measure takeover quality in prior literature, including crash rates, different statistics of velocity, acceleration, jerk, and TTC (Please refer to *McDonald et al.* (2019a) for the detailed list). This wide range of metrics makes it difficult to summarize findings in prior literature. There is an urgent need to examine if it is possible to propose a standard set of metrics for measuring takeover quality, and how to do so.

2.2.3.2 Effects of arousal

We found that there was no significant effect of arousal on takeover time. Prior research in manual driving showed that high arousal led to faster response time in hazard detection (*Trick et al.*, 2012; *Navarro et al.*, 2018; *Ünal et al.*, 2013). In takeover transitions, however, TORs serve as an attention management tool. Upon receiving a TOR, drivers are required to attend to the driving task. Moreover, in our study, drivers were engaged in a hands-free task (i.e., watching movies) before the TOR was issued. Without the need to physically end the task and put down the NDRT device, drivers could immediately switch their attention from the tablet to the driving scene. Our results showed that this process took less than 2 seconds, no matter in which emotional state drivers were. Recent studies comparing different types of NDRTs on takeover quality and timeliness showed that the types of NDRTs only influenced the takeover quality and not takeover time (*Gold et al.*, 2016; *Körber et al.*, 2016; *Zeeb et al.*, 2016; *Bueno et al.*, 2016; *Zeeb et al.*, 2017), providing further support for our findings.

To our surprise, the results suggest a trend that high arousal led to a smaller maximum resulting acceleration and thus better takeover quality. This implies that the benefits of high arousal in mobilizing attentional resources and effort for immediate actions could be reflected in takeover quality. Further research is needed to elucidate this effect.

2.2.3.3 Implications

Our findings have implications on the design of in-vehicle emotion regulation systems. Advances in machine learning enable accurate detection of drivers' emotional states in real time (*Izquierdo-Reyes et al.*, 2018; *Picard*, 2003). For example, if the system detects that a driver is a negative emotional state, strategies such as reappraisal and distraction (*Naragon-Gainey et al.*, 2017) can be used to help the driver manage his or her negative emotion, resulting in better takeover performance. Also, the AV could even delay or avoid handing over control.

2.2.3.4 Limitations and future studies

The present study has several limitations that should be taken into consideration. First, we only examined one type of takeover event, wherein the drivers perceived certain hazards and were expected to change lanes. Further research could be extended to other types of takeover events such as lane markings disappearing. Second, the study was conducted in a fixed-based desktop driving simulator with limited fidelity. Future studies can investigate the effects of emotional valence and arousal in a higher fidelity laboratory environment or a naturalistic driving environment. Third, drivers' emotional valence and arousal values were queried after a takeover event occurred. Although we followed a standard practice and asked the drivers to recall the movie clips, and based on which to indicate their emotional states prior to the takeover event, experiencing the event per se might influence drivers' perceptions of the movie clips. Further research could employ physiological measures of emotion, which could indicate drivers' emotional states non-intrusively before a takeover event. Meanwhile, eye-tracking metrics such as gaze behaviors could be recorded and analyzed in order to better understand drivers' attention allocation during conditionally automated driving.

2.2.4 Conclusion

Drivers have difficulty in takeover transitions, as they become increasingly decoupled from the operational level of driving. In response to this challenge, researchers have started to look into factors that could influence drivers' takeover performance. Despite the important role emotion plays in human-machine interaction and in manual driving, little is known about how emotion affects takeover performance in conditionally automated driving. By systematically manipulating drivers' emotional states, the current study extended earlier research by demonstrating how valence and arousal influence takeover time and quality. The benefits of positive emotions carry over from manual driving to conditionally automated driving, while the benefits of arousal do not. Moreover, our study provides empirical evidence that emotions we cannot simply apply the findings about emotions in manual driving to automated driving.

2.3 Experiment 2

Experiment 2 aimed to study the influence of drivers' cognitive load, oncoming traffic, and TOR lead time on their behavioral and psychophysiological responses to takeover requests.

2.3.1 Method

This research complied with the American Psychological Association code of ethics and was approved by the Institutional Review Board of the University of Michigan.

2.3.1.1 Participants

A total of 102 university students participated in the study (mean age = 22.9, standard deviation [SD] = 3.8; range = 18-38; 40 females and 62 males). All participants had normal or corrected-to-normal vision and a valid driver's license. On

Table 2.4: Participants' distribution of annual mileage and weekly mileage and average experience score with different driver assistance systems

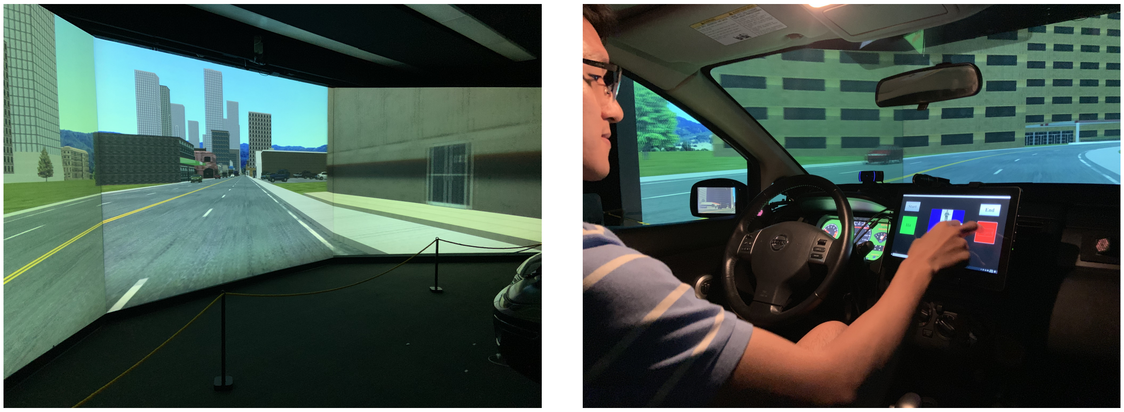
Annual mileage	N	Weekly mileage	N	Driving assistance system	Score
Less than 5,000 miles	34	Less than 50 miles	53	Cruise control	3.0
5,000 - 10,000 miles	33	50 - 100 miles	27	Adaptive cruise control	1.5
10,000 - 15,000 miles	25	100 - 150 miles	8	Lane-departure warning	1.8
15,000 - 20,000 miles	2	150 - 200 miles	6	lane-keeping assistance	1.5
20,000 - 25,000 miles	5	200 - 250 miles	8	Collision warning	1.9
More than 25,000 miles	3	More than 250 miles	2	Emergency braking	1.4

average, participants had held their driver's licenses for 4.9 years (SD = 3.2 years). Each participant received a compensation of $30 for about an hour of participation. A 5-point Likert scale was used to measure participants' experiences with various driver assistance features (1 indicated "never" and 5 indicated "always"). Table 2.4 showed participants' distribution of annual mileage and weekly mileage, as well as their average experience scores with different driver assistance systems.

2.3.1.2 Apparatus and stimuli

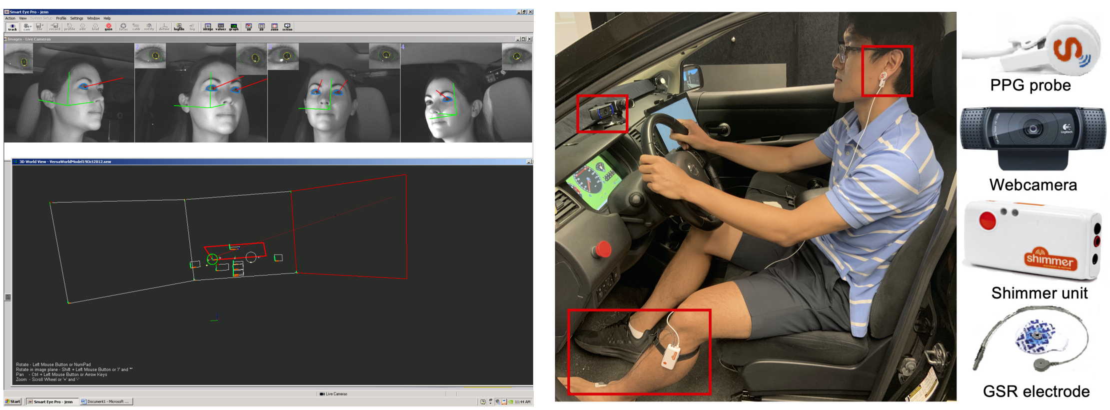
The study was conducted in a fixed-base driving simulator from Realtime Technologies Inc. (RTI, Michigan). The virtual world was projected on three front screens (16 feet away), one rear screen (12 feet away), and two side mirror displays (See Figure 2.7). There was a steering wheel and pedal system embedded in a Nisan Versa car model. The vehicle was programmed to simulate an SAE Level 3 automation, which handled the longitudinal and lateral control, navigation, and responded to traffic events. Participants could press the button on the steering wheel to activate the automated mode and engage in NDRTs. However, the automated mode would be deactivated automatically for drivers to take over control once the automated system failed to respond properly. At that moment, drivers would be alerted by an auditory warning "Takeover".

This simulator was equipped with a Smart Eye four-camera eye-tracking system (Smart Eye, Sweden) that provided live head-pose, eye-blink, and gaze data (Figure



(a) Driving simulator (b) In-vehicle setting

Figure 2.7: Driving simulator and Smarteye system



(a) (b)

Figure 2.8: (a) Smarteye. (b) Shimmer3 GSR+ unit and Logitech web camera.

2.8a). The sampling rate of the eye-tracking system was 120 Hz. The Shimmer3 GSR+ unit (Shimmer, MA, USA), including GSR electrodes and photoplethysmographic (PPG) probe, was used to collect GSR and HR data, with a sampling rate of 128 Hz. A Logitech web camera with a sampling rate of 30Hz was used to collect drivers' facial expressions (Figure 2.8b). The iMotions software (iMotions, MA, USA) was used for psychophysiological data synchronization and visualization in real time.

The NDRT utilized in the study was a visual N-back memory task (*Jaeggi et al.*, 2008). The stimulus consisted of nine (3×3) squares with two human figures randomly appearing in two out of the nine squares. Each stimulus was presented for 500 ms in

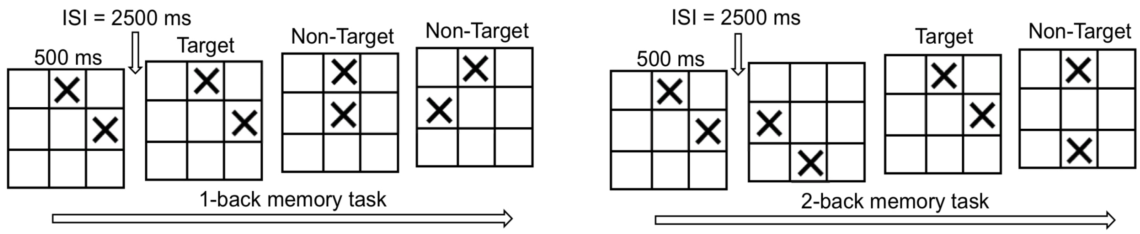


Figure 2.9: N-back memory task.

sequence with a 2500 ms interval (Figure 2.9). Participants were required to press the "Hit" button when the current stimulus was the same as the one presented N steps back in the sequence, and press the "Reject" button otherwise. With different N values (i.e., 1 and 2), participants were exposed to conditions with different cognitive load but the same manual and visual load. The task was running on an 11.6-inch touch screen tablet mounted in the center console of the vehicle.

2.3.1.3 Experimental design

The study employed a within-subjects design, with the driver's cognitive load, TOR lead time, and traffic density as independent variables. The cognitive load was manipulated via the difficulty of the NDRTs (low: 1-back memory task; high: 2-back memory task). The heavy- and no- traffic conditions had 15 and 0 oncoming vehicles per kilometer, respectively. The TOR lead time was around 4 or 7 seconds. Based on prior literature (*Rezvani et al.*, 2016; *Koo et al.*, 2016b; *Miller et al.*, 2016; *Molnar et al.*, 2018), eight takeover events were designed in urban and rural drives with typical roadway features (Table 2.5). The difficulty of the scenarios was designed to be approximately the same.

The order of cognitive load, traffic density, and TOR lead time was counterbalanced via an 8×8 balanced Latin Square across participants. Based on standard programming practices for the simulator, the order of scenario presentations was counterbalanced by having half of the participants drive from Event 1 to 8, and the

Table 2.5: Descriptions of takeover events

Event	World	Scenario type	Event descriptions
Event 1	Urban	Lane changing	Bicyclist in the lane ahead
Event 2	Urban	Lane keeping	Construction zone on the left
Event 3	Urban	Lane changing	Construction zone ahead
Event 4	Urban	Lane keeping	Sensor error on the curve
Event 5	Rural	Lane changing	Swerving vehicle ahead
Event 6	Rural	Lane keeping	No lane markings on the curve
Event 7	Rural	Lane keeping	Sensor error on the curve
Event 8	Rural	Lane changing	Police vehicle on shoulder

other half from Event 8 to 1. During the entire session, there were no other vehicles moving in the drivers' direction, and participants could avoid the objects in their lane by changing to the adjacent lane.

2.3.1.4 Dependent variables

We collected drivers' vehicle-related, NDRT-related, and psychophysiological measures (Table 2.6).

Takeover performance was measured in two aspects: timeliness and quality. With regard to takeover timeliness, takeover reaction time and eyes-on-road reaction time were measured. Takeover reaction time was defined as the time between the TOR and the start of maneuver. According to (*Gold et al.*, 2016), the threshold of maneuver was set as a 2 degree change of the steering wheel angle or a 1% change of the pedals, whichever was quicker. Eyes-on-road reaction time was the time between the TOR and the driver's first gaze vector on the road (*Eriksson et al.*, 2018).

Takeover quality was assessed by three measures: maximum resulting acceleration, maximum resulting jerk, and minimum TTC within the time window between the TOR and the end of the takeover process. The end of the takeover process was defined as either the endpoint of each takeover scenario or when participants re-engaged the vehicle, whichever was earlier. Consistent with prior research (*Gold et al.*, 2016), the endpoint of Events 1, 3, 5, and 8 was when the vehicle's center of gravity reached the boundary of the neighboring lane; the endpoint of Events 4, 6, and 7 was when the

Table 2.6: Dependent Variables

Dependent variables	Unit	Category	Explanation
Takeover reaction time	Seconds	Takeover timing	Time between TOR and start of maneuver
Eyes-on-road reaction time	Seconds	Takeover timing	Time between TOR and eyes on road
Maximum resulting acceleration/jerk	m/s^2, m/s^3	Takeover quality	Maximum resulting acceleration/jerk during takeover situation
Minimum time to collision	Seconds	Takeover quality	Minimum time to collision during takeover situation
Road deviation standard error (SE)	Centimeters	Takeover quality	Standard error of center road deviation during takeover situation
Reaction time in N-back task	Seconds	NDRT	The reaction time for the N-back memory task
Accuracy in N-back task	Percentage	NDRT	The accuracy for the N-back memory task
Heart rate variability	millisecond	Heart rate	Standard deviation of inter-beat-interval
Difference in average heart rate	beat per minute	Heart rate	Difference in average heart rate between NDRT and takeover stage
Mean phasic GSR	micro Siemens	GSR	Average GSR phasic activation
Maximum phasic GSR	micro Siemens	GSR	Maximum GSR phasic activation
Eyes-on-road time	percentage	Gaze behaviors	The time percentage while eyes are on the road
Blink number		Gaze behaviors	The number of blinks
Horizontal gaze dispersion	radian	Gaze behaviors	The standard deviation of gaze heading
Emotional valence	-100 to 100	Facial expressions	Signs indicate positive or negative emotions
Emotional engagement	0 to 100	Facial expressions	Increasing values signify increased emotional engagement

driver passed the exit point of the curve; and the endpoint of Event 2 was when the vehicle passed the construction zone location. However, participants were instructed to re-engage the vehicle as soon as they thought the vehicle was able to drive on its own. We used the same equations to calculate maximum resulting acceleration, maximum resulting jerk, and minimum time to collision as Experiment 1, which was introduced in Section 2.2.1.4.

In the present study, participants were informed that the NDRT would not stop automatically when an TOR was issued, and were instructed to end the NDRT by themselves. This setting was similar to the "no lockout" condition in the study by *Wandtner et al.* (2018), where participants had to make a trade-off decision between terminating the NDRT and taking over control of the vehicle immediately (*Wandtner et al.*, 2018). Drivers' reaction time and accuracy in NDRT were used as checks for cognitive load manipulation.

The psychophysiological measures included drivers' gaze behaviors, HR activities, GSRs, and facial expressions. PPG peaks were detected using an adaptive threshold method for heart rate extraction (*Shin et al.*, 2009). Heart rate variability was calculated as the standard deviation of RR intervals (i.e., the time elapsed between two successive R-waves on the electrocardiogram) (*Castaldo et al.*, 2017). In addition to directly measuring drivers' average heart rate in takeover stages relative to the NDRT stage, we also categorized such heart rate differences into three patterns, because they can reflect drivers' attentional styles during transitions, as introduced before. Heart rate acceleration/deceleration was defined as at least 2 heart beats per minute (bpm) increase/decrease from the NDRT stage to the takeover stage. No changes in heart rate indicated less than 2 bpm changes between the two stages (*Pohlmeyer and Coughlin*, 2008; *Reimer et al.*, 2011).

The raw GSR signals were decomposed into phasic and tonic components using the continuous decomposition analysis (CDA) via Ledalab in Matlab (*Benedek and

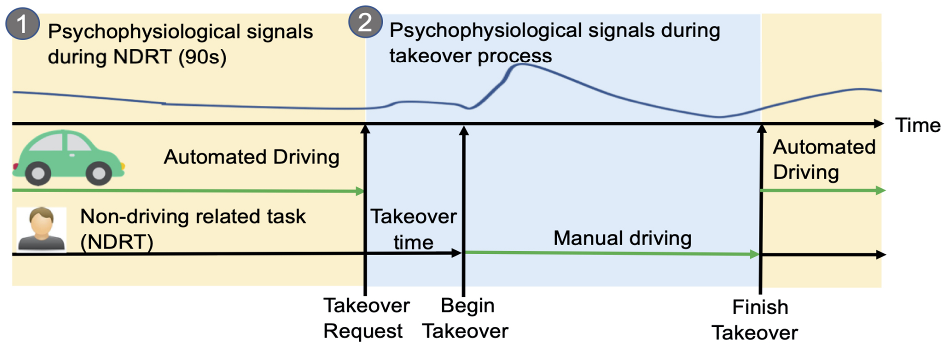


Figure 2.10: Two time windows to calculate measures from psychophysiological signals.

Kaernbach, 2010). Then maximum and mean phasic components were calculated for further analysis, as they were responsible for relatively rapid changes in response to specific events in the takeover transitions (*Wintersberger et al.*, 2018). For gaze behaviors, we calculated drivers' eyes-on-road time percentage, blink number, and horizontal gaze dispersion. Horizontal gaze dispersion was defined as the standard deviation of gaze heading. Drivers' emotional valence and engagement were extracted from their facial expressions using the iMotions Affectiva module to reflect how positive/negative and expressive their emotions were (*Stöckli et al.*, 2018; *Kulke et al.*, 2020).

We calculated the above-mentioned statistical psychophysiological measures using two time windows: the NDRT stage and the takeover stage (see Figure 2.10). The NDRT stage was approximately 90 seconds long. It started when the drivers were asked to initiate the NDRT and ended when the auditory "Takeover" alert was issued. The takeover stage started with "Takeover" alerts and ended when drivers negotiated takeover events and re-engaged the vehicle. In order to show the continuous takeover transition process, we also depicted the psychophysiological measures after TORs second by second when their main effects were significant.

2.3.1.5 Experimental procedure

After participants signed an informed consent form and completed an online demographics questionnaire, they were asked to track six targets on the front screen for eye-tracking calibration. Next, two GSR electrodes were attached to their left feet and the PPG probe to their left ear lobes. They were then introduced the N-back task and automated driving features of the simulator. Participants were told that there was no need to actively monitor the driving environments or take over control of the vehicle when no TOR was issued, as the vehicle was able to handle the situations by itself.

Participants had a 2-minute practice for the N-back memory task, followed by a 5-minute practice drive using the AV. Participants were informed that they would get an additional 20 dollars if their NDRT performance in the real experiment was ranked among the top 10. In the practice drive, they were asked to change lanes, engage the AV, perform the N-back task, and take over control of the vehicle. The takeover event was a scenario where traffic lights did not work and required drivers to observe the surroundings and take over control of the vehicle. During driving, participants were instructed to maintain the speed limit of 35 mph in urban and rural environments and 65 mph in highway environments and obey all traffic rules.

Each participant drove two experimental drives (15-20 minutes each), each containing four takeover events. At the beginning of the drives, participants were asked to activate the AV mode and then start the N-back task when the audio command "Please start the NDRT" was issued. After about a 90-second NDRT, a TOR was issued unexpectedly, and participants were required to terminate the NDRT manually and take over the control immediately. When participants thought they had negotiated the takeover event, they were free to activate AV mode. The experimenter would remind participants to re-engage the vehicle if they did not turn on the automation after the takeover event. The operation of NDRT, takeover, AV mode activation

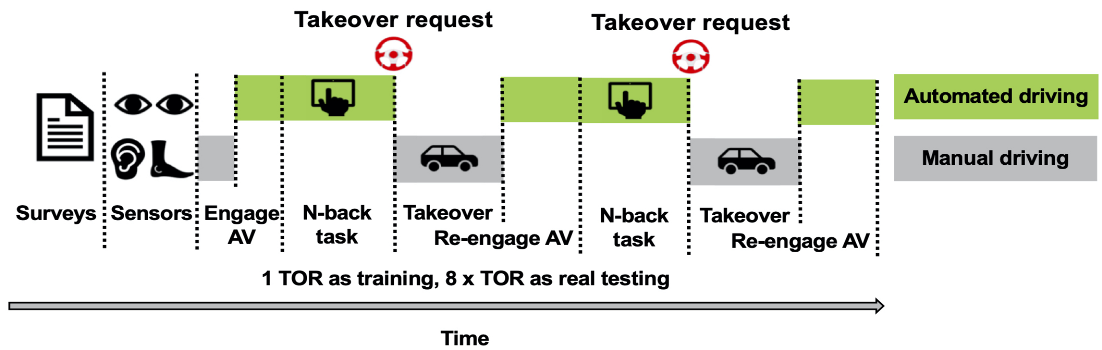


Figure 2.11: Sequence of events in one drive

process were repeated for each takeover event (See Figure 2.11).

2.3.1.6 Data analysis

Each participant experienced 8 scenarios, so 102 participants yielded a total of 816 (8×102) scenarios. Due to some participants' motion sickness and malfunctions of driving simulator and psychophysiological sensors (e.g., calibration failure of steering wheel and eye-tracking system, system freezing), 683 scenarios were available for further psychophysiological analysis and 640 scenarios were available for driving behavioral analysis.

Two types of linear mixed models were conducted using SPSS version 24 to examine effects on continuous dependent variables (Table 2.6). The first one used cognitive load, TOR lead time, traffic density, and their interactions as fixed effects, and the second one used time window (NDRT stage vs. takeover stage) as the fixed effect (only for psychophysiological measures, since driving behaviors only happened in takeover stage). Subjects were treated as random effects to resolve non-independence in all the models. Levene's tests were conducted to examine the assumption of homogeneity of variance. All the dependent variables showed equal variance across the cognitive load, traffic density, and TOR lead time levels. Although the Shapiro–Wilk tests showed that the assumption of normality was violated for some dependent variables (e.g., horizontal gaze dispersion), we argued that linear

Table 2.7: Mean and standard error of driving behaviors under ∼4s TOR lead time ($Mean \pm SE$)

	∼4s TOR lead time			
	No traffic		Heavy oncoming traffic density	
	Low cognitive load	High cognitive load	Low cognitive load	High cognitive load
Takeover RT (s)	2.32 ± 0.14	2.26 ± 0.14	2.2 ± 0.13	2.23 ± 0.11
Eyes-on-road RT (s)	1.05 ± 0.11	1.17 ± 0.09	0.84 ± 0.1	1.33 ± 0.09
Max resulting acceleration (m/s^2)	3.29 ± 0.31	3.21 ± 0.26	3.22 ± 0.29	3.71 ± 0.3
Max resulting jerk (m/s^3)	44.1 ± 10.1	36.2 ± 8.0	43.5 ± 8.6	45.2 ± 8.0
Minimum TTC (s)	1.39 ± 0.21	1.46 ± 0.16	1.3 ± 0.16	1.22 ± 0.18
Road deviation SE (cm)	1.00 ± 0.15	0.84 ± 0.07	0.81 ± 0.08	0.85 ± 0.07

mixed models can still be conducted because they are robust against violations of the assumptions of normality (*Gelman and Hill*, 2006). Meanwhile, if the main effects of independent variables on psychophysiological measures during the takeover stage were significant, we used pairwise t-tests to compare psychophysiological measures after TORs second by second to provide time-series insights. Since heart rate change pattern was a categorical variable, we used the chi-squared test to examine its dependence with independent variables, which could represent drivers' attentional styles in different conditions (*Pohlmeyer and Coughlin*, 2008; *Reimer et al.*, 2011). To increase the interpretation of psychophysiological results, Pearson correlation coefficients were examined to explore the relationships between emotions and physiological data. The significance level alpha was set at .05.

2.3.2 Results

2.3.2.1 Manipulation check

There were significant main effects of cognitive load on drivers' reaction time and accuracy in the NDRTs, $F(1, 552) = 37.5$, $p < .001$; $F(1, 553) = 158$, $p < .001$. Drivers had longer reaction time and lower accuracy in the 2-back task than the 1-back task, which indicated the success of our manipulation.

2.3.2.2 Takeover timeliness

Table 2.7 and 2.8 show the mean and standard error of driving behaviors.

Table 2.8: Mean and standard error of driving behaviors under ∼7s TOR lead time ($Mean \pm SE$)

	∼7s TOR lead time			
	No traffic		Heavy oncoming traffic density	
	Low cognitive load	High cognitive load	Low cognitive load	High cognitive load
Takeover RT (s)	2.33 ± 0.15	2.43 ± 0.14	2.11 ± 0.15	2.28 ± 0.14
Eyes-on-road RT (s)	1.16 ± 0.10	1.23 ± 0.10	1.00 ± 0.10	1.30 ± 0.09
Max resulting acceleration (m/s^2)	2.44 ± 0.25	2.46 ± 0.25	2.76 ± 0.28	2.39 ± 0.24
Max resulting jerk (m/s^3)	28.8 ± 7.1	28.0 ± 7.0	36.5 ± 7.7	26.4 ± 5.7
Minimum TTC (s)	3.34 ± 0.24	3.2 ± 0.25	2.68 ± 0.25	2.72 ± 0.23
Road deviation SE (cm)	0.78 ± 0.06	0.72 ± 0.06	0.83 ± 0.09	0.79 ± 0.05

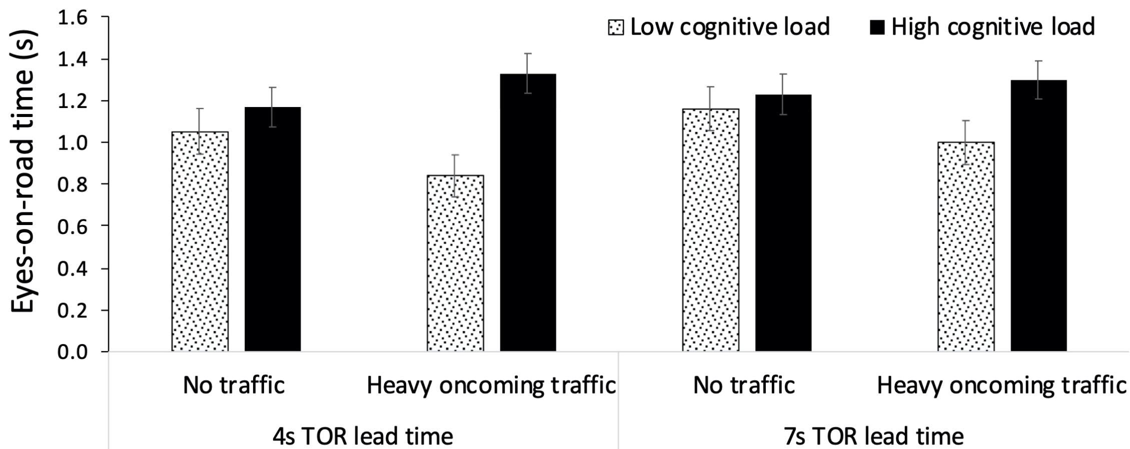


Figure 2.12: Eyes-on-road reaction time (s). Error bar indicates one standard error

There were no significant effect of cognitive load, $F(1,554) = .44$, $p = .51$; traffic density, $F(1,554) = 2.28, p = .13$; and TOR lead time, $F(1,555) = .30, p = .59$, on takeover time. None of their interaction effects were significant.

For eyes-on-road reaction time, there was a significant effect of cognitive load, $F(1,534) = 21.00$, $p < .001$. As shown in Figure 2.12, lower cognitive load led to shorter eyes-on-road reaction time. Moreover, there was a significant interaction effect between cognitive load and traffic density, $F(1,534) = 7.11, p = .01$. Simple effect analysis showed that lower cognitive load led to a shorter eyes-on-road time in heavy oncoming traffic condition, $p < .001$, while cognitive load did not influence eyes-on-road time when there was no oncoming traffic, $p = .18$.

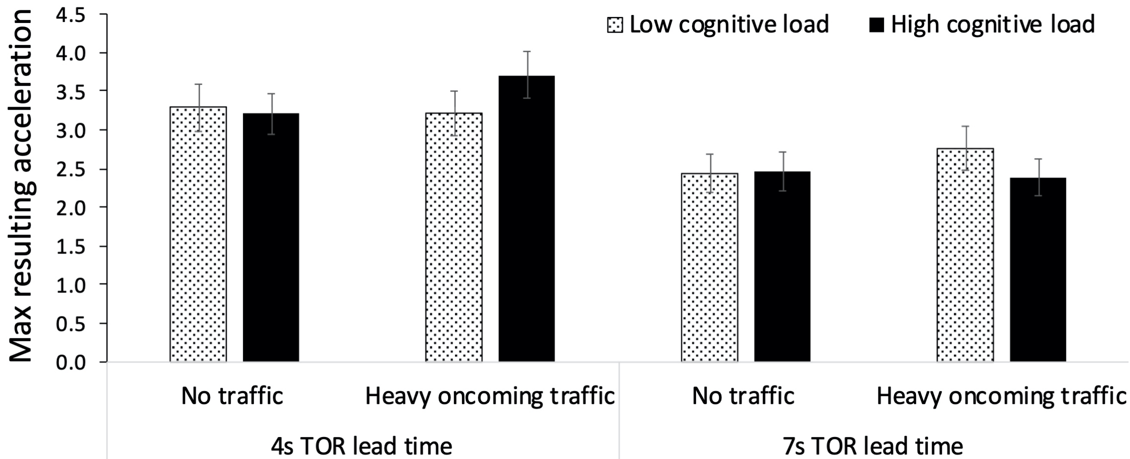


Figure 2.13: Maximum resulting acceleration (m/s^2). Error bar indicates one standard error

2.3.2.3 Takeover quality

The ∼4s TOR lead time resulted in a larger maximum resulting acceleration, $F(1, 555) = 23.23, p < .001$, and a larger maximum resulting jerk, $F(1, 555) = 5.64, p = .02$ (Figure 2.13). No other significant effects were found on maximum resulting acceleration, maximum resulting jerk, or standard error of road deviation.

In the four takeover events that required drivers to change lanes, drivers had shorter minimum TTC when oncoming traffic density was heavy, $F(1, 305) = 5.77, p = .02$, and when the TOR was ∼4s, $F(1, 305) = 115.32, p < .001$ (Figure 2.14).

2.3.2.4 Psychophysiological responses during NDRTs

During NDRT, there was a significant main effect of cognitive load on heart rate variability ($F(1, 586) = 5.17, p = .023$). Drivers had lower heart rate variability when they were in the condition of high cognitive load (Figure 2.15). All other main effects and interaction effects on heart rate variability were not significant, so they were not included in Figure 2.15.

As shown in Figure 2.16, drivers had lower horizontal gaze dispersion ($F(1, 586) = 108.75, p < .001$) and shorter eyes-on-road time ($F(1, 586) = 108.35, p < .001$) when

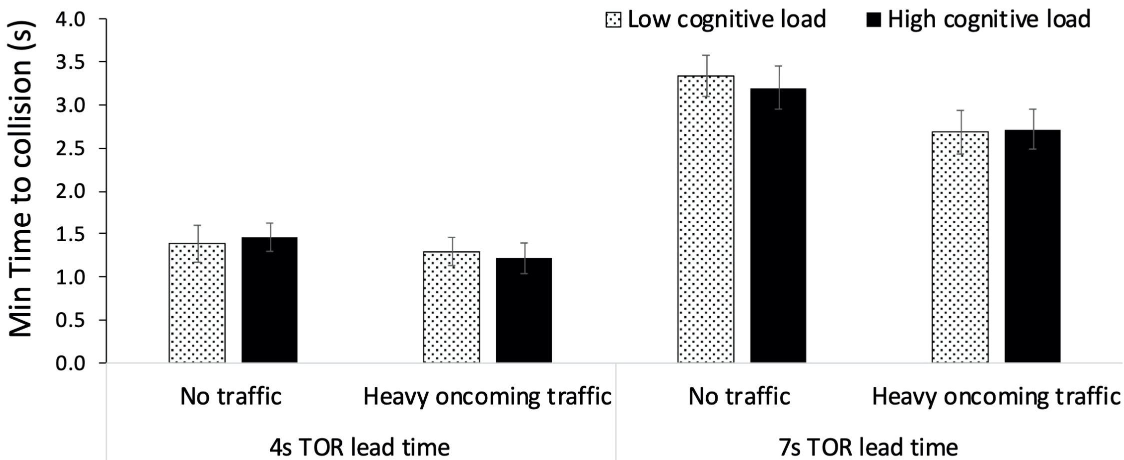


Figure 2.14: Minimum time to collision (s). Error bar indicates one standard error

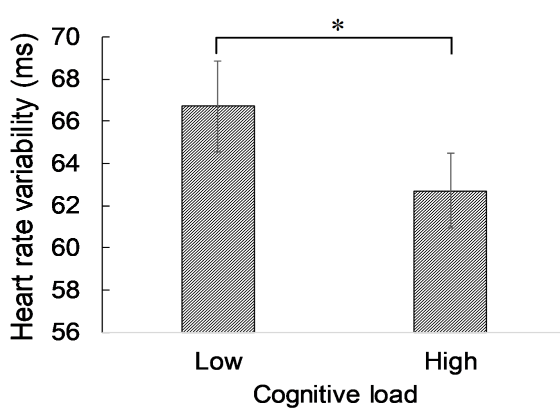


Figure 2.15: Heart rate variability (ms) during NDRTs by cognitive load.

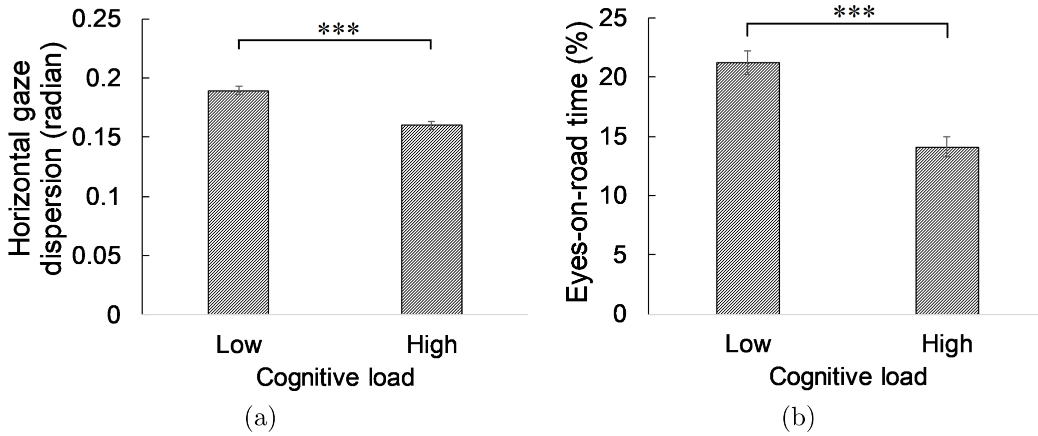


Figure 2.16: (a) Horizontal gaze dispersion (radian); (b) Eyes-on-road time percentage (%) during NDRT stage by cognitive load. TORs were issued at Time 0.

they were in high cognitive load. However, their blink number did not differ significantly between two cognitive load task conditions. The main effects of traffic density and TOR lead time and their interaction effects were not significant, and were not included in Figure 2.16.

2.3.2.5 Psychophysiological responses during takeover transitions

Only the main effect of TOR lead time on blink number was significant ($F(1, 588) = 6.11, p = .014$). We found that ∼4s TOR lead time led to fewer blink numbers than ∼7s TOR lead time in general during the takeover stage (Figure 2.17). If we analyzed the blink number second by second, as shown in Figure 2.18, we found that ∼4s TOR lead time significantly suppressed blinks at 2s, 3s, and 4s after TORs (2s: $t(90) = 2.96, p = .004$; 3s: $t(90) = 1.78, p = .05$; 4s: $t(90) = 4.51, p < .001$). Yet, no significant effects were found on the horizontal gaze dispersion.

Compared to the NDRT stage, drivers' mean phasic GSR was significantly higher in the takeover action stage ($F(1, 1275) = 44.43, p < .001$). As shown in Figure 2.19, drivers' GSR phasic activation increased after a TOR and reached a peak 5s after the alert. The main effects of TOR lead time on maximum and mean GSR phasic

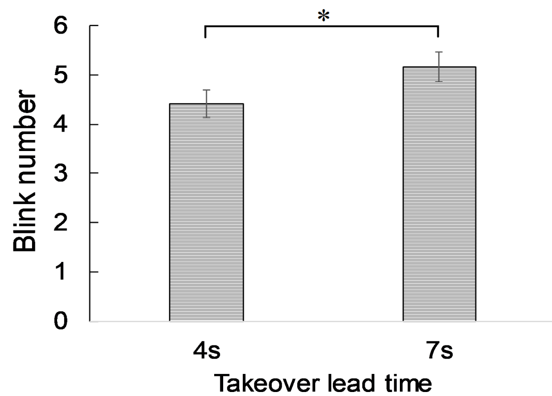


Figure 2.17: Blink number after TORs by TOR lead time.

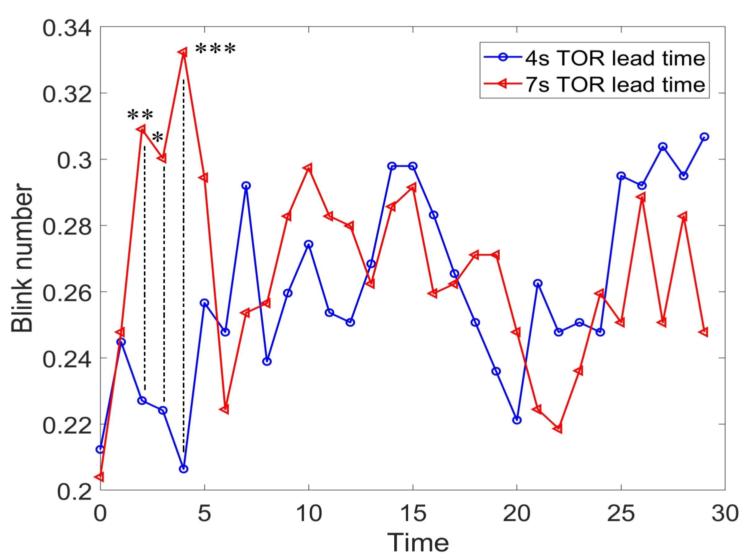


Figure 2.18: Blink number through the drives. TORs were issued at Time 0.

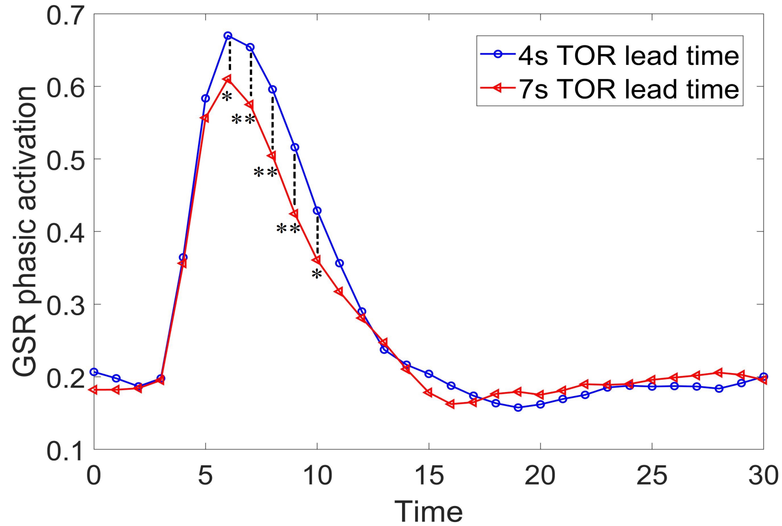


Figure 2.19: Mean GSR phasic activation (μS) through the drives. TORs were issued at Time 0.

activation were significant ($F(1, 587) = 8.80, p = .003$; $F(1, 591) = 4.92, p = .027$). Generally, ∼4s TOR lead time induced larger maximum and mean GSR phasic activation than ∼7s TOR led time during the whole takeover time window. Furthermore, we found that GSR phasic activation differences caused by TOR lead time appeared 5s after the TOR, lasted for 5s and disappeared 10s after the TOR (5s: $t(90) = 2.33, p = .022$, ; 6s: $t(90) = 2.87, p = .005$; 7s: $t(90) = 3.20, p = .002$; 8s: $t(90) = 3.14, p = .002$; 9s: $t(90) = 2.43, p = .017$). No other significant effects were found on the mean or maximum GSR phasic activation.

The main effects of cognitive load, traffic density, TOR lead time, and their interaction effects on exact values of heart rate changes (heart rate in the takeover stage minus NDRT stage) were not significant. As noted in the introduction, heart rate differences were then categorized into three patterns. Figure 2.20 shows the numbers of the three heart rate response patterns under different traffic density, TOR lead time and cognitive load conditions. Primarily, heart rate acceleration happened the most frequently when drivers switched from NDRTs to takeovers, followed by no changes, and heart rate deceleration. There was a significant main effect of traffic

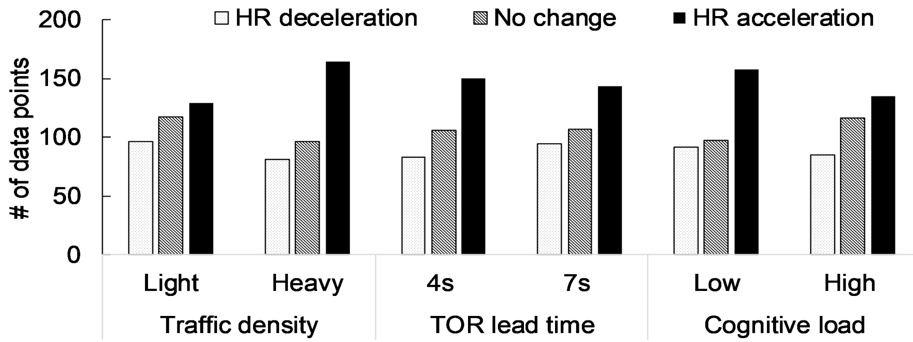


Figure 2.20: The number of takeover scenarios by independent variables and HR response pattern.

Table 2.9: Mean heart rate (and standard error) by traffic density group and HR response pattern.

Stage	Light traffic density			Heavy traffic density		
	HR deceleration (n = 96)	No changes (n = 117)	HR acceleration (n = 129)	HR deceleration (n = 81)	No changes (n = 96)	HR acceleration (n = 164)
NDRT	92.1 ± 3.1	80.3 ± 1.3	81.2 ± 1.8	90.0 ± 2.4	80.8 ± 2.0	81.6 ± 1.6
Takeover	85.2 ± 2.6	80.5 ± 1.3	91.1 ± 2.4	83.0 ± 2.0	80.9 ± 1.9	88.6 ± 1.8

density on heart rate response patterns ($\chi_2^2 = 7.54, p = .023$). In comparisons to light traffic density, significantly more heart rate acceleration patterns were found in the heavy traffic density condition (Table 2.9). As shown in Figure 2.21, such differences appeared at the 12th second after TORs and lasted until about the 27th second. Yet, the main effects of TOR lead time and cognitive load on heart rate response patterns were not significant.

2.3.2.6 Correlation matrix

The correlation matrix, shown in Table 2.10, indicates the relationships between drivers' physiological data and emotions in valence and engagement dimensions after TORs. We found that maximum and mean GSR phasic activation were negatively correlated with drivers' emotional valence, while blink number was positively correlated with drivers' emotional valence. In other words, the more negative emotions drivers had, the larger maximum and mean GSR phasic activation and lower blink

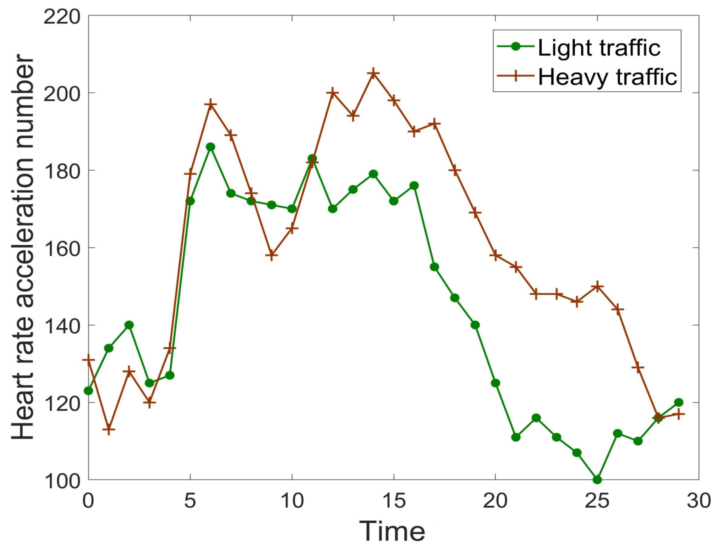


Figure 2.21: The number of heart rate acceleration patterns after TORs. TORs were issued at Time 0.

Table 2.10: Correlation matrix between drivers' physiological data, emotions, and subjective takeover performance.

	Horizontal gaze dispersion	Blink num	HR differences	Max GSR phasic	Mean GSR phasic
Valence	-.042	.123**	-.002	-.158**	-.107**
Engagement	-.017	.051	.09*	-.042	-.069

numbers they had after TORs. Meanwhile, drivers' engagement was significantly positively correlated with HR differences between takeover and NDRT stage.

2.3.3 Discussion

This experiment examined the effects of NDRTs, traffic density, and TOR lead time on drivers' driving behavioral and psychophysiological responses to TORs in simulated SAE Level 3 automated driving. We discuss how the three factors influenced different aspects of takeover transitions.

2.3.3.1 Driving behavioral responses

We manipulated drivers' cognitive load using an N-back memory task, and found that with a lower cognitive load, drivers had shorter eyes-on-road reaction time. We speculate that cognitive load influenced takeover actions by changing the monitoring behavior at the NDRT stage. When the cognitive load was low, drivers had the bandwidth to monitor the driving environment, stay prepared, and therefore responded quickly once the TOR was issued. More gaze behaviors could be analyzed in the future to validate the explanations.

Prior research on traffic density indicated that heavy traffic reduced takeover performance (*Gold et al.*, 2016, 2018a; *Radlmayr et al.*, 2014; *Körber et al.*, 2016; *Eriksson et al.*, 2018). However, in these studies, traffic density was confounded with the availability of escape paths. In the present study, all the other vehicles were in the oncoming traffic. This design allowed us to decouple traffic density and the availability of escape paths. We found that oncoming traffic density did not influence either takeover reaction time or maximum resulting acceleration/jerk. This result implies that heavy traffic density per se did not lead to increased use of braking or steering, but the available escape paths did. Nevertheless, as represented by minimum TTC, heavy oncoming traffic density did increase drivers' risk of collision through longer decision-making time.

Interestingly, we found an interaction effect between oncoming traffic density and cognitive load. Lower cognitive load shortened eyes-on-road reaction time only during heavy oncoming traffic. This suggests that participants adjusted their monitoring behavior according to the complexity of driving environments. When oncoming traffic density was heavy, low cognitive load allowed drivers to allocate their attention to the road in order to increase situational awareness and be prepared for potential events.

There were no effects of TOR lead time on takeover reaction time or eyes-on-road time. Generally, longer TOR lead time leads to longer takeover reaction times, but takeover reaction time is also influenced by other variables, such as scenario emergency (*McDonald et al.*, 2019b). A possible reason for the lack of significance here is that all the scenarios looked urgent, so drivers reacted reflectively and equally quickly, regardless of TOR lead time.

When TOR lead time was ∼7s, participants had smaller maximum resulting acceleration and jerk, and larger minimum time to collision. Maximum resulting acceleration and jerk have been utilized to identify shift safety, ride comfort (*Huang and Wang*, 2004) and driving aggressiveness (*Bagdadi and Várhelyi*, 2011; *Feng et al.*, 2017). The results indicate that ∼7s TOR lead time led to safer and and more comfortable takeover behavior and lower collision risk, which supports previous studies (*Gold et al.*, 2013; *Wan and Wu*, 2018; *Gold et al.*, 2016; *Körber et al.*, 2016).

2.3.3.2 Psychophysiological measures during NDRTs

During the NDRT stage with automated driving mode on, drivers were assigned N-back tasks on the tablet. Our results showed that drivers had lower heart rate variability when they were in the 2-back memory task than the 1-back memory task. Heart rate variability is a sensitive indicator of cognitive load (*Mehler et al.*, 2012; *Lei and Roetting*, 2011). Our findings aligned with previous research (*Bashiri and D Mann*, 2014; *Mehler et al.*, 2011), and implied drivers' high cognitive load in the 2-back memory task.

Meanwhile, we found that drivers had narrower horizontal gaze dispersion and spent less time monitoring the road when they were in the 2-back memory task. This can be explained from two aspects. First, the 2-back memory task required drivers to memorize more chucks and required more cognitive resources. Consistent with previous studies (*Wang et al.*, 2014; *Gold et al.*, 2016), narrower horizontal gaze dispersion indicated drivers' increased cognitive load in the 2-back memory task. Second, while more attentional resources were occupied by the 2-back memory task, drivers had fewer opportunities to monitor the driving environment. Their narrower horizontal gaze dispersion and less eyes-on-road time suggested reduced situational awareness of the driving environment (*Molnar*, 2017).

2.3.3.3 Psychophysiological measures during takeover transitions

Upon the TOR, drivers were required to terminate NDRTs, check the driving environment, and negotiate takeover scenarios appropriately. During this process, we found that drivers had lower blink numbers when TOR lead time was ∼4s. The number of blinks decreases when there is more information to be processed in a short time (*Veltman and Gaillard*, 1996). Thus, blink inhibition in ∼4s TOR lead time indicated that drivers paid greater attention to scenarios and utilized more efforts to support decision making and respond to urgent events. Meanwhile, we found that blink number was positively correlated with drivers' emotional valence detected by facial expressions. This suggested that the more blink suppression drivers had, the more negative emotions (e.g., stress) they had in the face of TORs (*Haak et al.*, 2009). However, we did not find significant differences in blink number in two different cognitive load conditions. This was probably because blink number was more sensitive to temporal demands (*Veltman and Gaillard*, 1996) than to cognitive demands. Meanwhile, we found a significantly negative correlation between drivers' subjective ratings of takeover performance and horizontal gaze dispersion. It is likely that drivers required wider horizontal gaze dispersion to process the driving information and negotiate takeover events in a worse takeover performance situation.

Regarding GSRs, drivers' phasic components increased significantly in response to TORs, which implies high emotional arousal in the face of unexpected events (*Boucsein*, 2012). In general, compared to ∼7s TOR lead time, drivers had larger maximum and mean GSR phasic activation in the ∼4s TOR lead time condition, indicating higher arousal when situations were more critical. However, a high arousal level could be associated with both positive and negative emotions. Therefore, we further investigated its correlation with drivers' emotional valence. We found that maximum and mean GSR phasic activation were negatively correlated with drivers' emotional valence. In other words, the higher arousal the drivers had in response to TORs, the more negative their emotions were. Following the previous studies (*Wandtner et al.*, 2018; *Morris et al.*, 2017; *Healey and Picard*, 2005), we interpreted that drivers experienced greater stress in the ∼4s TOR lead time condition, as indicated by the GSR phasic component and emotional valence.

As described in the results section, there were different patterns of drivers' average heart rate differences from NDRTs to the takeover stage. In general, heart rate acceleration happened the most frequently, which was associated with stimulus ignorance and environmental rejection (*Sokolov*, 1963; *Lacey*, 1967, 1970). Such an attentional pattern matched the takeover mechanism, as drivers were required to terminate or ignore their NDRTs for takeover actions at the moment of TOR. More interestingly, drivers showed more heart rate acceleration patterns in heavy traffic density. This meant that drivers selectively rejected and blocked out the overwhelming traffic information in attention-demanding situations. Even though we did not find any performance-level differences induced by traffic density (*Du et al.*, 2020c), heart rate measures explained drivers' attentional styles and revealed potential safety concerns with heavy traffic density during takeover transitions. Also, there was a significant positive correlation between drivers' engagement and heart rate changes from NDRT to takeover stages. The more heart rate acceleration drivers had, the more engaged they were in the takeover transitions, indicating that drivers were engaged in takeover actions while ignoring unnecessary traffic information in complex situations.

2.3.3.4 Time-series psychophysiological measures

The second-by-second analysis of psychophysiological measures allow us to understand drivers' responses to TORs in a continuous way. Using time-series data, we found that drivers' blink suppression happened 2s after TORs and lasted for 3 seconds. The onset of the significant differences at the 2nd second tended to be consistent with drivers' reaction time (average reaction time = 2.3 s in this study) (*Eriksson and Stanton*, 2017; *McDonald et al.*, 2019b). Once drivers started to take over control of the vehicle, their blinks were suppressed to extract the most important visual information and remove distracting information in the driving environment (*Bidder II and Tomlinson*, 1997). Yet, compatible with the characteristics of gaze behaviors in previous studies (*Alrefaie et al.*, 2019; *Kramer et al.*, 2013), such gaze reactions to TORs were rapid and could recover immediately when the complex driving information was processed.

With regard to GSR phasic activation, we found that drivers' phasic differences triggered by different lead times became significant 5s after the TOR, but only lasted for another 5 seconds and then became monotonous. This was likely because drivers perceived the event urgency differently at the time of TOR, but got used to it after they gradually negotiated takeover scenarios. This phenomenon was also consistent with the latency of GSRs responding to unexpected events, the rise time to the peak from the baseline, and the fall time returned to the baseline from the peak after unexpected events were resolved (*Boucsein*, 2012).

However, compared to other metrics, heart rate seemed to have a long latency before changes induced by TORs and such changes lasted for a long time, as shown in Figure 2.21. This is consistent with previous studies, as heart rate activities change gradually and required a longer time window to be stable (*Solovey et al.*, 2014; *Alrefaie et al.*, 2019).

In summary, drivers' psychophysiological response patterns in the time domain were rather different to the same TORs. Some responded immediately and recovered soon while others had long latency periods for responses and lasted for a long time. When we used the whole takeover transition period as the time window to calculate various measures for statistical analysis, it gave us an overview of drivers' states during takeover transitions. In contrast, analyzing the second-by-second time-series data gave us insights into their temporal changes and the optimal time window selection to improve the sensitivity and specificity of different psychophysiological measures.

2.3.3.5 Implications

Our findings have implications for the design of in-vehicle alert systems. While different levels of cognitive load, traffic density, and TOR lead time influence drivers' takeover performance differently, in-vehicle alert systems can be designed adaptively to match drivers' takeover readiness. For example, when oncoming traffic density is heavy, a gaze guidance system can be designed to highlight the most important features and support drivers in noticing potential hazards. When TOR lead time is short, a multimodal display combining auditory, visual, and tactile alerts can be issued to alarm drivers intensively and increase their alertness (*Politis et al.*, 2013). If driver monitoring systems detect that drivers are in high cognitive load induced by NDRTs, a regulatory warning can be issued to remind drivers to stop current NDRTs, keep relaxed, and prepare for potential takeover requests.

Psychophysiological measures indicated proactive responses induced by different NDRTs, traffic density and TOR lead time before performance behavior was observed. In summary, the inclusion of psychophysiological measures helped provide insights into the often unconscious mechanisms underlying the takeover performance behaviors. Therefore, such measures can help researchers understand the mechanisms of takeover transitions by complementing other vehicle-related measures and improve predictions of takeover performance proactively.

The reliable and valid assessment of drivers' internal states using psychophysiological measures can be the ground work to develop state detection and monitoring systems. Studies have shown that there are medium to strong associations between psychophysiological measures and drivers' states (*Du et al.*, 2020b; *Zhou et al.*, 2020a). Data from wearable devices can be used to train advanced machine learning models to indicate drivers' states in a continuous, non-obtrusive, proactive, and real-time way. Furthermore, according to monitoring results, an adaptive in-vehicle alert system can be designed to trigger warnings or intervene drivers when sub-optimal internal states are associated with potential hazards during the takeover transition period.

2.3.3.6 Limitations and future work

First, we used a high-fidelity fixed-base driving simulator to imitate takeover situations in a controlled laboratory and recruited younger adults as participants. Considering the experiment goals and time limitation, eight takeover scenarios were designed within about a 40-minute drive course. This is especially important when the psychophysiological measures collected are sensitive to various factors. However, the obtained results might be less ecologically valid than those obtained from on-road scenarios and across age groups. Future studies can replicate the experimental settings with naturalistic driving and recruit diverse drivers to see the robustness of psychophysiological measures. For example, drivers' individual demographic information and personalities may affect takeover performance and psychophysiological responses. Future research can also extend experiment duration to see the potential findings caused by drivers' drowsiness or boredom in automated driving mode.

Second, to interpret the psychophysiological data, we compared our results with well-established literature and provided insights on drivers' cognitive load, attention, and emotion states, as reflected by psychophysiological data throughout the takeover transitions. Correlation analysis between drivers' dimensional emotions and physiological data was also conducted to increase the validity and interpretability of results. Future study can collect more self-reported measures on internal states (e.g., situational awareness) to help interpret the results. It would also be valuable to examine the relationship between psychophysiological data and driving behaviors (e.g., minimum time to collision) to see whether psychophysiological data can be used to predict objective takeover performance.

Third, given the fact that drivers' internal states are associated with multiple psychophysiological measures, we used several of them to reliably measure subtle changes in drivers' cognitive load, attention, emotional states, and situational awareness. However, a variety of psychophysiological measures can be derived from the raw physiological signals. For example, in addition to emotional valence and engagement, emotional arousal can also be predicted from facial expressions using machine learning algorithms (*Zhou et al.*, 2020b). In future studies, more metrics, such as emotional arousal, frequency-domain HRV, and fixations can be potentially included.

2.3.4 Conclusion

This study systematically investigated the effects of drivers' cognitive load, TOR lead time, and traffic density on their driving behavioral and psychophysiological responses to takeover requests. The results have important implications for the design of in-vehicle alert systems to monitor driver behaviors, improve takeover readiness and optimize takeover performance. The findings will enhance the interaction between drivers and conditionally automated vehicles.

Generally, the results showed that drivers had worse takeover performance when they had a high cognitive load, short TOR lead time, and heavy oncoming traffic density. Interestingly, if drivers had low cognitive load, they paid more attention to driving environments and responded more quickly to takeover requests in high oncoming traffic density conditions. During the automated driving stage, we found that drivers had lower heart rate variability, narrower horizontal gaze dispersion, and shorter eyes-on-road time when they were in high cognitive load triggered by the 2-back memory task. Upon the TOR, $\sim$4s lead time led to inhibited blink numbers and larger maximum and mean GSR phasic activation, indicating higher emotional arousal and stress than $\sim$7s lead time. Meanwhile, heavy traffic density resulted in significantly frequent HR acceleration patterns than light traffic density, suggesting ignorance of overwhelming traffic information.

While driving behaviors alone give us insights into drivers' takeover performance, psychophysiological signals collected by non-invasive sensors allow us to estimate drivers' workload, emotions, attention, and situational awareness in a continuous and real-time manner. The findings can provide us a broad picture of driver states throughout the whole takeover stage, and inform the development of driver monitoring systems and the design of in-vehicle alert systems in SAE Level 3 automated driving.

2.4 Experiment 3

Experiment 3 aimed to systematically examine the effect of takeover scenarios and vehicle speed on drivers' behavioral and physiological responses to takeovers.

2.4.1 Method

This research complied with the American Psychological Association code of ethics and was approved by the Institutional Review Board of the University of Michigan.

2.4.1.1 Participants

A total of 40 university students (average age = 22.8 years, SD = 3.9; 20 females and 20 males) with normal or corrected-to-normal vision participated in the experiment. Participants were screened for valid US driver's license status and susceptibility to simulator sickness. The study lasted about 60 minutes, and each participant was compensated with $30 upon completion of the experiment.

2.4.1.2 Apparatus and stimuli

The study was conducted in the same fixed-base driving simulator as Experiment 2. The settings of vehicle, sensors, and NDRT were the same as in Experiment 2, as introduced in Section 2.3.1.2. Participants only conducted a visual 2-back memory task as the NDRT.

2.4.1.3 Experimental design

The experiment used a mixed design, with scenario type as the between-subjects variable and vehicle speed as the within-subjects variable. Two types of scenarios were designed on the basis of realistic situations and previous literature (*Lisetti and Nasoz*, 2004; *Uhrig et al.*, 2016; *Miller and Ju*, 2014; *Koo et al.*, 2016b; *Rezvani et al.*, 2016; *Zeeb et al.*, 2016), that is, lane keeping and lane changing scenarios (See Table 2.11). Participants were randomly assigned to one of the scenario types. Within the same scenario type, each participant went through two drives (two events in each drive), with one in the low vehicle speed (i.e., 35 mph) condition and the other in the high vehicle speed (i.e., 60 mph) condition. The order of the drives was counterbalanced among participants. Within the same scenario type, the order was counterbalanced by having half of the participants drive from Events 1 to 4, and the other half from Events 4 to 1.

The AV was always in the right lane prior to the TOR. According to the range of velodyne lidar (Velodyne Lidar, California), we set the distance between the AV and obstacle/entrance of the curve as 100 meters when the TOR was issued. Generally, there were 15 oncoming vehicles per kilometer as traffic (*Gold et al.*, 2016).

Table 2.11: Descriptions of takeover events

Event	Scenario type	Scenario descriptions
Event 1	Lane keeping	Sensor error on the left curve
Event 2		Right curve with construction zone on the left
Event 3		No lane markings on the left curve
Event 4		Sensor error on the right curve
Event 1	Lane changing	Stranded vehicle ahead
Event 2		Construction zone ahead
Event 3		Construction barrier ahead
Event 4		Police vehicle on shoulder

2.4.1.4 Dependent measures

We measured participants' objective takeover performance and physiological responses after TORs. We followed the same methods to calculate dependent variables as in Experiment 2. Table 2.6 lists the dependent variables. The definitions and equations to calculate these dependent variables were detailed in Section 2.3.1.4. Takeover performance consists of takeover timeliness (TOR reaction time) and takeover quality (maximum resulting acceleration and jerk, minimum time to collision, standard error of steering angle and road offset) (*Du et al.*, 2020d; *Feng et al.*, 2017; *Engström*, 2010). Consistent with prior research (*Gold et al.*, 2016), for the lane changing scenarios, the takeover actions ended when the vehicle's center of gravity reached the boundary of the neighboring lane; for the lane keeping scenarios, the takeover action ended when the driver passed the exit point of the curve. However, participants were instructed to re-engage the vehicle as long as they thought it was able to drive on its own. Hence, the takeover action ended earlier if participants re-engaged the vehicle before they reached the end point. Five crashes happened in lane changing scenarios, four in the high vehicle speed condition and one in the low vehicle speed condition. Participants either hit the objects or changed lanes on the shoulder during the collision. Under such situations, minimum TTC was treated as "not applicable", and other driving dynamic variables were calculated within the time window between the TOR and the end of takeover action. Physiological measures included drivers' gaze behaviors, galvanic skin responses, and heart rate (*Merat et al.*, 2012; *Gold et al.*, 2016; *Wintersberger et al.*, 2018; *Reimer et al.*, 2011). We calculated physiological measurements in the takeover stage (See Figure 2.10).

2.4.1.5 Experimental procedure

Upon arrival, participants signed an informed consent and filled out a demographic form. Experimenters attached two GSR electrodes to participants' left feet and the PPG probe to their left ear lobes. After sensor location adjustment and eye-tracking calibration, participants received a 5-minute training session, where they practiced how to keep lanes, change lanes and engage the automated driving mode by pressing a button on the steering wheel. They were asked to comply with all the traffic laws when they drove manually. They were informed that the speed limit was 35 mph in the urban and rural areas and 60 mph on the highway. Next, they started the NDRT and encountered an unexpected takeover event. The takeover event was the scenario where the traffic lights at the intersection did not work and required them to observe the surroundings and drive manually. Participants were told to re-engage the AV once they thought they had negotiated the situation.

After the training session, participants completed two experimental drives. As shown in Figure 2.22, each drive began with the command to activate the automated driving mode. Then there was an NDRT phase where participants were asked to do a visual 2-back memory task. The participants were informed that there was no need to monitor the environment when the AV was in the automated driving mode. Once a TOR was issued, participants were required to take over control of the vehicle immediately. They could hand back control to the AV after they negotiated the driving situation.

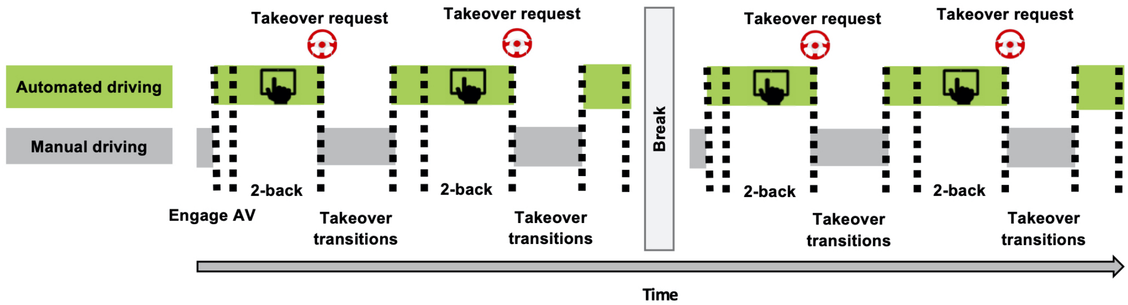


Figure 2.22: Sequence of takeover events in the experiment

2.4.1.6 Data analysis

Each participant experienced 4 events, resulting in 160 takeover events. Due to simulator and sensor malfunction, we excluded 9 events from driving behavior analysis and 15 events from physiological data analysis. We used the linear mixed model to analyze the relationship between the vehicle speed, scenario type, and the dependent variables. The scenario type, vehicle speed, and their two-way interactions were set as fixed effects. We used random intercept (participants had their own intercepts) but not random slope (participants had their own slopes) in the model development. A chi-square test was utilized to see the dependence between independent variables and heart rate change patterns. α was set at .05 for results to be reported as significant.

2.4.2 Results

Table 2.12 summarizes the mean and standard error of objective takeover performance and physiological measurements.

2.4.2.1 Takeover timeliness

As shown in Figure 2.23, drivers reacted to takeover events sooner in lane changing scenarios ($F(1, 38) = 14.79, p < .001$). The interaction effect between scenario type and speed ($F(1, 111) = 5.08, p = .026$) was significant. A simple effect analysis showed that, in lane keeping scenarios, drivers had shorter TOR reaction time in low

Table 2.12: Mean and standard error values of dependent measures

	Lane keeping scenarios		Lane changing scenarios	
	Low speed	High speed	Low speed	High speed
TOR reaction time (s)	2.87 ± 0.21	3.24 ± 0.28	2.05 ± 0.09	1.78 ± 0.08
Max resulting acc (m/s^2)	3.79 ± 0.27	2.70 ± 0.18	3.44 ± 0.53	6.56 ± 0.53
Max resulting jerk (m/s^3)	13.3 ± 3.9	10.2 ± 2.3	76.2 ± 17.4	97.3 ± 17.3
SE of steering angle (degree)	0.56 ± 0.05	0.11 ± 0.01	0.33 ± 0.04	1.03 ± 0.13
SE of road offset (centimeter)	0.89 ± 0.08	1.10 ± 0.16	2.41 ± 0.09	2.63 ± 0.12
Minimum time to collision (s)	NA	NA	1.63 ± 0.15	0.41 ± 0.10
Blink number	4.00 ± 0.60	5.26 ± 0.86	4.47 ± 0.78	2.82 ± 0.46
Horizontal gaze dispersion (radian)	0.19 ± 0.01	0.15 ± 0.01	0.22 ± 0.01	0.21 ± 0.01
Mean phasic GSR (μS)	0.31 ± 0.06	0.21 ± 0.03	0.29 ± 0.05	0.43 ± 0.07
Max phasic GSR (μS)	0.80 ± 0.15	0.77 ± 0.13	0.75 ± 0.13	0.89 ± 0.15

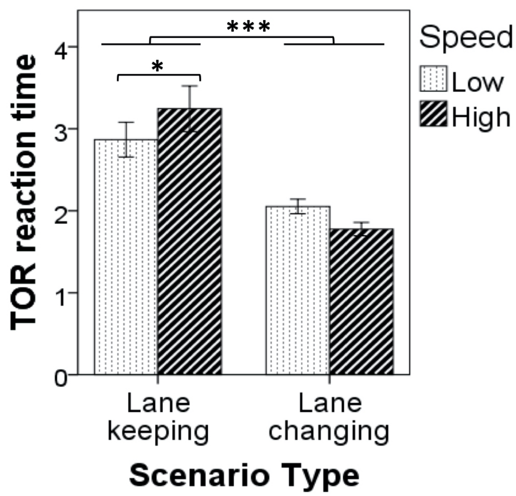
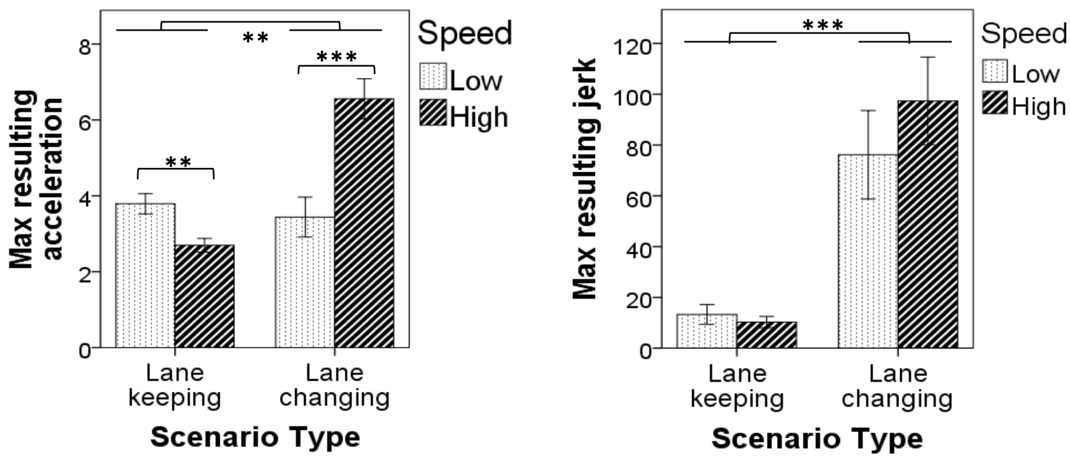


Figure 2.23: TOR reaction time (s)

speed conditions than high speed conditions (p = .046). No other significant effects were found.

2.4.2.2 Takeover quality

There were significant main effects of scenario type ($F(1, 38) = 8.58, p = .006$) and speed ($F(1, 110) = 15.49, p < .001$) on maximum resulting acceleration. Figure 2.24a shows that lane keeping scenarios and low speed both led to a smaller maximum resulting acceleration. In addition, the interaction effect between scenario type and speed (F(1, 110)=61.6, $p < .001$) was significant. Drivers had larger



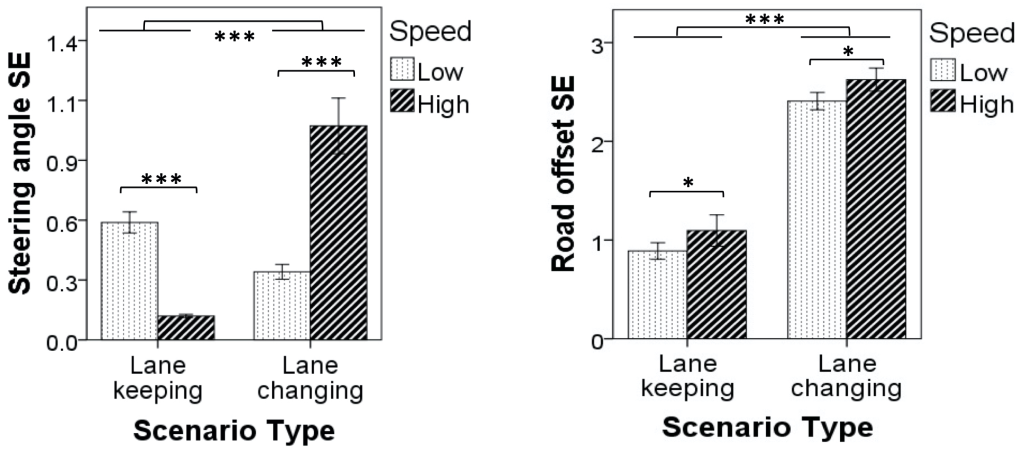
(a) (b)

Figure 2.24: Maximum resulting acceleration (m/s^2) and maximum resulting jerk (m/s^3)

maximum resulting acceleration in low speed conditions when scenarios were lane keeping (p = .006) but smaller maximum resulting acceleration in low speed conditions when scenarios were lane changing ($p < .001$). Regarding maximum resulting jerk, as shown in Figure 2.24b, only the main effect of scenario type was significant ($F(1, 38) = 16.14, p < .001$). Lane changing scenarios led to larger maximum resulting jerk compared to lane keeping scenarios.

As shown in Figure 2.25a, the SE of the steering angle in lane changing scenarios was larger than in lane keeping scenarios ($F(1, 32) = 22.49, p < .001$). The main effect of speed was not significant. Meanwhile, there was a significant interaction effect between scenario type and speed ($F(1, 108) = 65.37, p < .001$). Drivers had a larger SE of the steering angle in low speed conditions when scenarios were lane keeping ($p < .001$), but a smaller SE of the steering angle in low speed conditions when scenarios were lane changing ($p < .001$).

The main effects of scenario type ($F(1, 37) = 78.67, p < .001$) and vehicle speed ($F(1, 110) = 4.98, p = .028$) on the SE of road offset were significant. As indicated in Figure 2.25b, drivers had a smaller SE of road offset in low speed conditions and



(a) (b)

Figure 2.25: Standard error of steering wheel angle (degree) and road offset (centimeter)

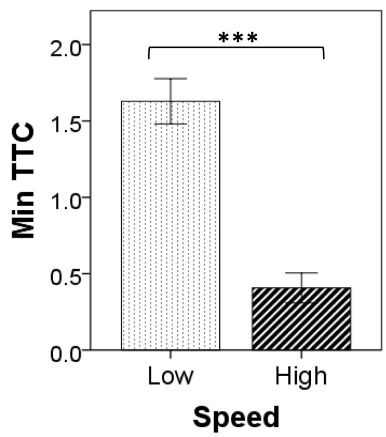


Figure 2.26: Minimum TTC (s) in different vehicle speed conditions

lane keeping scenarios. However, their interaction effect was not significant.

When the TOR was issued, the TTC was 3.73s in high speed conditions and 6.39s in low speed conditions. As shown in Figure 2.26, there was a significant main effect of speed ($F(1,54) = 62.27, p < .001$) on minimum TTC in lane changing scenarios. The minimum time to collision was shorter in high speed conditions than low speed conditions.

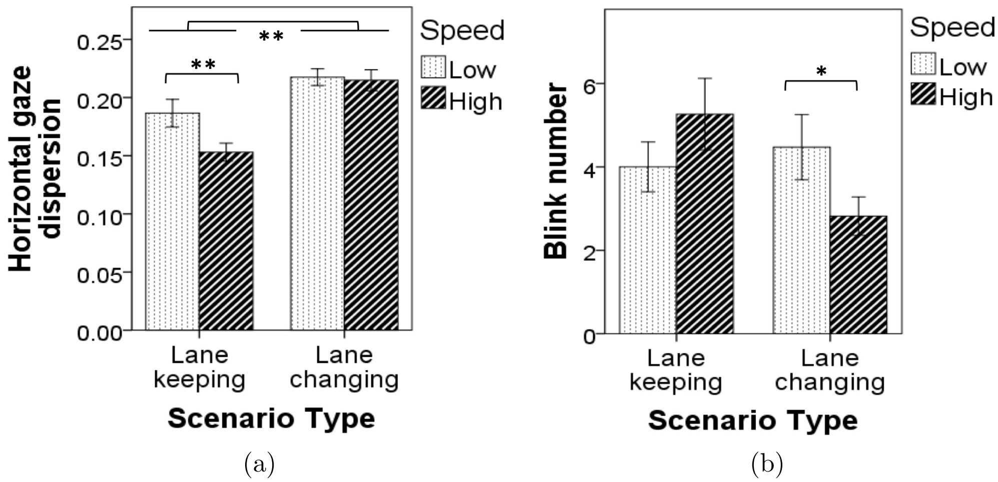


Figure 2.27: (a) Horizontal gaze dispersion (radian); (b) Blink number

2.4.2.3 Physiological responses

The main effects of scenario type ($F(1, 37) = 12.09, p = .001$) and vehicle speed ($F(1, 108) = 6.11, p = .015$) on horizontal gaze dispersion were significant. In general, drivers had wider horizontal gaze dispersion in low speed conditions than high speed conditions. Lane keeping scenarios led to narrower horizontal gaze dispersion than lane changing scenarios (Figure 2.27a). The interaction effect between scenario type and speed on horizontal gaze dispersion was also significant ($F(1, 108) = 4.88, p = .029$). Horizontal gaze dispersion was wider in low speed conditions in lane keeping scenarios ($p = .001$), but similar regardless of vehicle speed in lane changing scenarios (Figure 2.27a). Only the interaction effect between scenario type and speed on blink number was significant ($F(1, 107) = 9.37, p = .003$). High speed led to suppressed blinks compared to low speed in lane changing scenarios ($p = .014$), while speed did not influence blink number in lane keeping scenarios (Figure 2.27b).

As shown in Figure 2.28a, there was a significant interaction effect between scenario type and vehicle speed on drivers' mean GSR phasic activation ($F(1, 107) = 9.37, p = .003$). Compared to low speed conditions, drivers' mean GSR phasic activation in high speed conditions was significantly higher in lane changing scenarios

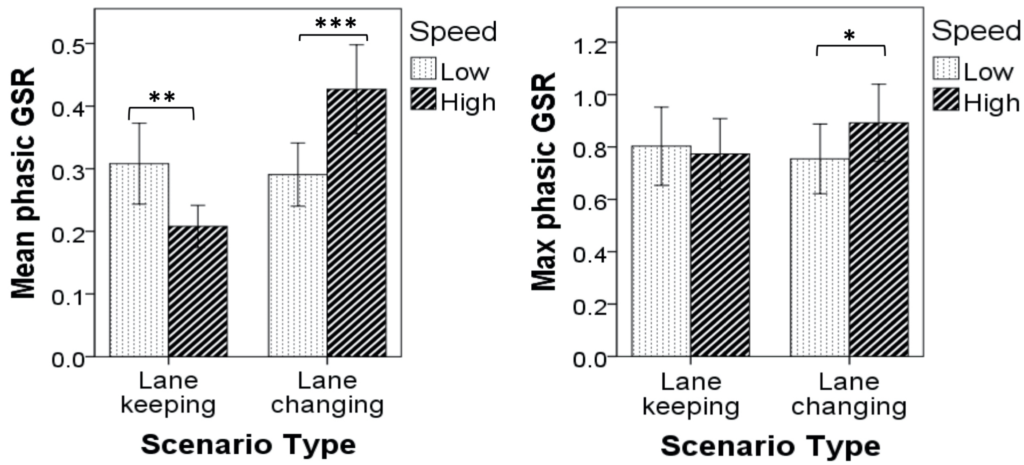


Figure 2.28: (a) Mean GSR phasic activation (μS); (b) Max GSR phasic activation (μS)

($p < .001$), but significantly lower in lane keeping scenarios ($p = .001$). The main effects of scenario type and vehicle speed on drivers' mean GSR phasic activation were both not significant. Regarding drivers' maximum GSR phasic activation, the interaction effect between scenario type and vehicle speed was marginally significant ($F(1, 105) = 3.65, p = .059$). Drivers had larger maximum GSR phasic activation in high speed conditions in lane changing scenarios ($p = .044$), but had similar maximum GSR phasic activation regardless of vehicle speed in lane keeping scenarios (Figure 2.28b).

Figure 2.29 shows the number of three heart rate change patterns before and after TORs under different scenario type and vehicle speed conditions. Heart rate response patterns were highly dependent on scenario types ($\chi^2_2 = 42.31, p < .001$). We found much more heart rate acceleration patterns in lane changing scenarios than lane keeping scenarios (Figure 2.29). Yet, there was no dependence between heart rate response patterns and vehicle speed.

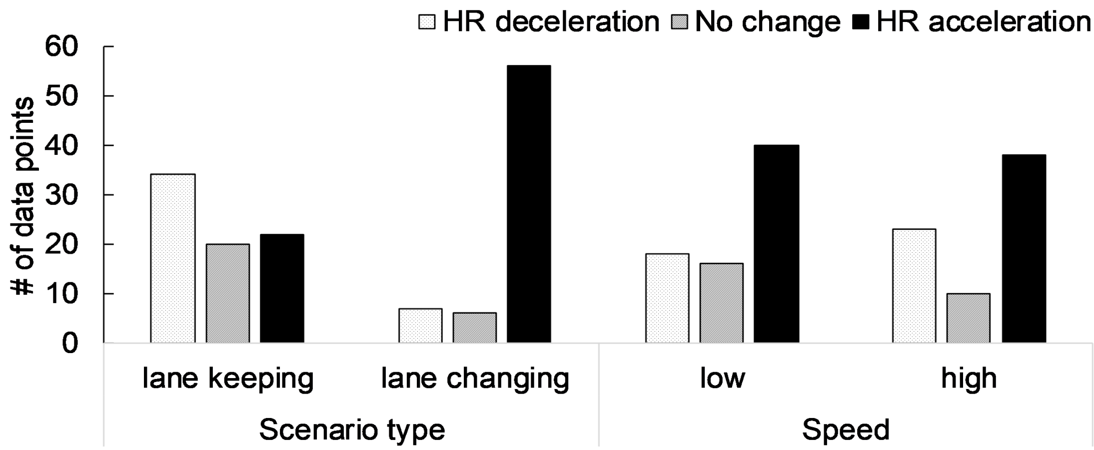


Figure 2.29: The number of data points by independent variables and HR response pattern

2.4.3 Discussion

This study investigated the effects of scenario type and vehicle speed on drivers' takeover performance and physiological responses.

2.4.3.1 TOR reaction time

In conditionally automated driving, once the driver hears the TOR request, s/he is expected to terminate the NDRT and use perceptual-motor calibration to take over control of the vehicle (*Mole et al.*, 2019). Two types of scenarios have different perceived situational criticality. In lane changing scenarios, the perception of the objects ahead triggers quick and reflexive motor behaviors, such as hands/feet back on the wheel/pedals at a moment's notice. In lane keeping scenarios, however, curves are not as visible as the objects ahead, and there are no other obvious contextual cues that request immediate motor control. Therefore, drivers' TOR reaction time was generally longer in lane keeping scenarios than lane changing scenarios. This is in accordance with the findings from *Dogan et al.* (2019), *Wu et al.* (2019) and *Radlmayr et al.* (2018).

Furthermore, we found that drivers' TOR reaction time was shorter in low speed than high speed condition for lane keeping scenarios, but similar in both speed conditions for lane changing scenarios. This can possibly be explained by the geometric design of roads in the study. When the vehicle was at a high speed (i.e., on the highway) in lane keeping scenarios, the road curves ahead were less sharp, which seemed to slow down drivers' vehicle control reactions. Yet, in lane changing scenarios, no matter which speed the AV had, the objects appeared 100 meters away when the TOR was issued, which activated a reflexive takeover reaction immediately. This aligns with a meta-analysis result that a higher urgency of situations is linked to shorter TOR reaction time (*Zhang et al.*, 2019).

2.4.3.2 Takeover quality

Drivers' specific takeover actions differed depending on the scenario types. For the lane changing scenarios, drivers were supposed to check the neighboring lanes and then rotate the steering wheel to change lanes for object avoidance. Braking was necessary to reduce the distance from objects ahead to ensure safety distance and gain more time for decision making, while acceleration was necessary to facilitate the lane changing process. Yet, in the lane keeping scenarios, drivers just needed to focus on the current lane and adjust the steering wheel to maintain lanes on the curvy road, although braking and acceleration could be applied for better adjustment and vehicle dynamics. Such strategic differences explain why the lane changing scenarios led to larger maximum resulting acceleration, maximum resulting jerk, SE of the steering angle and road offset compared with lane keeping scenarios.

With regard to vehicle speed, in general, drivers had larger maximum resulting acceleration, larger standard error of the steering wheel, and shorter minimum TTC in high speed conditions compared to low speed conditions. As the distance between the AV and ahead objects/curve entrance was fixed, high speed indicated short TOR lead time. Hence, our results were consistent with findings of previous studies that short TOR lead time was associated with shorter minimum TTC, greater maximum lateral and longitudinal acceleration (*Wan and Wu*, 2018; *Du et al.*, 2020c), and greater standard deviation of the steering wheel (*Mok et al.*, 2015).

Interestingly, there were significant interaction effects between vehicle speed and scenario type on maximum resulting acceleration and SE of the steering angle. Specifically, low speed led to larger maximum resulting acceleration and SE of the steering angle in lane keeping scenarios, but smaller maximum resulting acceleration and SE of the steering angle in lane changing scenarios. Notably, in lane keeping scenarios, low speed indicated urban and rural environments where curves had larger curvature than highway environments (*Aashto*, 2001). Thus, drivers needed to adjust the pedals and steering wheel more frequently, and used more effort to maintain lanes in low speed conditions.

2.4.3.3 Physiological responses

In lane changing scenarios, drivers not only needed to look at the forward roadway as in lane keeping scenarios, but also look around to see lane change possibilities. Such differences in scanning patterns explained why drivers had wider horizontal gaze dispersion in lane changing scenarios. High speed was associated with narrow horizontal gaze dispersion, suggesting high cognitive load for information processing and decision making and limited monitoring span in rapidly dynamic environment (*Engström et al.*, 2005; *Lemercier et al.*, 2014). This also explained why horizontal gaze dispersion was narrower in high speed conditions in lane keeping scenarios. Drivers had significant blink inhibition with high speed in lane changing scenarios, but similar blink numbers regardless of vehicle speed in lane keeping scenarios. This suggested an increased cognitive load with high speed in lane changing scenarios as criticality increased (*Merat et al.*, 2012).

Drivers' maximum and mean GSR phasic activation indicate drivers' arousal and stress in response to TORs (*Wintersberger et al.*, 2018). Consistent with a previous study (*Du et al.*, 2020a), drivers had larger mean and max GSR phasic activation with high speed (i.e., short TOR lead time) in lane changing scenarios, indicating high arousal and stress in such urgent situations. In contrast, high speed in lane keeping scenarios means a low curvature in highway. Thus, high speed in lane keeping scenarios led to lower mean GSR phasic activation, suggesting low arousal and stress as situations looked less critical.

Drivers' HR change patterns were significantly different between lane keeping scenarios and lane changing scenarios. Heart rate acceleration, suggesting defense responses that ignore environmental stimuli (*Lacey*, 1970; *Sokolov*, 1963), was observed more frequently in lane changing scenarios than lane keeping scenarios. As described before, lane changing scenarios required additional attention resources to identify traffic elements in the neighboring lanes while dealing with objects ahead in the same lane to keep safe distance. In such attention-demanding environments, defense responses suggested that drivers had to reject overwhelming traffic information to allow deeper cognitive processing, which may induce safety concerns during takeover transitions.

2.4.3.4 Implications

Given the fact that drivers' physiological data can indicate their cognitive load, emotional states, and attention allocation, physiological data during automated driving can be utilized in the future to develop driver monitoring systems and in-vehicle alert systems. For example, if the driver monitoring system indicates that drivers are in high cognitive load, the alert system can provide alerts for drivers to self-regulate their NDRTs. Specially, when the driving environment is complex (e.g., lane changing scenarios with high vehicle speed), the alert system can provide recommendations to help drivers maintain situational awareness of the driving environment and negotiate takeovers appropriately. While lane keeping scenarios lack visible contextual cues before an unforeseeable circumstance occurs, future studies can provide explanations for lane keeping scenarios at the time of TOR to enrich drivers' mental models.

2.4.3.5 Limitations and future work

There are several limitations that should be taken into consideration in the future as research opportunities. First, in order to ensure that objects were always right in front of the vehicle in lane changing scenarios, the objects were always on the right lane and available paths were always on the left. Future studies can increase lane changing directions to generalize results. Second, as a simulated driving study, this research involved no real threat to driver safety. On-road testing with fake objects in the test facility can be conducted in the future to increase the ecological validity of results. Third, only the aggregated results of physiological data were reported in this study. Future studies can look into the time-series data analysis of physiological data to gain more insights into drivers' states during dynamic takeover transitions.

2.4.4 Conclusion

Our study is critical to understanding how scenario type and vehicle speed influence takeover performance and physiological responses in conditionally automated driving. Results showed that lane changing scenarios, in general, led to shorter TOR reaction time, wider horizontal gaze dispersion, and many more HR acceleration patterns than lane keeping scenarios. The shorter TOR reaction time indicated higher criticality of scenarios, while takeover performance and physiological metrics matched corresponding takeover reactions in lane changing scenarios. With fixed distance between the AV and ahead objects/curvature entrance, high speed (i.e., short TOR lead time) generally led to worse takeover quality in the form of shorter minimum TTC, maximum resulting acceleration, and larger standard error of the steering wheel.

More interestingly, compared to low speed, high speed decreased drivers' takeover performance, and induced suppressed blinks and larger GSR phasic activation in lane changing scenarios, suggesting high cognitive load and stress in such attention-demanding and urgent situations. However, such effects were not significant or even reversed in lane keeping scenarios.

To sum up, our study systematically categorized takeover scenarios and examined the effects of scenario type and vehicle speed on drivers' takeover performance, as well as physiological responses to takeover requests. The findings of this study will add to the knowledge base on the role of different takeover scenarios in conditionally automated driving. It will help address safety concerns during takeover transitions and facilitate the adoption of automated vehicles.

CHAPTER 3

Predicting Drivers' Takeover Performance

3.1 Introduction

In response to takeover difficulty, previous studies mainly shed light on the relationships between certain factors and takeover performance (*Eriksson and Stanton*, 2017; *Gold et al.*, 2016; *Wan and Wu*, 2018; *Körber et al.*, 2016; *Helldin et al.*, 2013; *Gold et al.*, 2018a; *Radlmayr et al.*, 2014). Little effort has been made to integrate these findings into computational models that are capable of predicting drivers' takeover performance in real time. Those few exceptions (*Gold et al.*, 2018b; *Braunagel et al.*, 2017) that could predict drivers' takeover performance were developed and tested when drivers were engaged in different types of NDRTs (e.g., monitoring vs. reading), where apparent contextual cues existed to discriminate drivers' states. In daily life, even with a specific type of NDRTs such as writing an email, drivers' states can be rather different depending on the importance of the email. Also, some factors deliberately manipulated in the experiment settings, such as emotions, are not easily accessible in the real world. Although advanced wearable technology has made it convenient to collect drivers' physiological signals to reflect their cognitive and emotional states, only gaze behaviors were used in previous studies.

In the present study, therefore, we aimed to fill the research gap and to predict drivers' takeover performance. Our study contributes to the literature in three respects. First, our study aimed to predict drivers' takeover performance when they were engaged in a specific type of NDRTs with different levels of cognitive load. Second, in addition to gaze behaviors, we used drivers' heart rate indices and galvanic skin response indices to indicate their interaction with environments, which might improve prediction results. Third, our study employed a random forest model in addition to the machine learning models used in previous studies to predict takeover performance. Random forests have been proved to have great prediction performance for classification problems (*Zhou et al.*, 2020a; *McDonald et al.*, 2014). We introduce the model development procedure in detail below.

3.2 Takeover performance model performance development

3.2.1 Dataset

The data used in the development of algorithms were collected in Experiments 2 and 3, described in Chapter 2. Participants in both experiments wore the same set of physiological sensors. The similar experimental settings in both studies make it possible to combine the two datasets. At the same time, the varieties of takeover conditions in the two studies increased model generalizability. We collected drivers' physiological data, driving behaviors, and environment-related data. The physiological measures included heart rate indices, galvanic skin response indices and eye-tracking metrics. Because of malfunctions of the driving simulator and physiological sensors, data from 13 participants were excluded, and those of the other 129 participants (i.e., 828 takeover scenarios) were available for further analysis.

To develop the prediction model, we first pre-processed the raw data and then extracted 37 features and set the ground truth. Next, we used a 10-fold nested cross-validation method to tune hyper-parameters, train models, and predict test instances for model comparisons. Specifically, we resampled the training dataset and

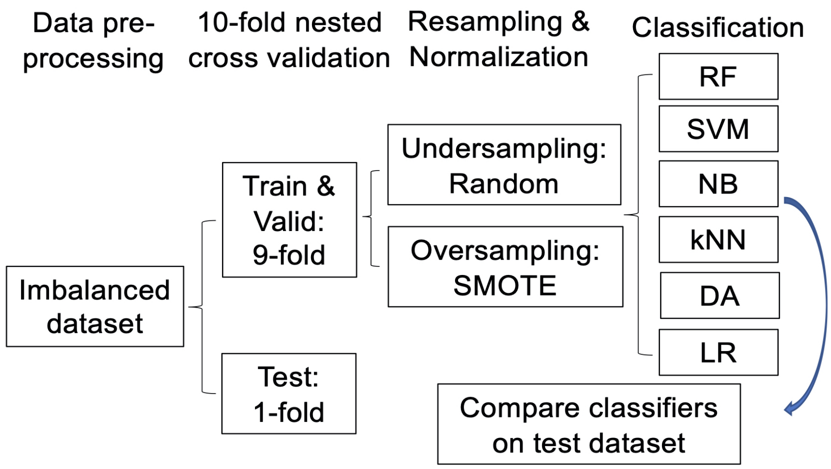


Figure 3.1: Modeling process (RF = random forest; SVM = support vector machine; NB = Naive Bayes; kNN = k-nearest neighbors; DA = discriminant analysis; LR = logistic regression).

normalized the entire dataset before performing the classification. Figure 3.1 shows the modeling process.

3.2.2 Data pre-processing

For GSR signals, we used continuous decomposition analysis (CDA) to decompose the GSR signal into phasic and tonic components, respectively, via Ledalab in Matlab (*Benedek and Kaernbach*, 2010). Then we used the phasic component for further feature extraction, because it is responsible for relatively rapid changes in response to specific events in the GSR signal (order of seconds). Heart rate measures were extracted from the raw RR interval using iMotions software. For eye-tracking data, only data points with high gaze quality value (threshold recommended by Smart Eye: .5) were recorded and used for analysis.

3.2.3 Feature generation and ground truth

To fit time series data into the supervised learning framework, we aggregated the values of physiological data within a sliding "time window" and calculated various statistics (*Anderson*, 2011). The end of the time window was the time of a TOR, and the start of the time window was X seconds before the TOR, ranging from 1 to 30 s. Model inputs included data on gaze behaviors, galvanic skin response indices, and heart rate indices, as well as environment factors. The generated features are listed in Table 3.1. A fixation is defined as "a relatively stable eye-in-head position within some threshold of dispersion (typically $\sim 2°$) over some minimum duration (typically 100-200 ms), and with a velocity below some threshold (typically 15-100° per second)" (*Jacob and Karn*, 2003). In the Smart Eye eye-tracking system, all frames with a gaze velocity below the fixation threshold (100° per second) were treated as a fixation. All frames with the gaze velocity above the saccade threshold (100° per second) were treated as a saccade. We categorized areas of interest (AOIs) into driving scenes, the NDRT tablet, and other areas. The number and average duration of fixations and saccades were accumulated within particular AOIs. The scan pattern was the probability of eyes switching from one AOI to another. Traffic density, TOR lead time, and scenario type were used to describe the driving environments because they indicated the predictability, criticality, and urgency of the takeover scenarios (*Gold et al.*, 2017). To reduce the potential impact of individual differences, we normalized the feature values across participants using the min-max normalization approach.

We used driving behaviors during takeover transitions to assess drivers' takeover performance. As shown in Table 3.2, for different takeover scenarios, we selected different metrics in the assessment. Minimum TTC was calculated only for the lane-changing scenarios, and standard deviation of road offset was calculated only for the lane-keeping scenarios. All the driving variables were calculated following prior studies (*Du et al.*, 2020d; *Clark and Feng*, 2017). If any of the calculated TOR reaction time, maximum resulting acceleration, or standard deviation of road offset values were larger than $\mu + 2\sigma$, we categorized a takeover transition as a bad performance. For minimum TTC, because the value of $\mu - 2\sigma$ was negative, we performed a log transformation first and categorized a takeover transition as bad if log (minimum

Table 3.1: Descriptions of generated features (HR = heart rate; min = minimum; max = maximum; GSR = galvanic skin responses; NDRT = non-driving-related task; TOR = takeover request).

Feature	Explanations
HR indices	Mean, min, max, and standard deviation of heart rate, inter-beat interval
GSR indices	Mean, max, and standard deviation of GSR in phasic component
GSR peak	The number of GSR peaks, and peak rise time
Fixation	Fixation number and duration in different areas of interests (AOIs) (i.e., driving scenes and NDRT tablet)
Saccade	Saccade number in different AOIs (i.e., driving scenes and NDRT tablet)
Pupil	The mean and standard deviation of pupil diameter in different AOIs (i.e., driving scenes and NDRT tablet)
Blink	The number of blinks
Gaze dispersion	Standard deviation of the values for gaze angle from right front (radians)
Eyes-on-the-road	The proportion of time that participants' gazes were on the road
Scan pattern	The probability of eyes switching from one AOI to another (i.e., the probability that drivers transited eyes from driving scenes to NDRT tablet, from NDRT tablet to driving scenes, or from other areas to driving scenes)
Traffic density	No or heavy oncoming traffic
Scenario type	Lane-keeping or lane-changing scenarios
TOR lead time	Short (3-4s) or long (6-7s) TOR lead time

Table 3.2: Takeover situations and corresponding driving behavior variables to determine takeover performance (TOR = takeover request; min = minimum; max = maximum; TTC = time to collision).

Reactions	Driving behavior variables (range for bad performance group)		
Lane changing	TOR reaction time ($> \mu + 2\sigma$)	Max resulting acceleration ($> \mu + 2\sigma$)	log(Min TTC) ($< \mu - 2\sigma$)
Lane keeping	TOR reaction time ($> \mu + 2\sigma$)	Max resulting acceleration ($> \mu + 2\sigma$)	Standard deviation of road offset ($> \mu + 2\sigma$)

TTC) was lower than $\mu - 2\sigma$ (*Braunagel et al.*, 2017). As long as one of the driving variables in a certain takeover scenario was categorized as a bad performance, we labeled the scenario as a bad takeover performance. Scenarios that led to collisions were also categorized as bad performances. Eventually, we got an imbalanced dataset with 109 "bad performance" labels and 719 "good performance" labels. The reasons that we used categorical takeover performance rather than individual driving variables as model output were that (1) it combined multiple aspects of driving behaviors and (2) it was easy to explain to drivers and more practical in guiding driver behaviors.

3.2.4 Model development

The takeover performance prediction model was trained with a random forest model considering the following justifications. First, as an ensemble method, random forests are robust for new data generalization and against training data overfitting (*Quinlan et al.*, 1996). Second, random forests can give us feature importance and makes models interpretable. Five other machine learning approaches mentioned in prior literature were applied for comparisons: k-nearest neighbors (kNN), support vector machine (SVM), Naive Bayes (NB), discriminant analysis (DA), and logistic regression (LR).

Considering the challenge of human behavior data collection, we used a 10-fold nested cross-validation method to train models and compare test results (*Lee et al.*, 2013; *Varma and Simon*, 2006). As shown in Figure 3.1, the 9-fold training and validation set ($N = 116$ subjects) was used to tune the hyper-parameters with the inner loop and then create classifiers. To handle the imbalanced dataset during the training, we employed a hybrid method of undersampling and oversampling (*Choirunnisa and Lianto*, 2018). The elimination process was done by deleting 300 good takeover performance scenarios randomly (*Prusa et al.*, 2015). Then we used the Synthetic Minority Over-sampling Technique (SMOTE) to create a balanced training and validation dataset with 678 data points (*Chawla et al.*, 2002). Table 3.3 demonstrates the training procedures of six machine learning approaches. The model assessment was based on the remaining 1-fold testing set (N = 13 subjects) with the outer loop. Notably, the subject data used for testing were not seen in the model training and validation stage. The random selection of 1-fold test dataset assumed that its distribution of good and bad takeover performance scenarios was similar to that of the whole dataset. With a 10-fold cross-validation, we could make sure all the data points in the dataset would appear once in the test dataset. The training and evaluation of the algorithm were implemented in Matlab 2018b (MathWorks, MA, USA).

3.2.5 Model evaluation

In a binary classification problem, there are four possible outcomes: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). TP is the number of positive samples predicted as a positive class, FP is the number of negative samples predicted as a positive class, FN is the number of positive class samples predicted as a negative class, and TN is the number of negative samples predicted as negative class. In this study, we used four classification evaluation indicators, namely precision, recall, accuracy, and F1-score, to carry out the evaluation of the model performance, which were defined as:

$$Precision = \frac{TP}{TP + FP} \tag{3.1}$$

Table 3.3: Machine learning techniques and training process

Machine learning approach	Techniques	Hyper-parameters
Support vector machine (SVM)	Embed the data in another dimensional space and find a soft margin that separates the classes with minimum classification error (*Chen et al.*, 2004)	Kernel, Regularization parameter
Naive Bayesian (NB)	Use maximum likelihood estimation to estimate parameters (i.e., prior probability and likelihood) (*Rish et al.*, 2001)	None
Random forest (RF)	Fit an algorithm on a set of bootstrapping samples (bagging) and predictors, i.e., randomly select training samples with replacement and take a random set of predictors at each node without replacement. Repeat many times to form an ensemble of trees (*Breiman*, 1996, 2001)	Tree number, Predictor number per split, Leaf size
k-nearest neighbor (kNN)	Calculate Euclidean distance between labeled and unlabeled points to find the k-nearest neighbors. Use the majority vote criteria to decide unlabeled points (*Keller et al.*, 1985)	k
Discriminant analysis (DA)	Find separating hyperplane using parameter estimation (*Friedman*, 1989)	Discriminant type, Regularization parameter
Logistic regression (LR)	Estimate the parameters of a logistic model (*Lee et al.*, 2006)	Regularization parameter

$$Recall = \frac{TP}{TP\ +\ FN} \tag{3.2}$$

$$Accuracy = \frac{TP\ +\ TN}{TP\ +\ FP\ + TN\ +\ FN} \tag{3.3}$$

$$F1 - score = \frac{2 \times Precision\ \times Recall}{Precision\ +\ Recall} \tag{3.4}$$

Precision manifests how well the model predicts (i.e., a measure of exactness) and recall manifests how well the model does not miss the target (i.e., a measure of completeness). The F1 measure is the weighted harmonic mean of the two, and represents a realistic measure of model performance.

The receiver operating characteristic (ROC) curve plots the true positive rate (TPR) against the false positive rate (FPR) at different thresholds (i.e., classifier boundary). The area under the curve (AUC) ranges from 0 to 1, and represents the degree of separability. A higher AUC value indicates better model performance. When AUC is 0.5, it means the model does not have any class separation capability.

3.3 Results

To improve the robustness of machine learning results, we ran the 10-fold cross-validation 30 times (i.e., 30 different random seeds) for every machine learning method at each time window. We first ran an omnibus analysis of variance (ANOVA) to compare the performance of the six machine learning methods. After that, we compared the random forest model with the other five methods using the pairwise t-test to see whether the random forest model had the best performance. Similarly, we compared the prediction results of the random forest model with different feature subsets against the full feature model using pairwise t-test. We examined the effects of time window and individual feature on random forest prediction performance using ANOVA. All post hoc comparisons used a Bonferroni α correction.

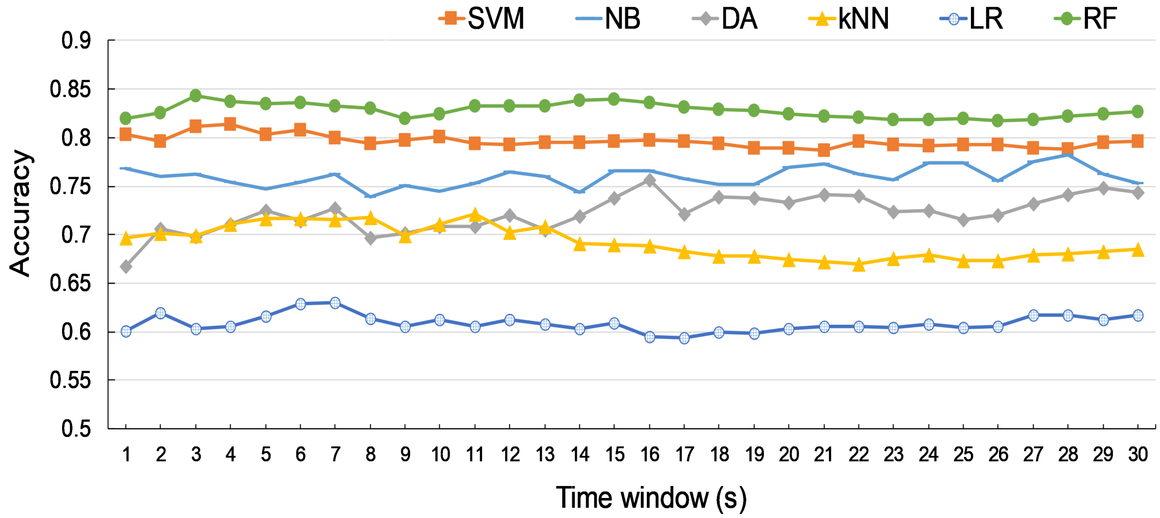


Figure 3.2: Prediction accuracy of six machine learning approaches under different time windows (SVM = support vector machine; NB = Naive Bayes; DA = discriminant analysis; kNN = k-nearest neighbors; LR = logistic regression; RF = random forest).

3.3.1 Model performance comparisons

Figures 3.2 and 3.3 show the average model accuracy and F1-score at different time windows. There was a main effect of machine learning approaches on the prediction accuracy ($F(5, 5399) = 13550, p < .001$) and F1-score ($F(5, 5399) = 4705, p < .001$). Table 3.4 shows the pairwise t-tests comparing the predictive performance of the random forest model with the other five models across different time windows. The results indicate that our proposed random forest model outperformed the other five models across time windows.

Figure 3.4 shows the ROC curves of the random forest and the other five machine learning approaches with the optimal hyper-parameters. The curve of the random forest is above and to the left of the other five curves at the majority of thresholds. Consistent with accuracy and F1-score results, the ROC curve comparisons demonstrated that the random forest model outperformed the other five models.

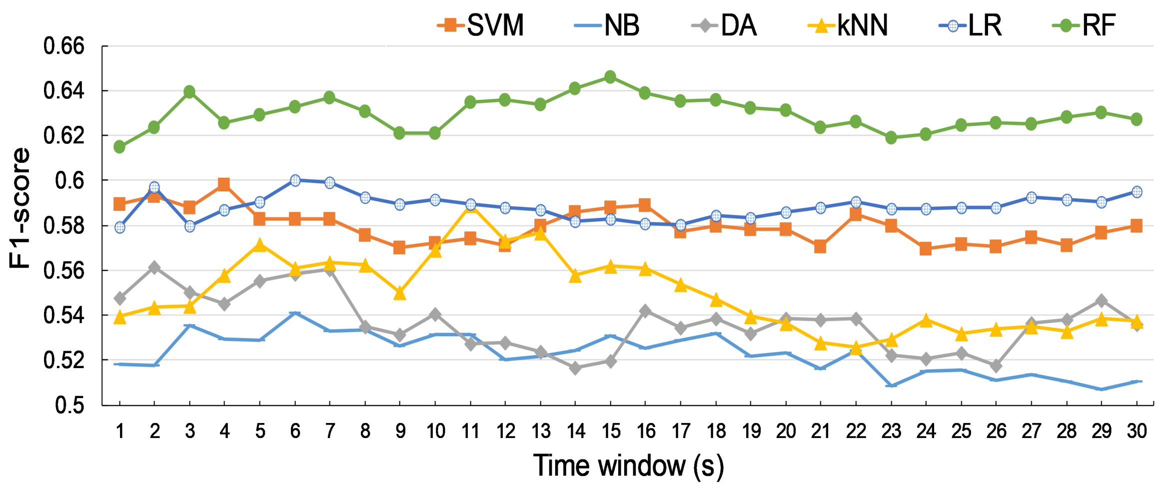


Figure 3.3: F1 scores of six machine learning approaches under different time windows (SVM = support vector machine; NB = Naive Bayes; DA = discriminant analysis; kNN = k-nearest neighbors; LR = logistic regression; RF = random forest).

Table 3.4: The mean prediction accuracy and F1-score of machine learning approaches across time windows and their comparisons to the random forest model.

Algorithm	Accuracy				F1-score			
	mean	SD	*t*-test statistic	p-value	mean	SD	*t*-test statistic	p-value
Random forest	.828	.012	-	-	.630	.015	-	-
Support vector machine	.796	.013	60.5	p<.001	.580	.019	72.4	p<.001
Naive Bayes	.760	.033	49.0	p<.001	.523	.022	107	p<.001
Discriminant analysis	.722	.021	134	p<.001	.537	.017	131	p<.001
k-nearest neighbor	.692	.020	209	p<.001	.550	.020	111	p<.001
Logistic regression	.609	.016	342	p<.001	.588	.009	74.5	p<.001

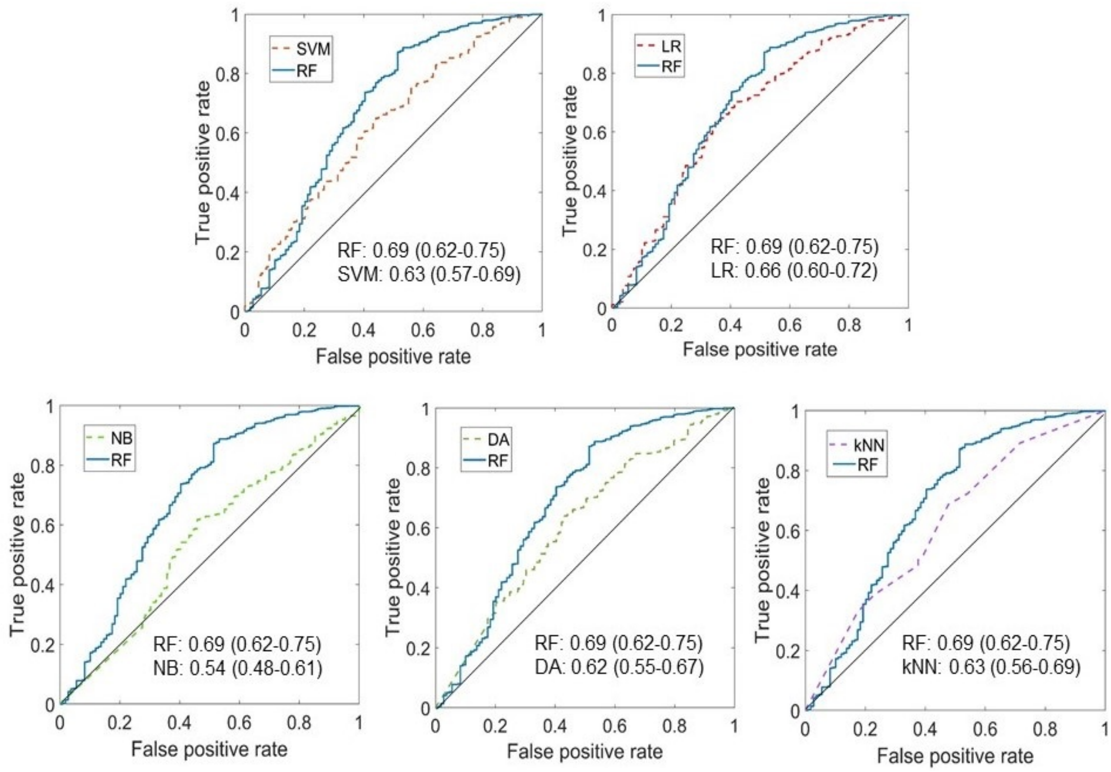


Figure 3.4: Receiver operating characteristic comparison plots for the random forest (RF) model and the five other models (SVM = support vector machine; LR = logistic regression; NB = Naive Bayes; DA = discriminant analysis; kNN = k-nearest neighbors). The bootstrapped (#1,000) confidence intervals are indicated within the parentheses.

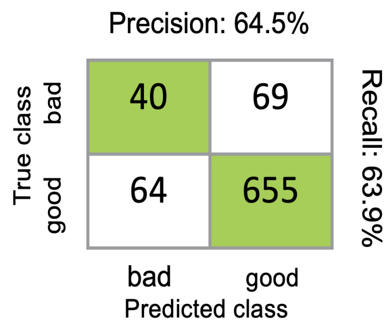


Figure 3.5: Confusion matrix when time window was 3s

3.3.2 Effects of window size on random forest prediction results

There was a main effect of time window on the random forest model accuracy ($F(29, 899) = 16, p < .001$) and F1-score ($F(29, 899) = 9, p < .001$). When applying an algorithm in real-world driving, a time window with shorter size and better prediction performance is preferred. According to Figures 6 and 7, we recommend 3 s as the optimal time window to predict takeover performance, with an average F1-score of 64.0% and accuracy of 84.3% (tuned hyper-parameters: the number of trees = 300; minimum leaf size = 2; the number of predictors per decision split = 6). Post hoc analysis showed that F1-score at the 3 s time window significantly outperformed the rest of the time windows except 5-8 s, 11-20 s, and 28-30 s (see Figure 3.3). Accuracy at the 3 s time window significantly outperformed the rest of the time windows except 4 s, 6 s, 11 s, and 13-16 s (see Figure 3.2).

3.3.3 The confusion matrix and feature importance

Figure 3.5 shows the confusion matrix when the time window was 3 s. The precision was 64.5% and the recall was 63.9%, accounting for balanced completeness and exactness of prediction.

Furthermore, by permuting the out-of-bag data (i.e., 36.8% of the total data that were not in the bootstrap samples) randomly across one predictor at a time and by measuring how much this permutation reduced the accuracy of the model,

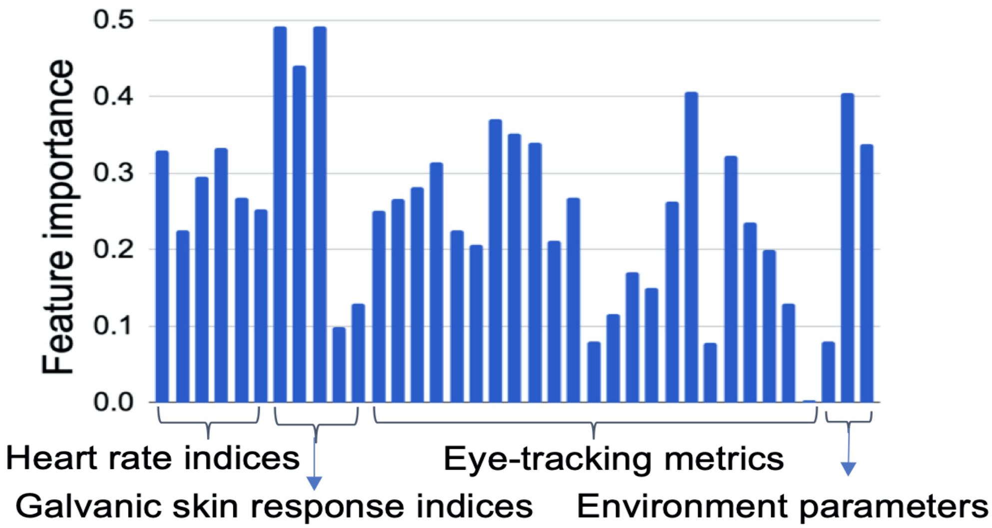


Figure 3.6: Feature importance when time window was 3s

we estimated the feature importance. The values indicate each feature's relative importance in predicting the takeover performance (the larger values are, the more important features are). Figure 3.6 illustrates the out-of-bag estimates of feature importance of the 37 predictor variables when the time window was 3 s. Table 3.5 lists the top 16 important predictor variables. As shown in the table, we found that some heart rate indices and GSR indices (e.g., maximum and mean phasic GSRs, mean of heart rate) were important in predicting takeover performance, but were not included in prior takeover performance algorithm development (*Gold et al.*, 2018b; *Braunagel et al.*, 2017).

3.3.4 Effects of features on random forest prediction results

The main effect of feature set on the model accuracy ($F(3, 119) = 304, p < .001$) and the F1-score ($F(3, 119) = 146, p < .001$) were significant at the 3 s time window. We found that the accuracy and F1-score of the random forest model using the full feature set were significantly higher than the accuracy and F1-score using other combinations of feature subsets at the 3 s time window (Figure 3.7 and Table 3.6). To be specific, if only environment factors were used as the features, the average prediction

Table 3.5: The top 16 important features when time window was 3s (GSR = galvanic skin response; NDRT = non-driving-related task).

Feature descriptions	Importance
Maximum of GSR in phasic component	.492
Mean of GSR in phasic component	.491
Standard deviation of GSR in phasic component	.441
Vertical gaze dispersion	.406
Scenario type	.404
Fixation duration	.371
Fixation duration on the driving scene	.352
Fixation duration on the NDRT	.341
Takeover request lead time	.338
Mean of inter-beat interval	.333
Mean of heart rate	.330
Eyes-on-the-road percentage	.323
Saccade number on the driving scenes	.314
Maximum heart rate	.295
Fixation number on the driving scenes	.282
Standard deviation of inter-beat interval	.268

Table 3.6: Random forest prediction accuracy and F1-score with different feature subsets at the 3 s time window and their comparisons to the full feature model.

Feature subsets	Accuracy				F1-score			
	mean	SD	*t*-test statistic	p-value	mean	SD	*t*-test statistic	p-value
All	.843	.010	-	-	.640	.015	-	-
Eye-tracking and environment	.818	.010	11.2	$p<.001$	.615	.013	10.9	$p<.001$
Physiological	.770	.020	17.2	$p<.001$	.563	.019	19.7	$p<.001$
Environment	.758	.005	42.7	$p<.001$	.611	.008	8.82	$p<.001$

accuracy and F1-score were only .758 and .611, respectively. If only physiological data were used as features, the average prediction accuracy was .770 and F1 score was 0.563. This suggests that a combination of environment features and features indicating drivers' states are necessary to build a model with high performance. The model using environment factors and eye-tracking metrics as features had an average accuracy of 0.818 and F1-score of 0.615 at the 3 s time window. After heart rate and galvanic skin response indices were added as features, the average model accuracy increased to 0.843, and the F1-score increased to 0.640.

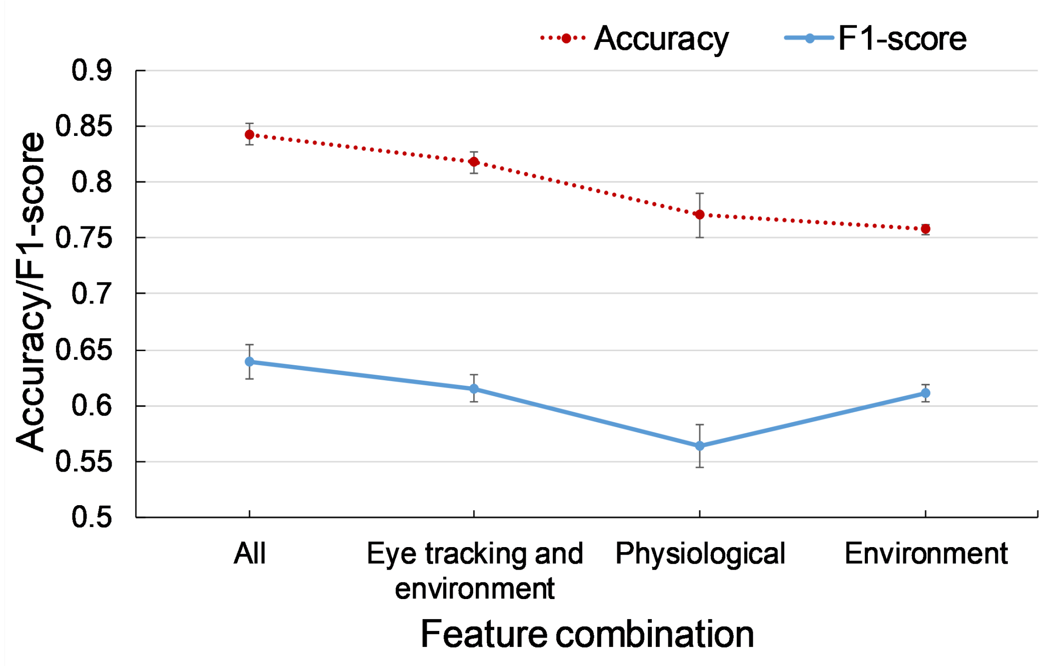


Figure 3.7: Prediction accuracy and F1-score of random forests with different feature subsets at the 3 s time window. Error bar indicates 1 standard deviation.

In addition, we ordered features according to the average feature importance values. Next, we built a random forest model with the most important feature, and then added features with lower importance one by one to build another 36 models. As shown in Figure 3.8, the model accuracy and F1-score generally increased at the beginning when more features were added, but reached a plateau when 16 or more features were included in the model. There was a main effect of feature numbers on the model accuracy ($F(36, 1109) = 3718, p < .001$) and F1-score ($F(36, 1109) = 293, p < .001$). Post hoc analysis showed that the F1-score of the full feature model was significantly higher than that for models with fewer than the top 9 important features, and accuracy of the full feature model was significantly higher than that of the models with fewer than the top 16 important features.

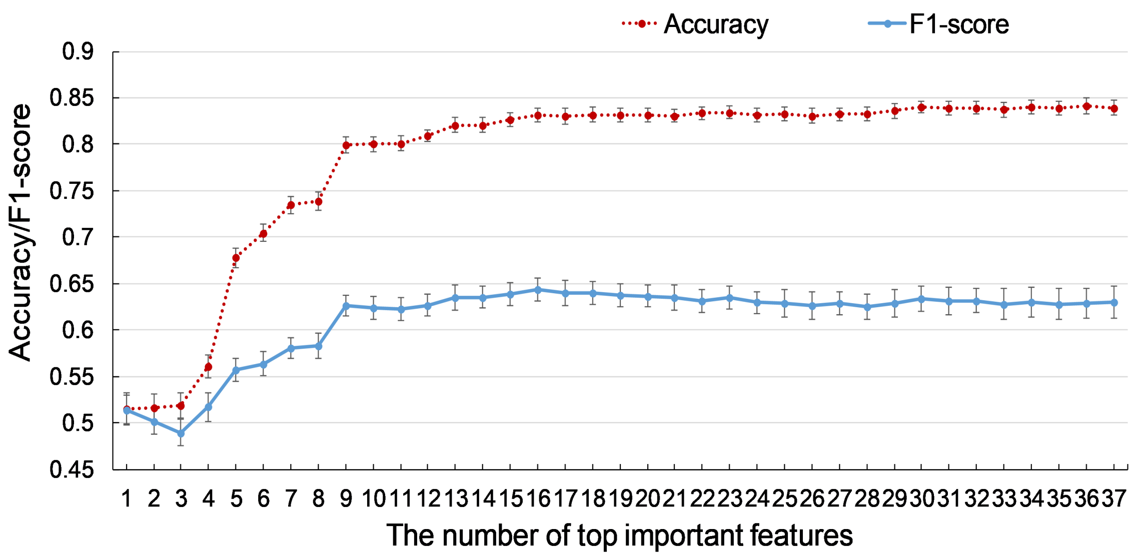


Figure 3.8: Model accuracy and F1-score with different numbers of top important features. Error bar indicates 1 standard deviation.

3.4 Discussion

3.4.1 Model performance comparisons

Our study compared the random forest model with the other five machine learning approaches used in prior literature for takeover performance prediction. As indicated by the results of model accuracy, F1-score, and ROC curve comparisons, the random forest approach outperformed the other classification approaches. Consistent with previous studies on drivers' fatigue and drowsiness detection (*McDonald et al.*, 2014; *Zhou et al.*, 2020a), the random forest approach also showed its supremacy for takeover performance prediction. It might be because random forests aggregate the results of many bootstrap aggregated (bagged) decision trees, which reduces the effects of overfitting and improves generalization.

3.4.2 Effects of window size on random forest prediction results

As the random forest outperformed other machine learning approaches, we examined the prediction performance of random forests under different time window sizes. The results showed that the window size significantly influenced random forest prediction performance. However, such a relationship was not linear. One of the explanations could be that we used a mixture of physiological signals as model inputs. Some physiological signals (e.g., pupil diameter) perform better with a shorter window size because they change rapidly according to the changes in the driver's cognitive workload (*Kramer et al.*, 2013). Some physiological signals (e.g., heart rate) perform better with a longer window size because they can provide an overall understanding of the driver's mental state (*Solovey et al.*, 2014). Future research is needed to explore model performance with customized time windows for different physiological signals.

It was important to find an optimal window size to calculate physiological features for model development in this study. Considering the implementation in real-world driving, a time window with shorter size and better prediction performance is preferred. Thus, we recommend 3 s as the optimal time window to predict takeover performance, with an accuracy of 84.3% and an F1-score of 64.0%. The post hoc analysis showed that the selection of time window for such performance is not unique. Time windows with a size of 6 s, 11 s, and 13-16 s led to similar prediction performance. Although the exact time window might be slightly different in the real world given the differences of situational and behavioral parameters, our study provides important insights on window size recommendation for the development of driver state detection systems.

Unlike previous studies, our model has a finer granularity and can predict drivers' takeover performance when they are engaged in a specific type of NDRTs with different levels of cognitive load. Such application differences make it infeasible to compare the exact accuracy and F1-score values with those in previous models. Because the test cases in our model prediction are from different participants and are not seen in the training set, our model can be used to predict takeover performance of a new driver who does not have historical data.

3.4.3 Effects of features on random forest prediction results

Drivers' galvanic skin responses, heart rate activities, and eye movements with a combination of environment factors were used to predict drivers' takeover performance. Building on *Braunagel et al.* (2017), we added GSR indices and HR indices for model development. Our results showed an improvement of model performance with a full set of features compared to other feature subsets (i.e., physiological data only, environment data only, eye-tracking and environment data). This aligns with the previous studies, because all these physiological signals reflected drivers' states and interactions with driving environments (*Mehler et al.*, 2012; *Radlmayr et al.*, 2014; *Wang et al.*, 2014; *Ratwani et al.*, 2010; *Young et al.*, 2013; *Bertola and Balk*, 2011).

Furthermore, we identified the most important features (e.g., maximum phasic GSR, gaze dispersion, scenario type, and mean of inter-beat interval) for model development. Although the model performance increased at the beginning as more features were added, it reached a plateau when 16 or more features were included. With the top 16 important features, we were able to develop a random forest model with comparable performance to the full feature model. Notably, the top 16 important features were extracted from galvanic skin responses, heart rate activities, eye movements, and environment factors, demonstrating the importance of all these data sources. Utilizing the advances of wearable technology and vehicle sensors, these features can be collected in a minimally invasive manner to predict drivers' takeover performance in real time.

3.4.4 Implications

Our study is a preliminary effort to predict drivers' takeover performance in designing advanced driver monitoring systems. With the advances of technologies in connected automated vehicle systems, real-time road environments such as traffic situations can be accessed easily in the future. Predictive model performance can be improved when data from various drivers engaging in different NDRTs in diverse environments are available for model training. The model outputs can contribute to the design of adaptive in-vehicle alert systems in conditionally automated driving. Specifically, if the system predicted that a driver would not be able to take over control successfully, a multi-modal display could be designed to help the driver realize the urgency of the event, increase situational awareness, and allocate attention properly. Eventually, it could improve drivers' takeover performance and enhance the safety and adoption of automated vehicles.

3.4.5 Limitations and future work

Several limitations should be taken into consideration in the future. First, this study used a snapshot of the time-series data as model inputs, without considering the complexity of sequence dependence among the data. Future studies could try a convolutional neural network (CNN) combined with long-short-term memory (LSTM) to predict drivers' takeover performance using a larger dataset. Second, the ground truth was determined by drivers' driving behaviors. It is necessary to propose a standard set of metrics for measuring takeover performance. An ensemble method combining subjective ratings, driving behaviors, and video coding can be explored to provide a more robust ground truth label of takeover performance. Third, instead of using dichotomous classification of takeover performance, we could increase the number of classes (e.g., bad, neutral, good; or very bad, bad, neutral, good, very good) or use regression to see model prediction power. Fourth, this study only recruited young adult participants with few AV experiences, and each participant only experienced four or eight takeover scenarios in the whole experiment. Future studies could recruit participants from different ages, AV experience levels, and training groups. Then the individual characteristics and power law of learning could be taken into account as model inputs to increase the generalization of models (*Forster et al.*, 2019).

3.5 Conclusion

This study developed a random forest model to predict drivers' takeover performance in conditionally automated driving. In contrast to previous models capable of predicting drivers' takeover performance when they performed different types of NDRTs, our model has a finer granularity and is able to predict takeover performance when drivers are engaged in a specific type of NDRTs. The results showed that the random forest classifier had an accuracy of 84.3% and an F1-score of 64.0% using a 3s time window, which outperformed other machine learning models used in prior studies. In addition, we identified the most important physiological measures for takeover performance prediction, and they can be used for developing in-vehicle monitoring systems. Such models can be used to guide the design of adaptive in-vehicle alert systems to improve takeover performance in conditionally automated driving in the future.

CHAPTER 4

Designing In-vehicle Alert Systems

4.1 Introduction

To help drivers take over control of the vehicle, this chapter presents two human-subject experiments on the design and evaluation of in-vehicle alert systems during takeover transitions. **Experiment 1** aimed to investigate drivers' acceptance (perceived usefulness, ease of use, attitude, and behavioral intention to use) of in-vehicle displays. We designed displays with different types of information and different modalities and evaluated them under different event criticality situations. **Experiment 2** aimed to examine the effects of display information, modality, and event criticality on drivers' performance expectancy, preference for displays, as well as their anxiety and preparedness during takeover transitions.

As the two experiments were conducted during the COVID-19 shelter-in-place season, only drivers' subjective ratings were collected via online studies. In the experiments, participants imagined themselves driving a Level 3 AV. The methods and results of each experiment are discussed and summarized in detail below.

4.2 Experiment 1

4.2.1 Method

This study complied with the American Psychological Association Code of Ethics and was approved by the institutional review board at the University of Michigan.

4.2.1.1 Participants

We conducted a power analysis using G*Power software. With a power value of 0.95, an effect size of 0.40, and α error probability of 0.05, we found that a sample size of 60 was needed for a mixed design.

Therefore, we recruited a total of 60 university students in the study (mean age = 24.03; SD = 5.34; range = 18-45; 25 females; 35 males). All the participants had the normal or corrected-to-normal vision, auditory acuity, and a valid driver's license. Participants were paid $10 for their 50-min online participation.

4.2.1.2 Apparatus and stimulus

The study made use of online videos and surveys. The videos simulated SAE Level 3 driving scenarios developed using an RTI fixed-base desktop driver simulator (RTI, MI, USA). In the automated driving mode, the AV handled longitudinal and lateral control, navigated directions, and followed traffic rules. The driver was not required to actively monitor the environment in this mode. However, when the AV met an unexpected scenario and was not able to negotiate the scenario, it would deactivate the automated driving mode and ask the driver to take control of the vehicle. The takeover request was issued in the form of a beep. As shown in Figure 4.1a, the speedometer and automated driving mode indicator were displayed on the dashboard and the view of the rear-view mirror was displayed on the top. The survey was developed using Qualtrics (Provo, UT, www.qualtrics.com).

(a) Simulated SAE Level 3 automated driving

(b) Experimental set up

Figure 4.1: Experimental settings

The NDRT was to type and chat with the experimenter using the Google Chat application on the phone. The chat topics were daily conversations related to weather, hobbies, movies, music, food, and so forth. As shown in Figure 4.1b, participants were asked to place their computers right in front of them, and their phones 45 degrees to the right, to simulate in-vehicle environments in the real world (i.e., their view of the phone did not overlap with their view of the driving scenes).

4.2.1.3 Experimental design

The study used a 3 (information type) $\times$ 3 (display modality) $\times$ 2 (event criticality) mixed design. The information type and event criticality were the within-subjects variables, and display modality was the between-subjects variable. Event criticality was manipulated by TOR lead time of scenarios, which was defined as the critical event onset for failures at the time of request (*Du et al.*, 2020d; *Eriksson and Stanton*, 2017). We set scenarios with 4s lead time as high criticality scenarios and 7s lead time as low criticality scenarios. There were three types of information: *why*, *what will*, and *why + what will*; and three levels of display modality: augmented reality, speech, and the combination of augmented reality and speech. The *why* information was the perception and comprehension of the elements in the environments, which

Table 4.1: Descriptions of lane changing takeover events and corresponding display information

Scenarios	*Why* in Speech	*What will* in Speech	*Why* in AR HUD	*What will* in AR HUD
Swerving vehicle ahead	"Swerving vehicle ahead"	"Change to right lane"	Highlight swerving vehicle	Right lane change sign
Construction zone ahead	"Construction zone ahead"	"Change to right lane"	Highlight construction zone	Right lane change sign
Police vehicle on the shoulder	"Police vehicle on shoulder"	"Change to left lane"	Highlight police vehicle	Left lane change sign
Exit highway	"Highway exit ahead"	"Change to exit lane"	Highlight exit sign	Right lane change sign
Road closure ahead	"Road closure ahead"	"Change to left lane"	Highlight road closure sign	Left lane change sign
Bicyclist ahead	"Bicyclist ahead"	"Change to left lane"	Highlight bicyclist	Left lane change sign

explained why there was a takeover request. The *what will* information consisted of the decision-making result, which recommended future actions to negotiate the takeover scenarios.

Based on prior literature (*Koo et al.*, 2016b; *Miller et al.*, 2016; *Rezvani et al.*, 2016; *Du et al.*, 2019a,b, 2020c), 12 takeover scenarios were designed in urban and rural areas with typical roadway features. Experimental conditions were randomly paired up with one lane keeping scenario and one lane changing scenario. The order of different conditions was random for each participant. The display information in both speech and augmented reality formats are presented in Tables 4.1 and 4.2. Two examples of AR HUDs in the experiment are shown in Figure 4.2.

The dependent variables were based on the Technology Acceptance Model (*Davis*, 1989). We used 7-point Likert scales (strongly disagree to strongly agree) to measure participants' perceived ease of use, perceived usefulness, behavioral intention to use, and attitude toward using the display (*Venkatesh and Davis*, 2000, 1996). We also asked participants to rank the usefulness of three types of information and answer an open-ended question about whether the displays helped them to take over and why.

Table 4.2: Descriptions of lane keeping takeover events and corresponding display information

Scenarios	*Why* in Speech	*What will* in Speech	*Why* in AR HUD	*What will* in AR HUD
Yellow lights flashing	"Yellow light flashing ahead"	"Proceed with caution	Highlight traffic light	Yield sign
Construction zone on the left	"Construction zone on the left"	"Stay in the lane"	Highlight construction zone	Arrows showing road direction
Speed limit failure	"Speeding on the curve"	"Slow down to 25 mph"	Highlight the speed limit sign	Slow down and speed limit sign
Red lights flashing	"Red light flashing ahead"	"Stop before proceeding"	Highlight traffic light	Stop sign
Pedestrian cut in	"Pedestrian in the road"	"Stop before proceeding"	Highlight Pedestrian	Stop sign
Sensor error	"Sensor error"	"Stay in your lane"	Sensor error symbol	Arrows showing road direction

(a) Police vehicle on shoulder

(b) Red light flashing ahead

Figure 4.2: Two examples of AR HUDs in the experiment

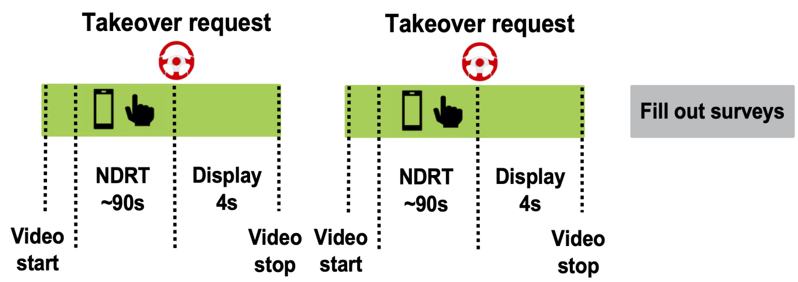


Figure 4.3: Experimental procedure

4.2.1.4 Experimental procedure

The study was conducted in the form of a Zoom video meeting. After participants clicked the "Agree" button on the consent form, the experimenter provided information on the experiment environment and stimuli, as well as instructions on tasks. In the real experiment, each participant experienced 6 blocks (2 event criticality levels × 3 information types). In each block, once the video started playing, participants were asked to imagine themselves as the driver of an SAE Level 3 AV (Figure 4.3). Then they needed to chat with the experimenter using the Google Chat application on their phone. Approximately 90 seconds later, an unexpected takeover request was issued, followed by displays in one of the conditions. This procedure was repeated again to cover both lane keeping and lane changing scenarios. At the end of the block, participants filled out the survey to indicate their subjective ratings of the alert systems and answer an open-ended question. We also captured their demographic information at the end.

4.2.1.5 Data analysis

We constructed linear mixed models using SPSS version 24 to examine effects on dependent variables. The display modality, information type, event criticality, and their interactions were treated as fixed effects. Subjects were treated as random effects to resolve non-independence in all the linear mixed models. Considering that the usefulness ranking was an ordinal variable, we used the Friedman test to examine the differences of usefulness ranking among conditions. For participants' answers to the open-ended question about whether the displays helped them to take over and why, we categorized the valence of the answers into three groups (i.e., positive, negative, and mixed), and then used a generalized estimating equation model to examine the effects of independent variables. Five answers were recorded as "N/A" to the open-ended question. Results are reported as significant for α less than .05 and as marginally significant for α less than 0.1 (*Gelman*, 2013). All post-hoc comparisons utilized a Bonferroni alpha correction.

4.2.2 Results

4.2.2.1 Measurement reliability and validity

The reliability (i.e. Cronbach's α values) of perceived usefulness ($\alpha = .898$), ease of use ($\alpha = .953$), behavioral intention to use ($\alpha = .868$), and attitude ($\alpha = .912$) exceeded the recommendation value ($\alpha = .7$) (*Carmines and Zeller*, 1979).

To assess the convergent validity of a construct, the AVE should be higher than 0.50 (*Fornell and Larcker*, 1981). When the AVE value is above 0.50, the variance explained by the construct is greater than the variance explained by measurement error, which indicates evidence of convergent validity of the construct. The AVEs of perceived usefulness, perceived ease of use, behavioral intention to use, and attitude were 0.55, 0.66, 0.55, and 0.46, respectively. Except attitude, all of the other variables were above the threshold of 0.50, as recommended by *Fornell and Larcker* (1981). To assess discriminant validity, we compared the correlations of all constructs with the square root of the AVE values of perceived usefulness, perceived ease of use, behavioral intention to use, and attitude. The correlation matrix, shown in Table 4.3, indicates that except for behavioral intention to use and attitude, correlations among constructs were well below the square root of AVEs, which is further evidence of discriminant

Table 4.3: Correlation matrix (Values on the diagonals within the parentheses represent the square root of the AVE for each factor)

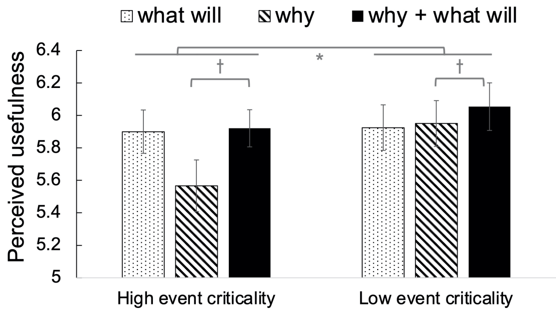
	Perceived usefulness	Perceived ease of use	Behavioral intention to use	Attitude
Perceived usefulness	(0.741)	.726**	.757**	.841**
Perceived ease of use		(0.815)	.526**	.715**
Behavioral intention to use			(0.739)	.794**
Attitude				(0.678)

validity among dependent measures. Given that the items of behavioral intention to use and attitude did not meet the validity requirement, we only analyzed the results of perceived usefulness and ease of use below.

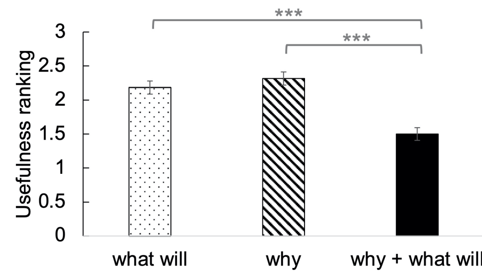
4.2.2.2 Effects on perceived usefulness

As shown in Figure 4.4a, the main effect of information type on perceived usefulness was marginally significant ($F(2, 289) = 2.548, p = .08$). The post-hoc analysis indicated that *why* only information led to marginally lower perceived usefulness compared to *why + what will* information ($p = .081$). The main effect of event criticality was also significant ($F(1, 289) = 4.564, p = .033$). High event criticality led to low perceived usefulness. All the other main effects and their interaction effects were not significant.

With regard to the usefulness ranking, the Friedman test showed that participants ranked three conditions of information types significantly differently ($\chi^2_2 = 23.03, p < .001$). Post-hoc comparisons indicated that participants gave the highest usefulness ranking when they were presented with *why + what will* information (*why + what will* vs. *why*: $Z = -4.045, p < .001$; *why + what will* vs. *what will*: $Z = -3.717, p < .001$; See Figure 4.4b). Nonetheless, there was no significant difference in usefulness ranking between *why* only information and *what will* only information ($Z = -.836, p = .403$).



(a)



(b)

Figure 4.4: Drivers' perceived usefulness (1 is the lowest usefulness) and usefulness ranking (1 is the highest ranking).

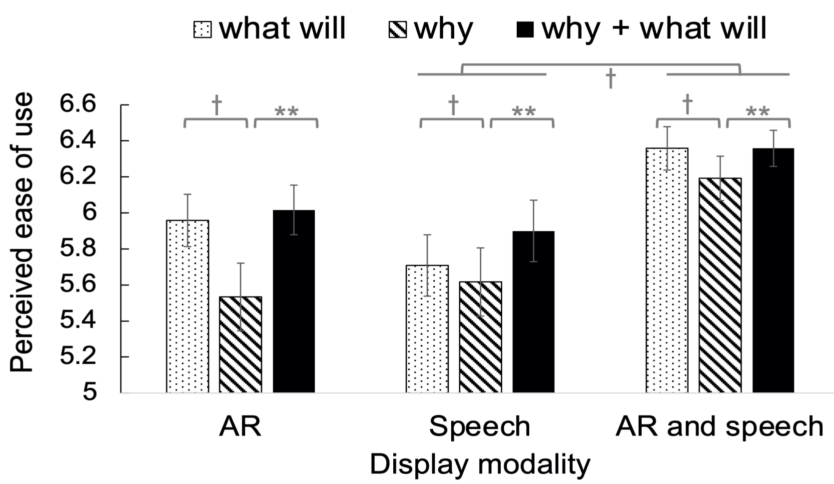


Figure 4.5: Drivers' perceived ease of use

4.2.2.3 Effects on ease of use

There was a significant main effect of information type on ease of use ($F(2, 289) = 4.977, p = .007$). To be specific, *why* only information led to the lowest ease of use (*why* vs. *what will*: $p = .077$; *why* vs. *why* + *what will*: $p = .008$; See Figure 4.5). The main effect of display modality was significant ($F(2, 57) = 3.409, p = .04$). The post-hoc analysis showed that the combination of AR HUD and speech led to marginally higher ease of use compared to speech ($p = .053$, See Figure 4.5). No other significance was found.

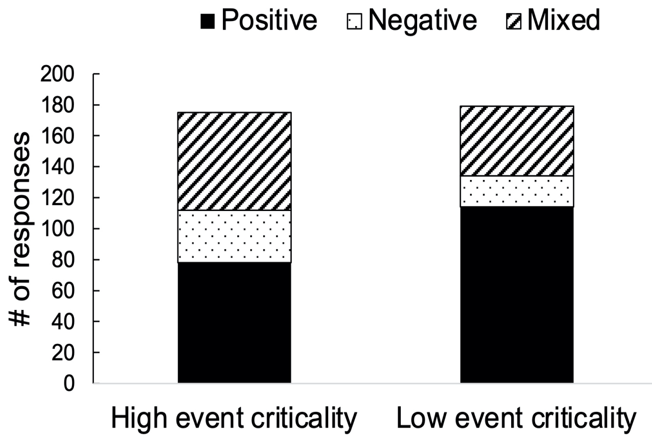


Figure 4.6: The valence distribution of drivers' answers to the open-ended question.

4.2.2.4 Interview results

There was a significant main effect of event criticality ($\chi_1^2 = 10.479, p = .001$) on the valence of participants' answers to the open-ended question (i.e., whether the displays helped them to take over and why). As shown in Figure 4.6, low event criticality led to more positive opinions of the displays than high event criticality. Participants' remarks are summarized below ("n" below indicates the number of responses; P indicates index of the participant).

Modality. Participants pointed out the pros and cons of different display modalities. For AR HUD, participants reported that it was "*clear*" (P7), "*simple*" (P33), and "*easy to visualize actions*" (P45). For speech, participants noted that it was "*clear*" (P57) and "*to the point*" (P9). Yet, the AR HUD showed potential issues such as "*low contrast to the background*" (n = 6) and "*small*" (n = 7). The problems of speech information included "*fast*" (n = 5), "*difficulty to understand*" (n = 13), and "*muffled*" (n = 7). When the information was presented in the combination of AR HUD and speech, the two modalities compensated for each other, providing more comprehensive descriptions and increasing the communication effectiveness. For example, P45 said: "*It helped identify where the obstacle is at on the road and tell me in the speech what the obstacle is for me to be prepared.*"

Information type. Participants who had positive opinions of display information stated that *why* information helped them "*focus on what's wrong*" (P22), "*direct attention to the right areas*" (P19), and "*identify what to look out for*" (P58). They described that *what will* information "*showed appropriate takeover actions*" (P34), "gave instructions on adjustments to make" (P19), and "*alerted what should be done*" (P14).

However, some participants expressed negative opinions of display information. With the *what will* only information, participants mentioned that the *why* information could also be helpful (n = 10). Similarly, with the *why* only information, participants reported that they also wanted the *what will* information (n = 9). Yet, we found 5 responses stating that the *what + why* information was too much. For example, P60 said: "*time to react was wasted by announcing the danger ahead without an immediate command.*"

Meanwhile, participants proposed other reasons why the display information was not helpful enough. First, participants may already have recognized the situation when the information was delivered (n = 10). Second, participants would have liked more information such as vehicles in the neighboring lanes to be highlighted (n = 4). Third, participants expressed concerns about the appropriateness of the recommended actions and reliability of the information in the long run (n = 13).

Event criticality. In high event criticality conditions, there were 44 responses reporting that participants did not have enough time to handle the takeover requests. For example, P32 said: "*I would have liked it to be a little earlier*" and P25 said: "*This display timing felt rushed.*" There were only 11 responses reporting that participants had "sufficient" time to respond (P25) when event criticality was high.

In low event criticality conditions, the pattern was opposite. We found 36 responses noting that the display "*was in time*" (P1) and "*gave enough time to respond*" (P16). There were only four responses stating that the display was "*at the* *last moment*" (P24) and "*too late*" (P23).

4.2.3 Discussion

This study investigated the impacts of different display information, modalities, and event criticality on the acceptance of an in-vehicle alert system designed to support takeover requests in simulated Level 3 automated driving.

4.2.3.1 Effects of display modality

Compared to the combination of AR HUD and speech, the speech only modality led to lower ease of use. Different from visual information, it takes several seconds to deliver speech messages. In our experiment, some participants mentioned that the speech was fast and difficult to understand. This may be because we set a relatively high speech speed to finish speech information delivery in a timely fashion during the takeover transitions. Such information-speed trade-off is inevitable when presenting information in speech modality. Some participants reported that the speech was muffled by the engine noise in the background. This may have resulted from the driving simulator settings in the experiment. Also, as mentioned by the existing literature, speech messages were somewhat annoying (*Nees et al.*, 2016). We speculate that the disadvantages of speech may be amplified in the long run.

Participants had similar perceived ease of use with AR HUD compared to the combination of AR HUD and speech. This means that without further speech explanations, the information from AR HUD was easy to understand and use. One possible reason may be that we primarily used official traffic signs and provided training on new ones at the beginning of the experiment. According to participants' answers to the open-ended question, we gathered some suggestions to improve the design of AR HUD in the future. For example, we can present information using symbols with a high contrast from the background.

4.2.3.2 Effects of information type

The *why* only information led to the lowest perceived usefulness and ease of use, consistent with existing studies (*Eriksson et al.*, 2018). The *why* information helps drivers to perceive the traffic elements and road situations. Yet, drivers need time and effort to process the information, assess the environment, and make decisions. The *what will* information indicates a higher level of information processing. Given that the information provided in the display was 100% accurate, it was beneficial to directly utilize the *what will* information in the display. This may be the reason why drivers found the *what will* information had comparable perceived usefulness and ease of use to the *why + what will* information. However, if participants were forced to rank the usefulness of the three information types, the *why + what will* information was ranked at the top of the list. This aligns with participants' responses to the open-ended questions, as they preferred more information when the *why* only information or the *what will* only information was presented. Although the interaction effects between event criticality and display information were not significant, there was a trend that rating differences of different information types were inflated in high event criticality conditions compared to the low event criticality conditions.

4.2.3.3 Effects of event criticality

In general, the usefulness of the displays depended on the event criticality. The displays had lower perceived usefulness when event criticality was high. This may be because drivers did not have enough time to process and utilize the display information in urgent scenarios. However, we failed to find any interaction effects between event criticality and display information/modalities. This could be explained by the fact that the dependent variables measured in this study were not sensitive to the effects of display modality/information in different event criticality conditions. Future studies can investigate the effects of display information and modality on other dependent variables of interest in different event criticality conditions.

4.2.3.4 Limitations and future work

The study has some limitations that can be addressed in future studies. First, due to COVID-19, we only collected subjective measurements using an online study. Future researchers can conduct lab studies or on-road studies to measure drivers' driving behaviors. The objective measurements can reflect the effects of information type and modality in a more comprehensive way. The on-road testings will inform us of the effectiveness of these systems in real-world conditions. Second, the event-related information provided in this study was 100% accurate. Information in real-world scenarios may involve some errors due to sensor limitations and complex driving environments. Future studies can consider the information reliability factor and measure drivers' trust in displays to see how results change. Third, this experiment studied the effects of display information and modality. Future studies can examine the effects of different display parameters (e.g., speech tone and speed) for a given display, and find the best ones to present the information.

4.2.4 Conclusion

This study compared different types of display information and modality to convey event-related information during takeover transitions. The results showed that presenting *why* only information was not sufficient considering drivers' perceived ease of use and perceived usefulness. The combination of speech and augmented reality modality was superior to the speech only modality from the perspective of ease of use. Also, displays showed higher perceived usefulness in low criticality scenarios. The results have implications for the design of in-vehicle displays during takeover transitions. The recommendations on the display information and modality can potentially speed up takeover reactions and improve takeover quality.

However, we failed to find any significant interaction effects between event criticality and display information/modalities. The possible reason may be that the dependent variables measured in this study were not sensitive to the effects of display modality/information in different event criticality conditions. Thus, we conducted Experiment 2 to investigate the effects of display information, modality, and event criticality on other dependent variables of interest.

4.3 Experiment 2

4.3.1 Method

This study complied with the American Psychological Association Code of Ethics and was approved by the institutional review board at the University of Michigan.

4.3.1.1 Participants

The survey was developed using Qualtrics (Provo, UT, www.qualtrics.com) and administered online in the U.S. population through Amazon Mechanical Turk (MTurk) (https://www.mturk.com). Amazon Mechanical Turk provides access to a virtual community of workers who are willing to complete human intelligence tasks (HITs) at their convenience. The qualification requirements to conduct the task include: (1) HIT Approval Rate (%) for all Requesters' HITs greater than 98; (2) Number of HITs Approved greater than 500; and (3) having a valid driver's license.

We conducted a power analysis using G*Power software. With a power value of 0.95, the small effect size of 0.25, and α error probability of 0.05, we found that a sample size of 303 was needed for a between-subjects design. A total of 515 subjects participated in the study (mean age = 41.6; SD = 12.7; range = 19-77; 267 males; 239 females). Participants were paid $1 for their online participation. After the attention check questions were analyzed, only 479 participants were available for further data analysis. Hence, 479 participants collected in the study were enough to gain power.

4.3.1.2 Apparatus and stimulus

The study used online videos and surveys. The video stimuli were the same as in Experiment 1, which was detailed in Section 4.2.1.2. Participants were asked to perform an NDRT while the vehicle was in automated mode. The NDRT was to engage in any task on the phone (e.g., play games, read the news, type). Similar to Experiment 1, participants were required to place their computers right in front of them and their phones 45 degrees to the right to simulate in-vehicle environments in the real world.

4.3.1.3 Experimental design

The study used a 3 (information type) $\times$ 3 (display modality) $\times$ 2 (event criticality) between-subjects design. Event criticality was manipulated by TOR lead time of scenarios, which was defined as the critical event onset for failures at the time of request (*Du et al.*, 2020d; *Eriksson and Stanton*, 2017). We set scenarios with 4s lead time as high criticality scenarios and 7s lead time as low criticality scenarios. There were three types of information: *why*, *what will*, and *why + what will*; and three levels of display modality: AR HUD, speech, and the combination of AR HUD and speech. The scenarios and corresponding display information in different modalities were the same as in Experiment 1, which was detailed in Section 4.2.1.3. Each participant was randomly assigned to one out of eighteen conditions. The number of participants in each condition is shown in Table 4.4, Table 4.5, and Table 4.6.

Table 4.4: Distributions of participants in augmented reality conditions

Modality	Augmented reality					
Information	What will		What will + Why		Why	
Event criticality	High	Low	High	Low	High	Low
Count	27	23	24	24	32	27

Table 4.5: Distributions of participants in speech conditions

Modality	Speech					
Information	What will		What will + Why		Why	
Event criticality	High	Low	High	Low	High	Low
Count	31	24	29	25	26	28

Table 4.6: Distributions of participants in the combination of speech and augmented reality conditions

Modality	Speech and augmented reality					
Information	What will		What will + Why		Why	
Event criticality	High	Low	High	Low	High	Low
Count	26	27	24	28	27	27

The dependent variables were based on the Unified Theory of Acceptance and Use of Technology, as introduced in Chapter 1. Given the fact that it was the first time that participants experienced the display and the product did not exist in the real world, Social Influence and Facilitating Conditions were not applicable to be measured in this experiment. In addition, we measured participants' anxiety and preference. The preference and anxiety questionnaire was adapted from a published model from the CHIMe Lab at Stanford University that is used to measure driver attitude (*Koo et al.*, 2016a; *Nass et al.*, 2005; *Takayama and Nass*, 2008). Anxiety comprised the averaged responses to four adjective items that described the AV experience: fearful, afraid, anxious, and uneasy. Preference for AV comprised the averaged responses to eight items: intelligent, effective, reliable, helpful, smart, dependable, high quality, and efficient. All the items were rated on seven-point rating scales (1: describes very poorly; 7: describes very well). We adapted the scale to reflect the in-vehicle human-machine interface context. We also asked participants how prepared they felt if they had to take control of the vehicle at the moment of notice, and used this to measure the degree of takeover preparedness.

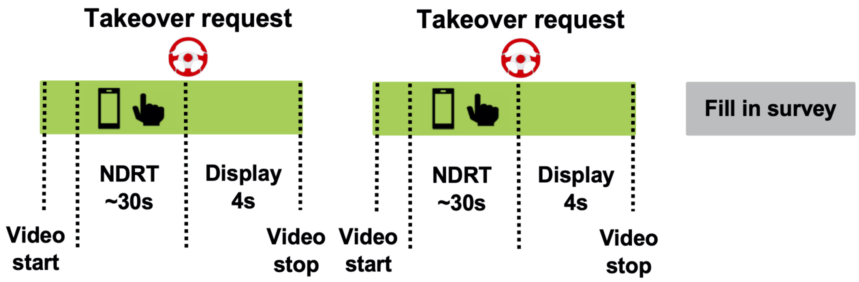


Figure 4.7: Experimental procedure

4.3.1.4 Experimental procedure

After participants clicked the "Agree" button on the consent form, they needed to read instructions on the experiment environment setup and the introduction of stimuli. When they passed the instruction screening questions, they were allowed to participate in the real experiment. In the real experiment, each participant was randomly assigned to one out of eighteen conditions. In each condition, once the video started playing, participants were asked to imagine themselves as the driver of an SAE Level 3 AV (Figure 4.7). Then they needed to play on their phone by reading the news or playing games, etc. Approximately 30 seconds later, an unexpected takeover request was issued, followed by displays in one of the conditions. This procedure was repeated again to cover both lane keeping and lane changing scenarios. At the end of the block, participants needed to fill in the survey to indicate their subjective ratings of the displays. Participants' demographic information was recorded in the end.

4.3.1.5 Data analysis

We constructed linear mixed models using SPSS version 24 to examine effects on dependent variables. The display modality, information type, event criticality, and their interactions were treated as fixed effects. Subjects were treated as random effects to resolve non-independence in all the models. Results are reported as significant for α less than .05. All post-hoc comparisons utilized a Bonferroni alpha correction.

Table 4.7: Correlation matrix (Values on the diagonals within the parentheses represent the square root of the AVE for each factor)

	Performance Expectancy	Effort Expectancy	Preference	Anxiety
Performance Expectancy	(0.774)	.739**	.724**	-.446**
Effort Expectancy	.739**	(0.679)	.692**	-.528**
Preference	.724**	.692**	(0.72)	-.459**
Anxiety	-.446**	-.528**	-.459**	(0.895)

4.3.2 Results

4.3.2.1 Measurement reliability and validity

The reliability of performance expectancy (α = .943), effort expectancy (α = .898), preference (α = .957), and anxiety (α = .962) exceeded the recommendation value (α = .7) (*Carmines and Zeller*, 1979).

To assess the convergent validity of a construct, the AVE should be higher than 0.50 (*Fornell and Larcker*, 1981). When the AVE value is above 0.50, the variance explained by the construct is greater than the variance explained by measurement error, which indicates evidence of convergent validity of the construct. The AVEs of performance expectancy, effort expectancy, preference, and anxiety were 0.598, 0.461, 0.519, and 0.801, respectively. Except for effort expectancy, all of the other variables were above the threshold of 0.50, as recommended by *Fornell and Larcker* (1981). To assess discriminant validity, we compared the correlations of all constructs with the square root of the AVE values of performance expectancy, effort expectancy, preference, and anxiety. The correlation matrix, shown in Table 4.7, indicates that except effort expectancy, correlations among constructs were below the square root of AVEs, which is further evidence of discriminant validity among dependent measures. Given that the effort expectancy items did not meet the validity requirement, we did not analyze their results below.

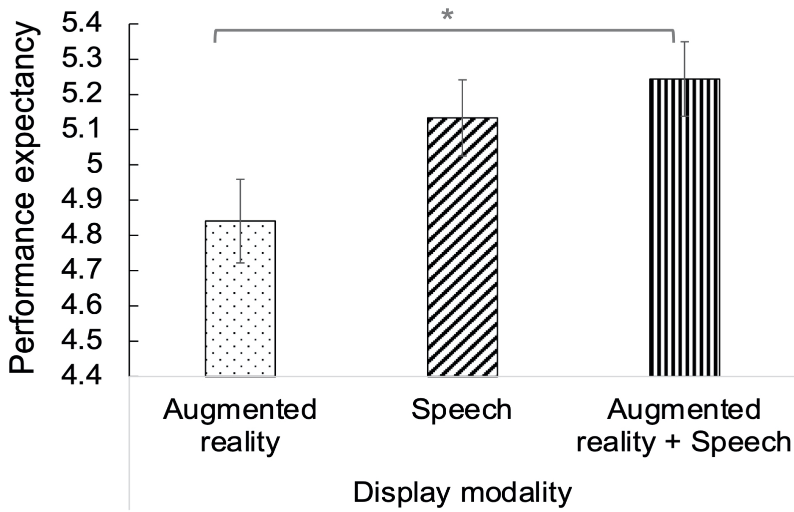


Figure 4.8: Drivers' performance expectancy in displays

4.3.2.2 Effects on performance expectancy

As shown in Figure 4.8, the main effect of modality on performance expectancy was significant ($F(2, 465) = 3.239, p = .04$). The post-hoc analysis indicated that drivers had lower performance expectancy towards AR HUD compared to the combination of AR HUD and speech conditions ($p = .043$). None of the other effects were significant.

4.3.2.3 Effects on preference

The interaction effect between display modality and event criticality was significant ($F(2, 465) = 3.969, p = .02$). The simple analysis result showed that drivers had the highest preference in the combination of AR HUD and speech conditions when event criticality was high (AR HUD + speech vs. AR HUD: $p = .005$; AR HUD + speech vs. speech: $p = .022$; See Figure 4.9). Yet, there was no significant preference difference between different display modality conditions in low event criticality conditions. No other significant effects were found.

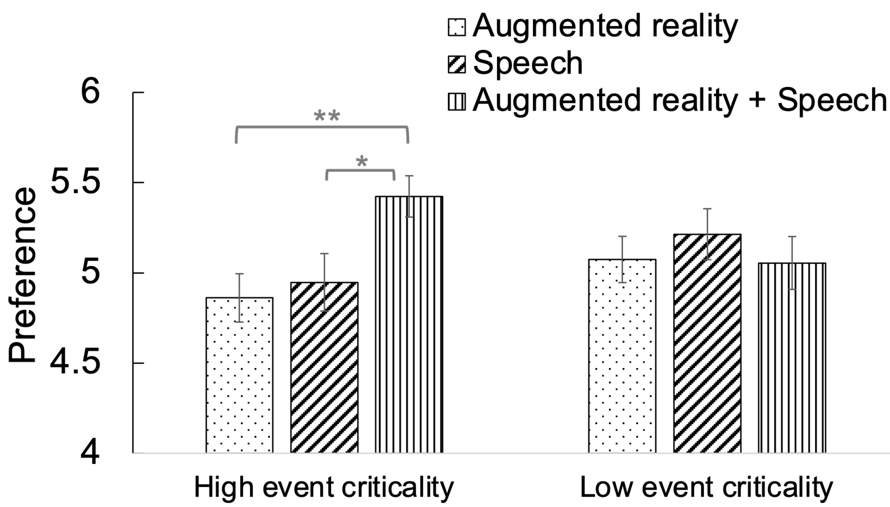


Figure 4.9: Drivers' preference for displays

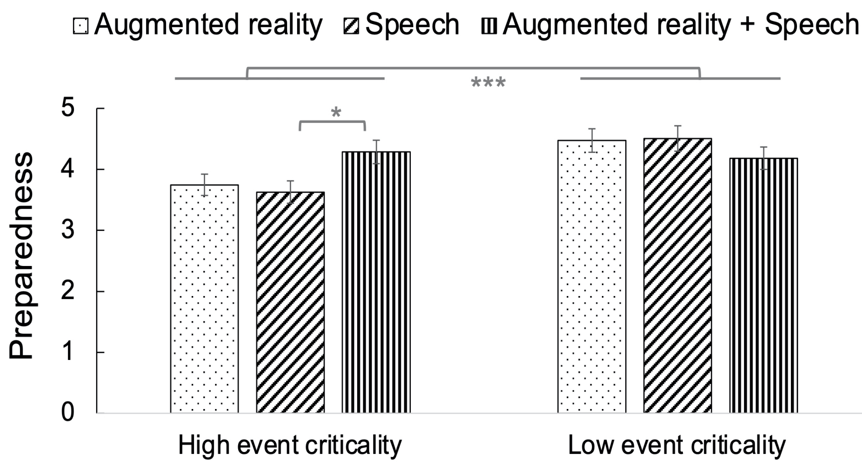


Figure 4.10: Drivers' takeover preparedness in different conditions

4.3.2.4 Effects on preparedness

There was a significant main effect of event criticality on drivers' takeover preparedness ($F(1, 465) = 10.881, p = .001$). Drivers felt more prepared in low event criticality conditions. As shown in Figure 4.10, the interaction effect between display modality and event criticality was significant ($F(2, 465) = 3.983, p = .019$). In high event criticality conditions, drivers felt less prepared in speech conditions than in the combination of AR HUD and speech conditions ($p = .038$). But such difference disappeared in low event criticality conditions.

4.3.2.5 Effects on anxiety

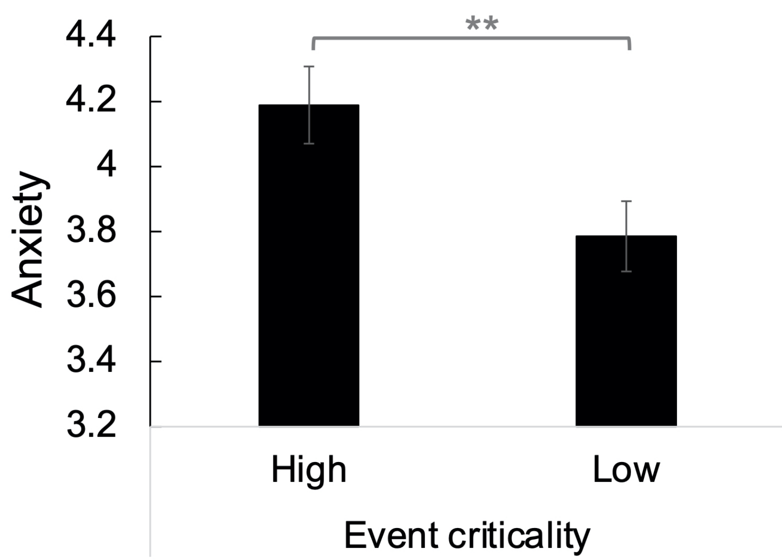
In general, drivers were more anxious in high event criticality conditions than low event criticality conditions ($F(1,465) = 7.035, p = .008$, See Figure 4.11a). As shown in Figure 4.11b, the interaction effect between modality and information was significant ($F(4,465) = 3.223, p = .013$). Only when information was presented in the speech modality, *why* only information led to higher anxiety compared to *why + what will* information ($p = .02$). The effects of information on drivers' anxiety were not significant in other modality conditions.

4.3.3 Discussion

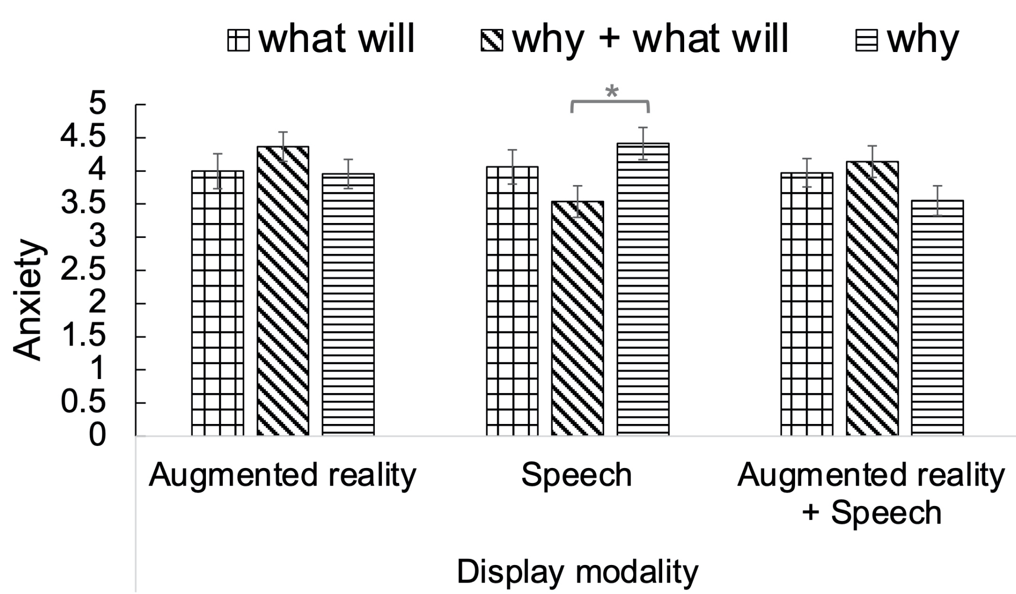
This experiment investigated how different display information, modality, and event criticality influenced drivers' performance expectancy, preference, anxiety, and preparedness towards in-vehicle alert systems after takeover requests in simulated SAE Level 3 automated driving.

4.3.3.1 Effects on preference and preparedness

When event criticality was high, in the combination of AR HUD and speech conditions, drivers felt more prepared to takeover control of the vehicle and had more preference compared to the AR HUD only or speech only conditions. Such a phenomenon disappeared when event criticality was low, which meant different display modalities did not influence drivers' preference and takeover preparedness in low event criticality conditions. This may be because drivers preferred a more direct and comprehensive display in an urgent situation. Yet, when the event was not that critical, drivers had more time and effort to perceive the driving environments, process the information, and make decisions on their own. They could also convert the information from one modality to another by themselves. Therefore, the display modality did not influence their preference and takeover preparedness in low event



(a)



(b)

Figure 4.11: Drivers' anxiety in different conditions

criticality conditions. Generally speaking, high event criticality led to lower takeover preparedness and higher anxiety. This is consistent with existing literature (*Du et al.*, 2020c).

4.3.3.2 Effects on anxiety

When different types of information were presented in the AR HUD or the the combination of AR HUD and speech conditions, drivers had similar anxiety during the takeover transitions. However, drivers' anxiety differed significantly when different types of information were presented in the speech modality. To be specific, when information was presented in the speech modality, drivers had more anxiety about the *why* only information compared to the *why + what will* information. This could potentially result from the reasons described below. Compared to other modalities, it took several seconds for speech information to be delivered. Although *why* only information helped drivers to perceive the traffic elements and road situations, they still needed to assess the environments and make decisions after environment perception. Thus, drivers might have been more anxious about the *why* only information when it was delivered in the speech modality.

4.3.3.3 Effects on performance expectancy

The results showed that drivers had higher performance expectancy with the the combination of AR HUD and speech conditions compared to AR HUD conditions. One potential reason could be as follows. Although the augmented reality displays overlapped with the driving environments and did not require much extra effort to access, sometimes drivers may not have understood the information directly. More intensive training on using augmented reality displays is needed before they are widely implemented on the market.

4.3.3.4 Limitations and future work

The findings of the present study should be interpreted in light of the following limitations. First, although we added attention check questions to exclude the participants who did not pay attention to the instructions and questions during the experiment, we were not able to ensure that participants fully followed our instructions such as playing on their phones in the automated driving mode. Second, we only collected subjective measurements using an online study. Future studies can conduct lab studies or on-road studies to provide more strict experimental control and measure drivers' driving behaviors. The objective measurements can reflect the effects of information type and modality in a more comprehensive way. Third, this study was a one-time online study. Future researchers can use a longitudinal study method to see how drivers' acceptance of technology changes over a given time period.

4.3.4 Conclusion

This study investigated how display information, modality, and event criticality influenced drivers' performance expectancy, preference, anxiety, and preparedness during takeover transitions. The results showed that when information was presented in the speech modality, drivers had more anxiety about *why* only information compared to *why + what will* information; when drivers were in high event criticality situations, they felt more prepared to take over control of the vehicle and had more preference for the combination of AR HUD and speech conditions than others.

The results have implications for the design of in-vehicle displays during takeover transitions. The recommendations on the display information and modality in different scenarios can enhance the interactions between AVs and drivers and can finally improve the takeover performance.

CHAPTER 5

Conclusion

5.1 Summary

To address takeover difficulty in conditionally automated driving and fill in the current research gaps, the aims of this dissertation research were to:

(1) Examine the effects of drivers' cognitive load, emotions, traffic density, TOR lead time on their driving behavioral (takeover timeliness and quality) and psychophysiological (eye movements, galvanic skin responses, and heart rate activities) responses to takeover requests.

(2) Develop computational models to predict drivers' takeover performance using drivers' physiological data and driving environment data via machine learning algorithms.

(3) Design in-vehicle alert systems with different display modalities and information types, and evaluate them in different takeover situations using human-subject experiments.

To meet Aim 1, we conducted three human-subject experiments. In Experiment 1, by systematically manipulating drivers' emotional states, we found that positive emotional valence led to smoother takeover behaviors, while emotional arousal did not influence takeover behaviors. In Experiment 2, we found that drivers had worse takeover performance when they had a high cognitive load, short takeover request lead time, and heavy oncoming traffic density in general. Interestingly, if drivers had low cognitive load, they paid more attention to driving environments and responded more quickly to takeover requests in high oncoming traffic density conditions. With regard to psychophysiological responses, we found that drivers had lower heart rate variability, narrower horizontal gaze dispersion, and shorter eyes-on-road time when they were in high cognitive load during the automated driving stage. Upon the TOR, $\sim$4s lead time led to inhibited blink numbers and larger maximum and mean GSR phasic activation, indicating higher emotional arousal and stress than $\sim$7s lead time. Meanwhile, heavy traffic density resulted in significantly frequent HR acceleration patterns than light traffic density, suggesting ignorance of overwhelming traffic information. In Experiment 3, we found that lane changing scenarios, in general, led to shorter TOR reaction time, wider horizontal gaze dispersion, and many more HR acceleration patterns than lane keeping scenarios. With a fixed distance between the AV and ahead objects/curvature entrance, high speed (i.e., short TOR lead time) generally led to worse takeover quality in the form of shorter minimum TTC, maximum resulting acceleration, and larger standard error of the steering wheel. More interestingly, compared to low speed, high speed decreased drivers' takeover performance and led to suppressed blinks and larger GSR phasic activation in lane changing scenarios, suggesting high cognitive load and stress in such attention-demanding and urgent situations. However, such effects were not significant or even reversed in lane keeping scenarios.

To meet Aim 2, we developed a random forest model to predict drivers' takeover performance in conditionally automated driving. In contrast to previous models capable of predicting drivers' takeover performance when they performed different types of NDRTs, our model has a finer granularity and is able to predict takeover performance when drivers are engaged in a specific type of NDRTs. The results showed that the random forest classifier has an accuracy of 84.3% and an F1-score of 64.0%

using a 3s time window, which outperformed other machine learning models used in prior studies. In addition, we identified the most important physiological measures and environment features for takeover performance prediction, and they can be used for developing in-vehicle monitoring systems.

To meet Aim 3, we compared different types of display information and modality to convey event-related information during takeover transitions. In Experiment 1, the results showed that presenting *why* only information was not sufficient considering drivers' perceived ease of use and perceived usefulness. The combination of speech and augmented reality modalities was superior to the speech only modality from the perspective of ease of use. In Experiment 2, the results showed that only when information was presented in the speech modality, drivers had more anxiety about the *why* only information compared to the *why + what will* information; only when drivers were in high event criticality situations, they felt more prepared to take over control of the vehicle and had more preference for the combination of augmented reality and speech conditions than others.

5.2 Intellectual merit and broad impact

The proposed work can add to the knowledge base in takeover response investigation, takeover performance prediction, and in-vehicle alert system design.

First, the findings of this dissertation will enhance the understanding of how drivers' emotions, cognitive load, traffic density, and scenario type influence their takeover responses during conditionally automated driving. While driving behaviors alone give us insights into drivers' takeover performance, psychophysiological signals collected by non-invasive sensors allow us to estimate drivers' workload, emotions, attention, and situational awareness in a continuous and real-time manner. The findings provide us a broad picture of driver states throughout the whole takeover process.

Second, by leveraging models and methods from both human factors and machine learning, computational models that are capable of predicting drivers' takeover performance in real time were developed. Being able to predict drivers' takeover performance will facilitate the development of in-vehicle monitoring systems and adaptive alert systems in conditionally automated driving.

Finally, findings from this research can be used to provide design recommendations to automated vehicle manufacturers on adaptive in-vehicle alert systems. This will improve drivers' takeover performance, enhance the interaction between drivers and vehicles, and eventually increase driving safety and adoption of AVs.

Our society will benefit through human behavior investigation, computational model development, and alert system design that will enable the avoidance of unexpected performance breakdown and increase safety and performance in particularly demanding environments and workplaces.

5.3 Future work

Through a series of empirical studies, this dissertation can enhance our understanding of takeover responses and address challenges during takeover transitions. However, as with all research, there are several limitations that should be taken into consideration in the future as research opportunities.

First, we used a high-fidelity fixed-base driving simulator to imitate takeover situations in a controlled laboratory and recruited younger adults as participants. This is especially important when the psychophysiological measures collected are sensitive to various factors. However, the obtained results might be less ecologically valid than those obtained from on-road scenarios and across age and driving experience groups. Future studies can replicate the experimental settings with naturalistic driving and recruit diverse drivers to see the robustness of results. For example, drivers' individual demographic information (e.g., driving experience), vehicle dynamics, and brands may affect takeover performance and psychophysiological responses. Meanwhile, given the fact that drivers' internal states are associated with multiple psychophysiological measures, we used several of them to reliably measure subtle changes in drivers' cognitive load, attention, emotional states, and situational awareness. However, a variety of psychophysiological measures can be derived from the raw physiological signals. In future studies, more metrics, such as EEG, emotional arousal, frequency-domain HRV, and fixations, can be potentially included.

Second, this dissertation used a snapshot of the time-series data as model inputs, without considering the complexity of sequence dependence among the data. Future studies could try a convolutional neural network combined with long-short-term memory to predict drivers' takeover performance using a larger dataset. Meanwhile, the ground truth was determined by drivers' driving behaviors. It is necessary to propose a standard set of metrics for measuring takeover performance. An ensemble method combining subjective ratings, driving behaviors, and video coding can be explored to provide a more robust ground truth label of takeover performance. Instead of using a dichotomous classification of takeover performance, we could increase the number of classes (e.g., bad, neutral, good; or very bad, bad, neutral, good, very good) or use regression to see model prediction power. With more participants recruited from different ages, AV experience levels, and training groups in the future, the individual characteristics and power law of learning could be taken into account as model inputs to increase the generalization of models.

Third, our display design and evaluation studies are a preliminary effort to develop adaptive in-vehicle alert systems. With the advances of technologies in connected automated vehicle systems, real-time road environments such as traffic situations can be accessed easily in the future. Predictive model performance can be improved when data from various drivers engaging in different NDRTs in diverse environments are available for model training. The model outputs can contribute to the design of adaptive in-vehicle alert systems in conditionally automated driving. Specifically, if the system predicts that a driver would not be able to take over control successfully, a multi-modal display with the most effective information could be designed to help the driver realize the urgency of the event, augment situational awareness, and allocate attention properly. This paves the way for future work in which these alert systems can be implemented and tested on a real autonomous driving system outside the driving simulator.

APPENDICES

APPENDIX A

Questionnaires for Experiment 1 in Chapter 4

A.1 Perceived usefulness survey

Using the circles below, please indicate how much you agree with the following statements about the autonomous vehicle in the two video clips above.

	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
Using the display will improve my takeover performance.	○	○	○	○	○	○	○
Using the display increases my safety during takeover transitions.	○	○	○	○	○	○	○
Using the display enhances my effectiveness to take over control of the vehicle.	○	○	○	○	○	○	○
I find the display to be useful during takeover transitions.	○	○	○	○	○	○	○

A.2 Perceived ease of use survey

Using the circles below, please indicate how much you agree with the following statements about the autonomous vehicle in the two video clips above.

	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
The display is clear and understandable.	○	○	○	○	○	○	○
Interacting with the display does not require a lot of mental effort.	○	○	○	○	○	○	○
I find the display easy to use.	○	○	○	○	○	○	○

A.3 Behavioral intention to use survey

Using the circles below, please indicate how much you agree with the following statements about the autonomous vehicle in the two video clips above.

	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
Assuming I have access to the display function in my automated vehicle, I intend to use it.	○	○	○	○	○	○	○
Given that I have access to the display function in my automated vehicle, I predict that I would use it.	○	○	○	○	○	○	○

A.4 Attitude towards using a technology survey

Using the circles below, please indicate how much you agree with the following statements about the autonomous vehicle in the two video clips above.

	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
Using the display for taking over control of the vehicle in automated driving is a good idea.	○	○	○	○	○	○	○
Using the display for taking over control of the vehicle in automated driving is a wise idea.	○	○	○	○	○	○	○
I likethe idea of using the display during takeover transitions in automated driving.	○	○	○	○	○	○	○
Using the display during takeover transitions in automated driving would be a pleasant experience.	○	○	○	○	○	○	○

A.5 Usefulness ranking

From the videos you have watched, please rank the different versions of context information according to what you think was more useful (1 most useful - 3 least useful).

What: The perception of the elements in the environment	1
What + How: The perception of the elements in the environment and the projection of action in the near future.	2
How: The projection of action in the near future	3

APPENDIX B

Questionnaires for Experiment 2 in Chapter 4

B.1 Anxiety survey

How well do the following adjectives describe your feelings when you imagined yourself taking over control of the vehicle with the displays in the two video clips above?

	Extremely poorly	Very poorly	Poorly	Neutral	Well	Very well	Extremely well
Fearful	○	○	○	○	○	○	○
Afraid	○	○	○	○	○	○	○
Anxious	○	○	○	○	○	○	○
Uneasy	○	○	○	○	○	○	○

B.2 Performance expectancy survey

Using the circles below, please indicate how much you agree with the following statements about the autonomous vehicle in the two video clips above.

	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
I would find the display useful in my driving during takeover transitions.	○	○	○	○	○	○	○
Using the display would enable me to react to unsafe takeover conditions more quickly.	○	○	○	○	○	○	○
Using the display would improve my takeover performance.	○	○	○	○	○	○	○
If I use the display, I will decrease my risk of being involved in an accident.	○	○	○	○	○	○	○
Please select "Disagree" option for this question.	○	○	○	○	○	○	○

B.3 Effort expectancy survey

Using the circles below, please indicate how much you agree with the following statements about the autonomous vehicle in the two video clips above.

	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
My interaction with the display would be clear and understandable.	○	○	○	○	○	○	○
It would be easy for me to become skillful at using the display.	○	○	○	○	○	○	○
I would find the display difficult to use.	○	○	○	○	○	○	○
Learning to use the diplay would be easy for me.	○	○	○	○	○	○	○

B.4 Preference survey

How well do the following adjectives describe the displays of the autonomous vehicle in the two video clips above?

	Extremely poorly	Very poorly	Poorly	Neutral	Well	Very well	Extremely well
Intelligent	○	○	○	○	○	○	○
Efficient	○	○	○	○	○	○	○
Smart	○	○	○	○	○	○	○
High Quality	○	○	○	○	○	○	○
Reliable	○	○	○	○	○	○	○
Dependable	○	○	○	○	○	○	○
Effective	○	○	○	○	○	○	○
Helpful	○	○	○	○	○	○	○
Select "Neutral"	○	○	○	○	○	○	○

B.5 Takeover preparedness survey

How prepared would you feel if you had to take control of the vehicle at the moment of notice you just experienced?

	Extremely Bad	Very bad	Bad	Neutral	Well	Very well	Extremely well
Preparedness	○	○	○	○	○	○	○

REFERENCES

REFERENCES

AAA Foundation for Traffic Safety (2016), Prevalence of Self-Reported Aggressive Driving Behavior: United States, 2014, *Tech. Rep. Washington, D.C. AAA Foundation for Traffic Safety.*

Aashto, A. (2001), Policy on geometric design of highways and streets, *American Association of State Highway and Transportation Officials, Washington, DC*, *1*(990), 158.

Abdu, R., D. Shinar, and N. Meiran (2012), Situational (state) anger and driving, *Transportation Research Part F: Traffic Psychology and Behaviour*, *15*(5), 575–580.

Abraham, H., C. Lee, S. Brady, C. Fitzgerald, B. Mehler, B. Reimer, and J. F. Coughlin (2017), Autonomous vehicles and alternatives to driving: trust, preferences, and effects of age, in *Proceedings of the transportation research board 96th annual meeting (TRB'17).*

Alrefaie, M. T., S. Summerskill, and T. W. Jackon (2019), In a heart beat: Using driver's physiological changes to determine the quality of a takeover in highly automated vehicles, *Accident Analysis & Prevention*, *131*, 180–190.

Anderson, T. W. (2011), *The statistical analysis of time series*, vol. 19, John Wiley & Sons.

Baek, S. J., H. Yun, and J. H. Yang (2019), How do humans respond when automated vehicles request an immediate vehicle control take-over?, in *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings*, pp. 341–345.

Bagdadi, O., and A. Várhelyi (2011), Jerky driving—an indicator of accident proneness?, *Accident Analysis & Prevention*, *43*(4), 1359–1363.

Bagdadi, O., and A. Várhelyi (2013), Development of a method for detecting jerks in safety critical events, *Accident Analysis & Prevention*, *50*, 83–91.

Barrett, L. F. (1998), Discrete emotions or dimensions? the role of valence focus and arousal focus, *Cognition & Emotion*, *12*(4), 579–599.

Bashiri, B., and D. D Mann (2014), Heart rate variability in response to task automation in agricultural semi-autonomous vehicles, *The Ergonomics Open Journal*, *7*(1).

Bazilinskyy, P., S. M. Petermeijer, V. Petrovych, D. Dodou, and J. C. de Winter (2018), Take-over requests in highly automated driving: A crowdsourcing survey on auditory, vibrotactile, and visual displays, *Transportation research part F: traffic psychology and behaviour*, *56*, 82–98.

Benedek, M., and C. Kaernbach (2010), A continuous measure of phasic electrodermal activity, *Journal of neuroscience methods*, *190*(1), 80–91.

Bertola, M. A., and S. A. Balk (2011), Eyes on the road: A methodology for analyzing complex eye tracking data, *Proceedings of the 19th Annual Southeast SAS Users Group (SESUG)(Alexandria, VA)*.

Bidder II, W. H., and A. Tomlinson (1997), A comparison of saccadic and blink suppression in normal observers, *Vision research*, *37*(22), 3171–3179.

Bishop, R. (2019), Is 2020 the year for eyes-off automated driving?, *Forbes*.

Bolon-Canedo, V., N. Sanchez-Marono, and A. Alonso-Betanzos (2011), Feature selection and classification in multiple class datasets: An application to kdd cup 99 dataset, *Expert Systems with Applications*, *38*(5), 5947–5957.

Boucsein, W. (2012), *Electrodermal activity*, Springer Science & Business Media.

Bradley, M. M., and P. J. Lang (1994), Measuring emotion: the self-assessment manikin and the semantic differential, *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59.

Braunagel, C., W. Rosenstiel, and E. Kasneci (2017), Ready for take-over? a new driver assistance system for an automated classification of driver take-over readiness, *IEEE Intelligent Transportation Systems Magazine*, *9*(4), 10–22.

Breiman, L. (1996), Bagging predictors, *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001), Random forests, *Machine learning*, *45*(1), 5–32.

Bueno, M., E. Dogan, F. H. Selem, E. Monacelli, S. Boverie, and A. Guillaume (2016), How different mental workload levels affect the take-over control after automated driving, in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2040–2045, IEEE.

Carmines, E. G., and R. A. Zeller (1979), *Reliability and validity assessment*, vol. 17, Sage publications.

Castaldo, R., L. Montesinos, T. S. Wan, A. Serban, S. Massaro, and L. Pecchia (2017), Heart rate variability analysis and performance during a repeated mental workload task, in *EMBEC & NBC 2017*, pp. 69–72, Springer.

Chan, M., and A. Singhal (2013), The emotional side of cognitive distraction: Implications for road safety, *Accident Analysis & Prevention*, *50*, 147–154.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002), Smote: synthetic minority over-sampling technique, *Journal of artificial intelligence research*, *16*, 321–357.

Chen, C.-F., and P.-C. Chen (2011), Applying the tam to travelers' usage intentions of gps devices, *Expert Systems with Applications*, *38*(5), 6217–6221.

Chen, D.-R., Q. Wu, Y. Ying, and D.-X. Zhou (2004), Support vector machine soft margin classifiers: error analysis, *Journal of Machine Learning Research*, *5*(Sep), 1143–1175.

Choirunnisa, S., and J. Lianto (2018), Hybrid method of undersampling and oversampling for handling imbalanced data, in *2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pp. 276–280, IEEE.

Clark, H., and J. Feng (2017), Age differences in the takeover of vehicle control and engagement in non-driving-related activities in simulated driving with conditional automation, *Accident Analysis & Prevention*, *106*, 468–479.

Collet, C., A. Clarion, M. Morel, A. Chapon, and C. Petit (2009), Physiological and behavioural changes associated to the management of secondary tasks while driving, *Applied ergonomics*, *40*(6), 1041–1046.

Damböck, D., T. Weißgerber, M. Kienle, and K. Bengler (2012), Evaluation of a contact analog head-up display for highly automated driving, in *4th International Conference on Applied Human Factors and Ergonomics. San Francisco. USA*, Citeseer.

Davis, F. D. (1989), Perceived usefulness, perceived ease of use, and user acceptance of information technology, *MIS quarterly*, pp. 319–340.

Deffenbacher, J. L., D. M. Deffenbacher, R. S. Lynch, and T. L. Richards (2003), Anger, aggression, and risky behavior: a comparison of high and low anger drivers, *Behaviour Research and Therapy*, *41*(6), 701–718.

Dingus, T. A., F. Guo, S. Lee, J. F. Antin, M. Perez, M. Buchanan-King, and J. Hankey (2016), Driver crash risk factors and prevalence evaluation using naturalistic driving data, *Proceedings of the National Academy of Sciences of the United States of America*, *113*(10), 2636–2641.

Dogan, E., V. Honnêt, S. Masfrand, and A. Guillaume (2019), Effects of non-driving-related tasks on takeover performance in different takeover situations in conditionally automated driving, *Transportation research part F: traffic psychology and behaviour*, *62*, 494–504.

Dolcos, F., K. S. LaBar, and R. Cabeza (2004), Dissociable effects of arousal and valence on prefrontal activity indexing emotional evaluation and subsequent memory: an event-related fmri study, *Neuroimage*, *23*(1), 64–74.

Du, N., J. Ayoub, F. Zhou, A. Pradhan, L. Robert Jr, D. Tilbury, E. Pulver, and X. J. Yang (2019a), Examining the impacts of drivers' emotions on takeover readiness and performance in highly automated driving, *HFES 2019*.

Du, N., J. Haspiel, Q. Zhang, D. Tilbury, A. K. Pradhan, X. J. Yang, and L. P. Robert Jr (2019b), Look who's talking now: Implications of av's explanations on driver's trust, av preference, anxiety and mental workload, *Transportation research part C: emerging technologies*, *104*, 428–442.

Du, N., X. J. Yang, and F. Zhou (2020a), Psychophysiological responses to takeover requests in conditionally automated driving, *Accident Analysis & Prevention*, *148*, 105,804.

Du, N., F. Zhou, E. Pulver, D. Tilbury, L. P. Robert, A. K. Pradhan, and X. J. Yang (2020b), Predicting takeover performance in conditionally automated driving, in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–8.

Du, N., F. Zhou, E. Pulver, D. M. Tilbury, L. P. Robert, A. K. Pradhan, and X. J. Yang (2020c), Evaluating effects of cognitive load, takeover lead time, and traffic density ondrivers' takeover performance in conditionally automated driving, *AutomotiveUI '20*, doi:https://doi.org/10.1145/3409120.3410666.

Du, N., F. Zhou, E. Pulver, D. M. Tilbury, L. P. Robert, A. K. Pradhan, and X. J. Yang (2020d), Examining the effects of emotional valence and arousal on takeover performance in conditionally automated driving, *Transportation research part C: emerging technologies*, *112*, 78–87.

Ekman, P., W. V. Freisen, and S. Ancoli (1980), Facial signs of emotional experience., *Journal of Personality and Social Psychology*, *39*(6), 1125.

Engström, J. (2010), Scenario criticality determines the effects of working memory load on brake response time, in *Proceedings of the European Conference on Human Centred Design for Intelligent Transport Systems*, pp. 25–36.

Engström, J., E. Johansson, and J. Östlund (2005), Effects of visual and cognitive load in real and simulated motorway driving, *Transportation research part F: traffic psychology and behaviour*, *8*(2), 97–120.

Eriksson, A., and N. A. Stanton (2017), Takeover time in highly automated vehicles: noncritical transitions to and from manual control, *Human Factors*, *59*(4), 689–705.

Eriksson, A., S. M. Petermeijer, M. Zimmermann, J. C. De Winter, K. J. Bengler, and N. A. Stanton (2018), Rolling out the red (and green) carpet: supporting driver decision making in automation-to-manual transitions, *IEEE Transactions on Human-Machine Systems*, *49*(1), 20–31.

Feng, F., S. Bao, J. R. Sayer, C. Flannagan, M. Manser, and R. Wunderlich (2017), Can vehicle longitudinal jerk be used to identify aggressive drivers? an examination using naturalistic driving data, *Accident Analysis & Prevention*, *104*, 125–136.

Fornell, C., and D. F. Larcker (1981), Evaluating structural equation models with unobservable variables and measurement error, *Journal of marketing research*, *18*(1), 39–50.

Forster, Y., F. Naujoks, A. Neukum, and L. Huestegge (2017), Driver compliance to take-over requests with different auditory outputs in conditional automation, *Accident Analysis & Prevention*, *109*, 18–28.

Forster, Y., S. Hergeth, F. Naujoks, M. Beggiato, J. F. Krems, and A. Keinath (2019), Learning to use automation: Behavioral changes in interaction with automated driving systems, *Transportation research part F: traffic psychology and behaviour*, *62*, 599–614.

Fredrickson, B. L., and C. Branigan (2005), Positive emotions broaden the scope of attention and thought-action repertoires, *Cognition & Emotion*, *19*(3), 313–332.

Friedman, J. H. (1989), Regularized discriminant analysis, *Journal of the American statistical association*, *84*(405), 165–175.

Gao, H., A. Yüce, and J.-P. Thiran (2014), Detecting emotional stress from facial expressions for driving safety, in *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 5961–5965, IEEE.

Gelman, A. (2013), Commentary: P values and statistical practice, *Epidemiology*, *24*(1), 69–72.

Gelman, A., and J. Hill (2006), *Data analysis using regression and multilevel/hierarchical models*, Cambridge university press.

Ghazizadeh, M., Y. Peng, J. D. Lee, and L. N. Boyle (2012), Augmenting the technology acceptance model with trust: Commercial drivers' attitudes towards monitoring and feedback, in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 56, pp. 2286–2290, Sage Publications Sage CA: Los Angeles, CA.

Gilboy, J. (2021), Honda's now selling the world's first production car with level 3 self-driving tech, *The Drive*.

Gold, C., D. Damböck, L. Lorenz, and K. Bengler (2013), "take over!" how long does it take to get the driver back into the loop?, in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 57, pp. 1938–1942, SAGE Publications Sage CA: Los Angeles, CA.

Gold, C., M. Körber, D. Lechner, and K. Bengler (2016), Taking over control from highly automated vehicles in complex traffic situations: the role of traffic density, *Human Factors*, *58*(4), 642–652.

Gold, C., F. Naujoks, J. Radlmayr, H. Bellem, and O. Jarosch (2017), Testing scenarios for human factors research in level 3 automated vehicles, in *International conference on applied human factors and ergonomics*, pp. 551–559, Springer.

Gold, C., R. Happee, and K. Bengler (2018a), Modeling take-over performance in level 3 conditionally automated vehicles, *Accident Analysis and Prevention*, *116*, 3–13, doi:10.1016/j.aap.2017.11.009.

Gold, C., R. Happee, and K. Bengler (2018b), Modeling take-over performance in level 3 conditionally automated vehicles, *Accident Analysis & Prevention*, *116*, 3–13.

Groeger, J. A. (2013), *Understanding driving: Applying cognitive psychology to a complex everyday task*, Routledge.

Gross, J. J., and R. W. Levenson (1997), Hiding feelings: the acute effects of inhibiting negative and positive emotion., *Journal of Abnormal Psychology*, *106*(1), 95.

Haak, M., S. Bos, S. Panic, and L. Rothkrantz (2009), Detecting stress using eye blinks and brain activity from eeg signals, *Proceeding of the 1st driver car interaction and interface (DCII 2008)*, pp. 35–60.

Hancock, G., P. Hancock, and C. Janelle (2012), The impact of emotions and predominant emotion regulation technique on driving performance, *Work*, *41*(Supplement 1), 3608–3611.

Hayward, J. C. (1972), Near miss determination through use of a scale of danger.

Healey, J. A., and R. W. Picard (2005), Detecting stress during real-world driving tasks using physiological sensors, *IEEE Transactions on intelligent transportation systems*, *6*(2), 156–166.

Helldin, T., G. Falkman, M. Riveiro, and S. Davidsson (2013), *Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving*, 210–217 pp., ACM, New York, New York, USA, doi:10.1145/2516540.2516554.

Hergeth, S., L. Lorenz, and J. F. Krems (2017), Prior familiarization with takeover requests affects drivers' takeover performance and automation trust, *Human Factors*, *59*(3), 457–470.

Hidalgo-Muñoz, A. R., A. J. Béquet, M. Astier-Juvenon, G. Pépin, A. Fort, C. Jallais, H. Tattegrain, and C. Gabaude (2019), Respiration and heart rate modulation due to competing cognitive tasks while driving, *Frontiers in Human Neuroscience*, *12*, 525.

Hu, H., Z. Zhu, Z. Gao, and R. Zheng (2018), Analysis on biosignal characteristics to evaluate road rage of younger drivers: A driving simulator study, in *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 156–161, IEEE.

Huang, Q., and H. Wang (2004), Fundamental study of jerk: evaluation of shift quality and ride comfort, *Tech. rep.*, SAE Technical Paper.

Izquierdo-Reyes, J., R. A. Ramirez-Mendoza, M. R. Bustamante-Bello, J. L. Pons-Rovira, and J. E. Gonzalez-Vargas (2018), Emotion recognition for semi-autonomous vehicles framework, *International Journal on Interactive Design and Manufacturing (IJIDeM)*, *12*(4), 1447–1454.

Jacob, R. J., and K. S. Karn (2003), Eye tracking in human-computer interaction and usability research: Ready to deliver the promises, in *The mind's eye*, pp. 573–605, Elsevier.

Jaeggi, S. M., M. Buschkuehl, J. Jonides, and W. J. Perrig (2008), Improving fluid intelligence with training on working memory, *Proceedings of the National Academy of Sciences*, *105*(19), 6829–6833.

Jeon, M. (2017), *Emotions and affect in human factors and human-computer interaction*, Academic Press.

Jeon, M., B. N. Walker, and J.-B. Yim (2014), Effects of specific emotions on subjective judgment, driving performance, and perceived workload, *Transportation Research Part F: Traffic Psychology and Behaviour*, *24*, 197–209.

Keller, J. M., M. R. Gray, and J. A. Givens (1985), A fuzzy k-nearest neighbor algorithm, *IEEE transactions on systems, man, and cybernetics*, (4), 580–585.

Kensinger, E. A. (2004), Remembering emotional experiences: The contribution of valence and arousal, *Reviews in the Neurosciences*, *15*(4), 241–252.

Koo, J., J. Kwac, W. Ju, M. Steinert, L. Leifer, and C. Nass (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.

Koo, J., D. Shin, M. Steinert, and L. Leifer (2016a), Understanding driver responses to voice alerts of autonomous car operations, *International Journal of Vehicle Design*, *70*(4), 377–392.

Koo, J., D. Shin, M. Steinert, and L. Leifer (2016b), Understanding driver responses to voice alerts of autonomous car operations, *International Journal of Vehicle Design*, *70*(4), 377–392.

Körber, M., C. Gold, D. Lechner, and K. Bengler (2016), The influence of age on the take-over of vehicle control in highly automated driving, *Transportation Research Part F: Traffic Psychology and Behaviour*, *39*, 19–32.

Kramer, S. E., A. Lorens, F. Coninx, A. A. Zekveld, A. Piotrowska, and H. Skarzynski (2013), Processing load during listening: The influence of task characteristics on the pupil response, *Language and cognitive processes*, *28*(4), 426–442.

Kuehn, M., T. Vogelpohl, and M. Vollrath (2017), Takeover times in highly automated driving (level 3), in *25th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration*.

Kulke, L., D. Feyerabend, and A. Schacht (2020), A comparison of the affectiva imotions facial expression analysis software with emg for identifying facial expressions of emotion, *Frontiers in Psychology*, *11*, 329.

Lacey, J. I. (1967), Somatic response patterning and stress: Some revisions of activation theory, *Psychological stress: Issues in research*, pp. 14–37.

Lacey, P. (1970), Some autonomic-central nervous system inter-relations., *Physiological correlates of emotion.*, pp. 205–227.

Langlois, S., and B. Soualmi (2016), Augmented reality versus classical hud to take over from automated driving: An aid to smooth reactions and to anticipate maneuvers, in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1571–1578, IEEE.

Langlois, S., T. N. That, and P. Mermillod (2016), Virtual head-up displays for augmented reality in cars: a user testing to validate the congruence, in *Proceedings of the European Conference on Cognitive Ergonomics*, pp. 1–8.

Larue, G. S., A. Rakotonirainy, N. L. Haworth, and M. Darvell (2015), Assessing driver acceptance of intelligent transport systems in the context of railway level crossings, *Transportation Research Part F: Traffic Psychology and Behaviour*, *30*, 1–13.

Lee, J. J., B. Knox, J. Baumann, C. Breazeal, and D. DeSteno (2013), Computationally modeling interpersonal trust, *Frontiers in psychology*, *4*, 893.

Lee, S.-I., H. Lee, P. Abbeel, and A. Y. Ng (2006), Efficient l˜ 1 regularized logistic regression, in *AAAI*, vol. 6, pp. 401–408.

Lei, S., and M. Roetting (2011), Influence of task combination on eeg spectrum modulation for driver workload estimation, *Human factors*, *53*(2), 168–179.

Lemercier, C., et al. (2014), Inattention behind the wheel: How factual internal thoughts impact attentional control while driving, *Safety science*, *62*, 279–285.

Li, S., P. Blythe, W. Guo, and A. Namdeo (2018), Investigation of older driver's takeover performance in highly automated vehicles in adverse weather conditions, *IET Intelligent Transport Systems*, *12*(9), 1157–1165.

Libby Jr, W. L., B. C. Lacey, and J. I. Lacey (1973), Pupillary and cardiac activity during visual attention, *Psychophysiology*, *10*(3), 270–294.

Lisetti, C. L., and F. Nasoz (2004), Using noninvasive wearable computers to recognize human emotions from physiological signals, *EURASIP Journal on Applied Signal Processing*, *2004*, 1672–1687.

Lohani, M., B. R. Payne, and D. L. Strayer (2019), A review of psychophysiological measures to assess cognitive states in real-world driving, *Frontiers in human neuroscience*, *13*, 57.

Louw, T., G. Kountouriotis, O. Carsten, and N. Merat (2015), Driver Inattention During Vehicle Automation: How Does Driver Engagement Affect Resumption Of Control?

Louw, T., G. Markkula, E. Boer, R. Madigan, O. Carsten, and N. Merat (2017), Coming back into the loop: Drivers' perceptual-motor performance in critical events after automated driving, *Accident Analysis & Prevention*, *108*, 9–18.

Luo, R., et al. (2019), Toward real-time assessment of workload: A bayesian inference approach, in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, pp. 196–200, SAGE Publications Sage CA: Los Angeles, CA.

Ma, S., W. Zhang, Z. Yang, C. Kang, C. Wu, C. Chai, J. Shi, Y. Zeng, and H. Li (2020), Take over gradually in conditional automated driving: The effect of two-stage warning systems on situation awareness, driving stress, takeover performance, and acceptance, *International Journal of Human–Computer Interaction*, pp. 1–11.

Madigan, R., T. Louw, M. Dziennus, T. Graindorge, E. Ortega, M. Graindorge, and N. Merat (2016), Acceptance of automated road transport systems (arts): an adaptation of the utaut model, *Transportation Research Procedia*, *14*, 2217–2226.

Madigan, R., T. Louw, M. Wilbrink, A. Schieben, and N. Merat (2017), What influences the decision to use automated public transport? using utaut to understand public acceptance of automated road transport systems, *Transportation research part F: traffic psychology and behaviour*, *50*, 55–64.

Madigan, R., T. Louw, and N. Merat (2018), The effect of varying levels of vehicle automation on drivers' lane changing behaviour, *PloS one*, *13*(2), e0192,190.

McDonald, A. D., J. D. Lee, C. Schwarz, and T. L. Brown (2014), Steering in a random forest: Ensemble learning for detecting drowsiness-related lane departures, *Human factors*, *56*(5), 986–998.

McDonald, A. D., H. Alambeigi, J. Engström, G. Markkula, T. Vogelpohl, J. Dunne, and N. Yuma (2019a), Toward Computational Simulations of Behavior During Automated Driving Takeovers: A Review of the Empirical and Modeling Literatures, doi:10.1177/0018720819829572.

McDonald, A. D., H. Alambeigi, J. Engström, G. Markkula, T. Vogelpohl, J. Dunne, and N. Yuma (2019b), Toward computational simulations of behavior during automated driving takeovers: A review of the empirical and modeling literatures, *Human factors*, *61*(4), 642–688.

Mehler, B., B. Reimer, J. F. Coughlin, and J. A. Dusek (2009), Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers, *Transportation Research Record*, *2138*(1), 6–12.

Mehler, B., B. Reimer, and J. A. Dusek (2011), Mit agelab delayed digit recall task (n-back), *Cambridge, MA: Massachusetts Institute of Technology*, p. 17.

Mehler, B., B. Reimer, and J. F. Coughlin (2012), Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: an on-road study across three age groups, *Human factors*, *54*(3), 396–412.

Meng, F., and C. Spence (2015), Tactile warning signals for in-vehicle systems, *Accident analysis & prevention*, *75*, 333–346.

Merat, N., A. H. Jamson, F. C. Lai, and O. Carsten (2012), Highly automated driving, secondary task performance, and driver state, *Human factors*, *54*(5), 762–771.

Miller, D., and W. Ju (2014), Does the first officer concur? shared control with smart vehicle systems, in *Adjunct Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 1–6, ACM.

Miller, D., M. Johns, B. Mok, N. Gowda, D. Sirkin, K. Lee, and W. Ju (2016), Behavioral measurement of trust in automation: the trust fall, in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 60, pp. 1849–1853, SAGE Publications Sage CA: Los Angeles, CA.

Miranda Correa, J. A., M. K. Abadi, N. Sebe, and I. Patras (2018), Amigos: A dataset for affect, personality and mood research on individuals and groups, *IEEE Transactions on Affective Computing*, pp. 1–1, doi:10.1109/TAFFC.2018.2884461.

Mok, B., M. Johns, K. J. Lee, D. Miller, D. Sirkin, P. Ive, and W. Ju (2015), Emergency, automation off: Unstructured transition timing for distracted drivers of automated vehicles, in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pp. 2458–2464, IEEE.

Mole, C. D., O. Lappi, O. Giles, G. Markkula, F. Mars, and R. M. Wilkie (2019), Getting back into the loop: the perceptual-motor determinants of successful transitions out of automated driving, *Human factors*, *61*(7), 1037–1065.

Molnar, L. J. (2017), Age-related differences in driver behavior associated with automated vehicles and the transfer of control between automated and manual control: a simulator evaluation, *Tech. rep.*, University of Michigan, Ann Arbor, Transportation Research Institute.

Molnar, L. J., L. H. Ryan, A. K. Pradhan, D. W. Eby, R. M. S. Louis, and J. S. Zakrajsek (2018), Understanding trust and acceptance of automated vehicles: An exploratory simulator study of transfer of control between automated and manual driving, *Transportation research part F: traffic psychology and behaviour*, *58*, 319–328.

Morris, D. M., J. M. Erno, and J. J. Pilcher (2017), Electrodermal response and automation trust during simulated self-driving car use, in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 61, pp. 1759–1762, SAGE Publications Sage CA: Los Angeles, CA.

Naragon-Gainey, K., T. P. McMahon, and T. P. Chacko (2017), The structure of common emotion regulation strategies: A meta-analytic examination, *Psychological Bulletin*, *143*(4), 384–427, doi:10.1037/bul0000093.

Nass, C., I.-M. Jonsson, H. Harris, B. Reaves, J. Endo, S. Brave, and L. Takayama (2005), Improving automotive safety by pairing driver emotion and car voice emotion, in *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, pp. 1973–1976, ACM.

Naujoks, F., C. Mai, and A. Neukum (2014), The effect of urgency of take-over requests during highly automated driving under distraction conditions, *Advances in Human Aspects of Transportation*, *7*(Part I), 431.

Naujoks, F., C. Purucker, K. Wiedemann, A. Neukum, S. Wolter, and R. Steiger (2017), Driving performance at lateral system limits during partially automated driving, *Accident Analysis & Prevention*, *108*, 147–162.

Navarro, J., F. Osiurak, and E. Reynaud (2018), Does the tempo of music impact human behavior behind the wheel?, *Human Factors*, *60*(4), 556–574.

Nees, M. A., B. Helbein, and A. Porter (2016), Speech auditory alerts promote memory for alerted events in a video-simulated self-driving car ride, *Human factors*, *58*(3), 416–426.

Pêcher, C., C. Lemercier, and J.-M. Cellier (2009), Emotions drive attention: Effects on driver's behaviour, *Safety Science*, *47*(9), 1254–1259.

Petermeijer, S., P. Bazilinskyy, K. Bengler, and J. De Winter (2017), Take-over again: Investigating multimodal and directional tors to get the driver back into the loop, *Applied ergonomics*, *62*, 204–215.

Petersen, L., L. Robert, X. J. Yang, and D. Tilbury (2019), Situational Awareness, Driver's Trust in Automated Driving Systems and Secondary Task Performance, *SAE International Journal of Connected and Automated Vehicles*, *2*(2).

Picard, R. W. (2003), Affective computing: challenges, *International Journal of Human-Computer Studies*, *59*(1-2), 55–64.

Pohlmeyer, B. M. B. R. A., and J. Coughlin (2008), The association between heart rate reactivity and driving performance under dual task demands in late middle age drivers, *Advances in Transportation Studies, Int. J., Special Issue,(2008)*, pp. 53–70.

Politis, I., S. Brewster, and F. Pollick (2013), Evaluating multimodal driver displays of varying urgency, in *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 92–99.

Prusa, J., T. M. Khoshgoftaar, D. J. Dittman, and A. Napolitano (2015), Using random undersampling to alleviate class imbalance on tweet sentiment data, in *2015 IEEE international conference on information reuse and integration*, pp. 197–202, IEEE.

Quinlan, J. R., et al. (1996), Bagging, boosting, and c4. 5, in *AAAI/IAAI, Vol. 1*, pp. 725–730.

Radlmayr, J., C. Gold, L. Lorenz, M. Farid, and K. Bengler (2014), How traffic situations and non-driving related tasks affect the take-over quality in highly automated driving, in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 58, pp. 2063–2067, Sage Publications Sage CA: Los Angeles, CA.

Radlmayr, J., F. M. Fischer, and K. Bengler (2018), The influence of non-driving related tasks on driver availability in the context of conditionally automated driving, in *Congress of the International Ergonomics Association*, pp. 295–304, Springer.

Rahman, M. M., M. F. Lesch, W. J. Horrey, and L. Strawderman (2017), Assessing the utility of tam, tpb, and utaut for advanced driver assistance systems, *Accident Analysis & Prevention*, *108*, 361–373.

Ratwani, R. M., J. M. McCurry, and J. G. Trafton (2010), Single operator, multiple robots: an eye movement based theoretic model of operator situation awareness, in *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, pp. 235–242, IEEE Press.

Reimer, B., B. Mehler, J. F. Coughlin, N. Roy, and J. A. Dusek (2011), The impact of a naturalistic hands-free cellular phone task on heart rate and simulated driving performance in two age groups, *Transportation research part F: traffic psychology and behaviour*, *14*(1), 13–25.

Rezvani, T., K. Driggs-Campbell, D. Sadigh, S. S. Sastry, S. A. Seshia, and R. Bajcsy (2016), Towards trustworthy automation: User interfaces that convey internal and external awareness, in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 682–688, IEEE.

Richardson, N. T., L. Flohr, and B. Michel (2018), Takeover requests in highly automated truck driving: How do the amount and type of additional information influence the driver–automation interaction?, *Multimodal Technologies and Interaction*, *2*(4), 68.

Rish, I., et al. (2001), An empirical study of the naive bayes classifier, in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, pp. 41–46.

Robert, L. P. (2019), Are automated vehicles safer than manually driven cars?, *AI & SOCIETY*, pp. 1–2.

Rowe, G., J. B. Hirsh, and A. K. Anderson (2007), Positive affect increases the breadth of attentional selection, *Proceedings of the National Academy of Sciences*, *104*(1), 383–388.

Russell, J. A. (1980), A circumplex model of affect., *Journal of Personality and Social Psychology*, *39*(6), 1161.

Schartmüller, C., K. Weigl, P. Wintersberger, A. Riener, and M. Steinhauser (2019), Text comprehension: Heads-up vs. auditory displays: Implications for a productive work environment in sae level 3 automated vehicles, in *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 342–354.

Schmidt, E. A., M. Schrauf, M. Simon, M. Fritzsche, A. Buchner, and W. E. Kincses (2009), Drivers' misjudgement of vigilance state during prolonged monotonous daytime driving, *Accident Analysis & Prevention*, *41*(5), 1087–1093.

Shabanpour, R., N. Golshani, A. Shamshiripour, and A. K. Mohammadian (2018), Eliciting preferences for adoption of fully automated vehicles using best-worst analysis, *Transportation research part C: emerging technologies*, *93*, 463–478.

Shen, S., and D. M. Neyens (2014), Assessing drivers' performance when automated driver support systems fail with different levels of automation, in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 58, pp. 2068–2072, Sage Publications Sage CA: Los Angeles, CA.

Shin, H. S., C. Lee, and M. Lee (2009), Adaptive threshold method for the peak detection of photoplethysmographic waveform, *Computers in biology and medicine*, *39*(12), 1145–1152.

Sibi, S., H. Ayaz, D. P. Kuhns, D. M. Sirkin, and W. Ju (2016), Monitoring driver cognitive load using functional near infrared spectroscopy in partially autonomous cars, in *2016 IEEE Intelligent Vehicles Symposium (IV)*, pp. 419–425, IEEE.

Society of Automotive Engineers (2018), Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems.

Sokolov, E., and N. Paramonova (1961), Progressive changes in the orienting reflex in man during the development of sleep inhibition., *Zhurnal vysshei nervnoi deiatelnosti imeni IP Pavlova*, *11*, 217–226.

Sokolov, E. N. (1963), Higher nervous functions: The orienting reflex, *Annual review of physiology*, *25*(1), 545–580.

Solovey, E. T., M. Zec, E. A. Garcia Perez, B. Reimer, and B. Mehler (2014), Classifying driver workload using physiological and driving performance data: two field studies, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 4057–4066, ACM.

Sparrow, R., and M. Howard (2017), When human beings are like drunk robots: Driverless vehicles, ethics, and the future of transport, *Transportation Research Part C: Emerging Technologies*, *80*, 206–215, doi:10.1016/J.TRC.2017.04.014.

Steinhauser, K., F. Leist, K. Maier, V. Michel, N. Pärsch, P. Rigley, F. Wurm, and M. Steinhauser (2018), Effects of emotions on driving behavior, *Transportation Research Part F: Traffic Psychology and Behaviour*, *59*, 150–163.

Stöckli, S., M. Schulte-Mecklenbeck, S. Borer, and A. C. Samson (2018), Facial expression analysis with affdex and facet: A validation study, *Behavior research methods*, *50*(4), 1446–1460.

Takayama, L., and C. Nass (2008), Assessing the effectiveness of interactive media in improving drowsy driver safety, *Human factors*, *50*(5), 772–781.

Thirunavukkarasu, G. S., H. Abdi, and N. Mohajer (2016), A smart hmi for driving safety using emotion prediction of eeg signals, in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 004,148–004,153, IEEE.

Trick, L. M., S. Brandigampola, and J. T. Enns (2012), How fleeting emotions affect hazard perception and steering while driving: The impact of image arousal and valence, *Accident Analysis & Prevention*, *45*, 222–229.

Uhrig, M. K., N. Trautmann, U. Baumgärtner, R.-D. Treede, F. Henrich, W. Hiller, and S. Marschall (2016), Emotion elicitation: A comparison of pictures and films, *Frontiers in psychology*, *7*, 180.

Ünal, A. B., D. de Waard, K. Epstude, and L. Steg (2013), Driving with music: Effects on arousal and performance, *Transportation Research Part F: Traffic Psychology and Behaviour*, *21*, 52–65, doi:https://doi.org/10.1016/j.trf.2013.09.004.

Underwood, G., P. Chapman, S. Wright, and D. Crundall (1999), Anger while driving, *Transportation Research Part F: Traffic Psychology and Behaviour*, *2*(1), 55–68.

van den Beukel, A. P., and M. C. van der Voort (2013), The influence of time-criticality on situation awareness when retrieving human control after automated driving, in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pp. 2000–2005, IEEE.

Varma, S., and R. Simon (2006), Bias in error estimation when using cross-validation for model selection, *BMC bioinformatics*, *7*(1), 91.

Veltman, J., and A. Gaillard (1996), Physiological indices of workload in a simulated flight task, *Biological psychology*, *42*(3), 323–342.

Venkatesh, V., and F. D. Davis (1996), A model of the antecedents of perceived ease of use: Development and test, *Decision sciences*, *27*(3), 451–481.

Venkatesh, V., and F. D. Davis (2000), A theoretical extension of the technology acceptance model: Four longitudinal field studies, *Management science*, *46*(2), 186–204.

Venkatesh, V., M. G. Morris, G. B. Davis, and F. D. Davis (2003), User acceptance of information technology: Toward a unified view, *MIS quarterly*, pp. 425–478.

Walch, M., K. Lange, M. Baumann, and M. Weber (2015), Autonomous driving: investigating the feasibility of car-driver handover assistance, in *Proceedings of the 7th international conference on automotive user interfaces and interactive vehicular applications*, pp. 11–18.

Wan, J., and C. Wu (2018), The effects of lead time of take-over request and non-driving tasks on taking-over control of automated vehicles, *IEEE Transactions on Human-Machine Systems*, (99), 1–10.

Wandtner, B., N. Schömig, and G. Schmidt (2018), Effects of non-driving related task modalities on takeover performance in highly automated driving, *Human factors*, *60*(6), 870–881.

Wang, Y., B. Reimer, J. Dobres, and B. Mehler (2014), The sensitivity of different methodologies for characterizing drivers' gaze concentration under increased cognitive demand, *Transportation research part F: traffic psychology and behaviour*, *26*, 227–237.

Wintersberger, P., A. Riener, and A.-K. Frison (2016), Automated driving system, male, or female driver: Who'd you prefer? comparative analysis of passengers' mental conditions, emotional states & qualitative feedback, in *Proceedings of the 8th international conference on automotive user interfaces and interactive vehicular applications*, pp. 51–58.

Wintersberger, P., A. Riener, C. Schartmüller, A.-K. Frison, and K. Weigl (2018), Let me finish before i take over: Towards attention aware device integration in highly automated vehicles, in *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pp. 53–65.

Wintersberger, P., A.-K. Frison, A. Riener, and T. v. Sawitzky (2019), Fostering user acceptance and trust in fully automated vehicles: Evaluating the potential of augmented reality, *PRESENCE: Virtual and Augmented Reality*, *27*(1), 46–62.

Wright, T. J., R. Agrawal, S. Samuel, Y. Wang, S. Zilberstein, and D. L. Fisher (2017), Effects of alert cue specificity on situation awareness in transfer of control in level 3 automation, *Transportation research record*, *2663*(1), 27–33.

Wright, T. J., R. Agrawal, S. Samuel, Y. Wang, S. Zilberstein, and D. L. Fisher (2018), Effective cues for accelerating young drivers' time to transfer control following a period of conditional automation, *Accident Analysis & Prevention*, *116*, 14–20.

Wu, C., H. Wu, N. Lyu, and M. Zheng (2019), Take-over performance and safety analysis under different scenarios and secondary tasks in conditionally automated driving, *IEEE Access*, *7*, 136,924–136,933.

Young, K. L., P. M. Salmon, and M. Cornelissen (2013), Missing links? the effects of distraction on driver situation awareness, *Safety science*, *56*, 36–43.

Zeeb, K., A. Buchner, and M. Schrauf (2015), What determines the take-over time? an integrated model approach of driver take-over after automated driving, *Accident Analysis & Prevention*, *78*, 212–221.

Zeeb, K., A. Buchner, and M. Schrauf (2016), Is take-over time all that matters? the impact of visual-cognitive load on driver take-over quality after conditionally automated driving, *Accident Analysis & Prevention*, *92*, 230–239.

Zeeb, K., M. Härtel, A. Buchner, and M. Schrauf (2017), Why is steering not the same as braking? the impact of non-driving related tasks on lateral and longitudinal driver interventions during conditionally automated driving, *Transportation Research Part F: Traffic Psychology and Behaviour*, *50*, 65–79.

Zhang, B., J. de Winter, S. Varotto, R. Happee, and M. Martens (2019), Determinants of take-over time from automated driving: A meta-analysis of 129 studies, *Transportation research part F: traffic psychology and behaviour*, *64*, 285–307.

Zhou, F., A. Alsaid, M. Blommer, R. Curry, R. Swaminathan, D. Kochhar, W. Talamonti, L. Tijerina, and B. Lei (2020a), Driver fatigue transition prediction in highly automated driving using physiological features, *Expert Systems with Applications*, p. 113204.

Zhou, F., S. Kong, C. C. Fowlkes, T. Chen, and B. Lei (2020b), Fine-grained facial expression analysis using dimensional emotion model, *Neurocomputing*, *392*, 38–49.